The New Encyclopaedia Britannica

in 30 Volumes

MACROPÆDIA
Volume 15

Knowledge in Depth

FOUNDED 1768
15 TH EDITION

Encyclopaedia Britannica, Inc.
William Benton, Publisher, 1943–1973
Helen Hemingway Benton, Publisher, 1973–1974

Chicago
Auckland/Geneva/London/Manila/Paris/Rome
Seoul/Sydney/Tokyo/Toronto

First Edition	1768-1771
Second Edition	1777-1784
Third Edition	1788-1797
Supplement	1801
Fourth Edition	1801-1809
Fifth Edition	1815
Sixth Edition	1820-1823
Supplement	1815-1824
Seventh Edition	1830-1842
Eighth Edition	1852-1860
Ninth Edition	1875-1889
Tenth Edition	1902-1903

Printed in U.S.A.

Library of Congress Catalog Card Number: 81-70041
International Standard Book Number: 0-85229-400-X

Proboscidea

The order Proboscidea comprises three suborders and about 300 species of terrestrial mammals. All but two species, the Asiatic, or Asian, elephant ***(Elephas maximus)*** and the African elephant ***(Loxodonta africana)***,are extinct. The elephants are the largest surviving land animals and, among the mammals, are exceeded only by the whales in size.

The Proboscidea are characterized by columnar limbs, bulky bodies, and elongated snouts. In recent forms, testes are internal. The snout is a long boneless proboscis, or trunk; it is a combination of the upper lips, palate, and nostrils. Some of the incisor teeth develop into tusks. One extinct suborder (Deinotherioidea) lost the upper tusks; certain others have lost the lower ones and evolved upper tusks of dentine from which the enamel has partially or completely disappeared. The canine teeth were generally repressed in all groups, and the cheek teeth developed rows of blunt cones or ridges. In later forms, the temporary teeth were replaced by permanent ones, which are pushed by an escalator-like movement along a horizontal plane, so that the front teeth were replaced by teeth moving forward from the rear. The skull, which originally was elongated, became shorter, higher, and bulkier in later forms. The back of the eye orbit remained open instead of forming a complete bony ring, and the nasal opening in all Proboscidea is at a higher horizontal plane than the eye sockets. The neck shortened as the animals evolved larger, higher bodies and an elongated trunk that also functions as a hand. The skull has enlarged out of proportion to the brain in order to serve as an anchor for the trunk and to support the heavy dentition. This order occurs in all the continents except Australia. Fossils of proboscideans provide valuable information about early humans who were their contemporaries.

Drawing by Christian D. Olsen

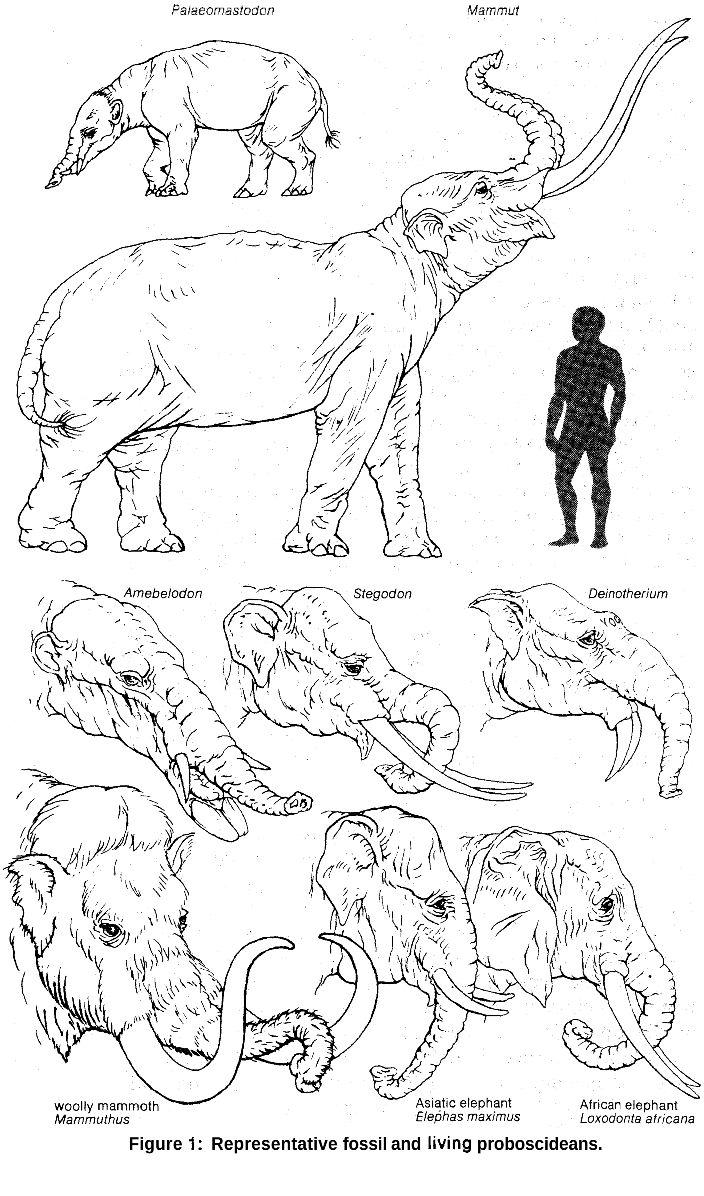


Figure 1: Representative fossil and living proboscideans.

GENERAL FEATURES

Size range and distribution. In Europe, the landmass that broke up to form the islands of the Mediterranean Sea harboured proboscideans. Three fossil species have been found in Malta; one had had a height of 2.1 metres, or 6.9 feet ***(Palaeoloxodon mindriensis)***, another a height of 1.5 metres, or 4.9 feet ***(P. melitensis)***, and the third was less than one metre, or about three feet ***(P. falconeri)***. ***P. creticus*** of Crete was 1.5 metres, and ***P. cypriotis*** of Cyprus was 0.9 metre (three feet) in height. In North America, a ***Mammuthus*** isolated on Santa Rosa Island, off the coast of southern California, was probably derived from ***Mammuthus meridionalis***, a species that stood **4.2** metres (13.8 feet) at the shoulders.

Early human records of elephants

Elephas maximus asurus lived in Iran and Syria. Early drawings of the animal and fragmentary skeletal remains indicate that it was the largest subspecies of the Asian elephant. The war elephants employed by Pyrrhus in 255 BC and engraved upon Roman seals show animals of unusual size. "Sarus," which signified "the Syrian," was the outstanding animal in the elephant battle squadron of the Carthaginian general Hannibal. In 1500 BC elephants ***(Elephas maximus rubridens)*** existed in China as far north as Anyang, in northern Honan Province. Writings from the 14th century state that elephants were still to be found in Kwangsi Province.

Man as well as other environmental factors exterminated the woolly mammoth ***(Mammuthus primigenius)*** and the imperial mammoth ***(M. imperator)*** about 10,000 years ago. Several races of the living species of Asian and African elephants also died out by about 1500 BC. The small North African race became extinct by the 2nd century AD, and some of the American mastodons, such as ***Cuvieronius postremus*** of South America, died out as recently as the 4th century AD. The large African bush elephants ***(Loxodonta africana)*** were exterminated from the Transvaal in South Africa early in the 20th century, but they still occur over much of the continent south of the Sahara Desert. A smaller elephant inhabits the forests of western equatorial Africa, particularly in the Congo region. It is considered by some to be a subspecies (***Loxodonta africana cyclotis)*** of the African elephant; others believe it to be represented by several subspecies; still others consider it to be a separate species (L. ***cyclotis)***.

In Asia, elephants have been exterminated from Iran, Iraq, Afghanistan, the northwestern part of India, and from much of the Malay Peninsula, Java, and the greater part of Borneo and Sri Lanka (formerly Ceylon). Isolated colonies remain in forest areas of Mysore in the peninsular part of India, in Assam, Nepal, Burma, Malaya, southern China, Sri Lanka, Sumatra, Borneo, and other islands of the East Indies.

Capture of elephants

Importance to man. Elephants constitute the chief source of commercial ivory. Because of the continuing demand for this commodity, the animals are in danger of extermination. Elephant "pearls" consist of concentric layers of ivory deposited over a foreign object that has been intruded into the soft ivory at the base of the growing tusk. From early times elephants have been used as beasts of burden in India and Burma. Since they do not breed freely in captivity, new stock for domestication is often captured from wild herds. One method is to drive them through a funnel-shaped stockade into a small enclosure; trained tame elephants help to subdue, noose, and train young captives for service. The training and handling of an elephant are usually entrusted to one man, called the mahout in India and the oozie in Burma; an elephant and its keeper frequently became inseparable companions.

Trained elephants carry humans in a howdah, or miniature hunting lodge, on their backs. They are used to move timber or other heavy materials.

In modern times, African elephants have been trained for labour only since late in the 19th century; however, they were used by the Carthaginians in wars with the Romans. Elephants depicted on Roman medals of the 2nd century AD have heads and bodies of the Asian but ears of the African species. This suggests that both species at that time were tamed. Elephants were used as executioners in Roman amphitheatres and for military pageants. They are still used in exhibitions at circuses, carnivals, and zoological parks. In warfare elephants have been used to drag heavy equipment, especially through mud and up steep slopes. As late as World War II, they were of value in military movements over the mud-clogged roads of Southeast Asia.

The association between man and elephants goes far back into mythology, and a rich folklore has developed. Bracelets of hairs from the tail of a freshly-killed or living elephant are prized as good-luck charms.

Attempts to locate the legendary "graveyards" to which old elephants allegedly resort when near death have been, for the most part, unsuccessful. The groups of buried elephant remains that have been found probably represent sites where elephants drowned in bogs or rivers or perished from imbibing poisonous water.

NATURAL HISTORY

Reproduction and life cycle. Tuskers, or tusked bulls, occasionally fight brief, savage duels that may end in death for the defeated animal. A duel between tuskless elephants may last for days, with occasional periods of rest. The female selected from the herd by the winner often makes an apparent attempt to escape from him. After a brief preliminary courting, the male mounts the female from behind, leaning over her back and either gripping her body or resting his forefeet upon her pelvis, and assumes a standing posture. Copulation lasts for about 20 seconds, with very little movement or noise. Mating continues promiscuously for about two days, after which the most powerful bull drives off the others and remains with the cow for about three weeks. The period of gestation varies from 20 months for a female calf to 22 months for a male. When parturition is about to occur, the herd surrounds the cow, who assumes a squatting position while giving birth.

In regions where large carnivores, such as tigers, prey upon newborn elephants, the cow seeks a female associate. The mother and the other elephants in the herd blow dust upon the moist, newborn calf to dry it. Two hours after birth, the baby is able to stand and is suckled. The mother and calf then join the herd.

Tame cows begin breeding at the age of eight or nine years; tame bulls begin when about 11 or 12 years of age. The interval between the birth of successive calves is 'about four years. In captivity cows are known to continue bearing calves until 60 or 70 years of age.

Appearance of elephant calf

The newborn elephant is about one metre (three feet) high and weighs about 90 kilograms (200 pounds). It is covered with yellow and brown hair. After a few months the hair on some parts of the body is as long as that in the extinct mammoth. In E. *maximus* there is also a pinkish patch around each eye, and when the calf is about five months old, a faint whitish patch develops on each cheekbone. As this patch spreads, similar patches appear upon the trunk and ears. In the more easterly races of Malaya and Sumatra there are only a few gray spots. The hoof nails, which are dark at first, gradually become lighter. When the elephant is about eight years old, a thick oily secretion known as musth, or must, begins to flow from a gland in the temporal, or temple, region. It occurs in both sexes, but is more active in males. The secretion increases each year until it drips into the elephant's mouth. Some authorities believe that the function of the secretion is to inhibit feeding; others believe it has some effect on sexual activity.

An elephant is not fully grown until it is about 25 years old. In the wild, the average life-span is about 80 years, but under optimum conditions an elephant may live for 120 years.

Behaviour. The organization of an elephant herd is often according to age and sex. In Elephas, although herds of 10 or 15 females and their young may appear to be under the leadership of a large female, and their organization matriarchal, young adult males are always in the vicinity, as is the real leader of the entire group, a large male. The leader may be accompanied by one young adult male who acts as a scout, warning the leader of danger. The herd also has a system of scouts, and, before emerging into an open area, one of the scouts usually explores it. If no danger is apparent, he signals by trumpeting to the herd to advance. Individuals often serve as guards while the rest of the herd feeds or bathes.

The herd is held together both by blood relationship and by a strong sense of companionship. If an individual is injured, three or four others surround it, shielding it from danger, supporting it, and helping it to move away. A calf that has lost its mother is adopted by the other cows in the herd even if they have their own calves to raise.

Ecology. Elephants clear paths through forests that are too dense for other animals. Many modern roads in elephant-inhabited countries originated in this manner. Elephants browse to a height of about five metres (16 feet), thereby increasing the amount of sunlight available for shrubs. Their uprooting of grass and roots aerates the soil and stimulates the growth of plants that replace the ones devoured. Mud wallows frequented by elephants are fertilized by their excreta.

Feeding habits

An elephant may destroy or discard as much vegetation as it consumes. An adult may eat between 250 to 350 kilograms (550 to 750 pounds) of solid food each day. When grazing, the animal uses its trunk or forefoot to gather grass, which is slapped against a forelimb to rid it of sand. In rainy weather, when soil is more difficult to shake off, the animal browses. Asian elephants break off branches; African elephants are more likely to push over trees. The wood apple (Feronia elephantorum) is a favourite food of the Asian elephant. The animal also eats wild rice that grows in forest lakes and various other aquatic plants. The African bush elephant eats the fruit of various palm trees. The spongy wood of the baobab tree provides some water during periods of drought.

FORM AND FUNCTION

Extant forms. The adult elephant has a tuft of hair at the tip of the tail and sometimes a patch of hair on the head. The limbs are adapted for bearing great weight: the legs are straight and pillar-like, and the bones of the joints are flat at the articular surfaces. Each toe has a heavy hoof nail; the weight is borne on thick pads behind the toes. The nose and upper lip are extended into a long trunk, which contains the nasal passages and has nostrils at the tip. Water for drinking is sucked into the trunk and then discharged directly into the mouth. The trunk is used for placing food into the mouth, for spraying and dusting the body, for lifting or moving heavy objects, and even for throwing objects at man.

The upper second incisors are typically developed into ivory tusks, the longest and heaviest teeth of any living animal. Canine teeth are absent. The complex molars are of the high-crowned type, with transverse rows of enamel ridges on the grinding surface, which is often traversed by a longitudinal median groove. There is normally only one complete functional tooth and half of a second one at a time on each side of each jaw. These are replaced horizontally from the rear as they wear away.

The Asian elephant (Elephas *maximus*) and its races are distinguished from the African elephants by being somewhat smaller and by having relatively small ears with the upper edge curled forward. The head is more domed, is structurally more complex, and has a greater development of diploe, or bony cell cavities. E. *maximus* also has high-crowned teeth and a relatively smooth trunk with a single fingerlike projection at the tip. Adult bulls weigh up to six tons and commonly stand 3.3 metres (10.8 feet) at the withers. Only the males develop tusks, which average 1.5 metres (4.9 feet) in length, the pair weighing about 32 kilograms (71 pounds). These develop flat, longitudinal planes of wear near the tip. Tusks 2.7 metres (8.9 feet) long and 68 kilograms (150 pounds) in weight have been recorded. About 90 percent of the males of the Ceylonese race lack tusks; Sumatran elephants are of slighter build and have longer trunks.

In contrast to the Asian species, the African bush elephant is generally larger. This and the forest form have extremely large ears (one metre in breadth), with the upper edge curled backward, and a roughened, heavily ringed trunk with two projections at the tip. The molars are of coarser construction, have fewer ridges, are less crenulated (scalloped), and consist of a thicker surface of enamel over thick plates of dentine. The top of the head is not domed, and the forehead is more convex. The tusk tips are usually conical, the legs are longer, and the eyes are relatively larger than those of the Asian elephant.

The average height of adult bull bush elephants is 3.3 metres (10.8 feet) at the shoulder, and the average weight is six tons. Cows are about 0.6 metre (two feet) shorter. Both sexes possess tusks, which average about 1.8 metres (5.9 feet) long, the pair weighing 36 to 55 kilograms (79 to 121 pounds). A pair of tusks in the British Museum weighs about 133 kilograms (293 pounds); the larger one of the two is 3.5 metres (11.5 feet) long, with a basal circumference of 46 centimetres (18 inches). The largest elephant on record, a bull bush elephant killed in the Cuando river district of southeastern Angola in 1955, is on display at the Smithsonian Institution in Washington, D.C.; it probably weighed 10 tons when alive and stood four metres (13 feet) at the shoulder. Adult forest elephants are about 2.1 metres (6.9 feet) tall and weigh 1,225 kilograms (2,700 pounds), with slender tusks that are often more than three metres long. The ears are relatively small and smoothly rounded at the margins.

Albinos

Albino, or "white," elephants occasionally occur, especially in Thailand and Burma, where they are regarded by some as semisacred.

The mammoth. The genus Mammuthus contains some of the largest members of the family Elephantidae. Certain ones reached a height of over 4.2 metres (13.8 feet) at the shoulders. One was M. meridionalis of Asia and Europe, and the other was M. imperator, which entered North America during the upper Pleistocene.

The genus also contains one of the most specialized members of the family, the woolly mammoth, M. *primigenius*, which probably became extinct about 10,000 years ago. It inhabited the sub-Arctic area of Asia and Europe and eventually entered North America over the Bering Strait; it travelled southward across western North America almost to Wyoming, then spread eastward toward Lake Michigan. Its height of 3.3 metres (10.8 feet) at the shoulders equalled that of a large *Elephas maximus,* but the body was relatively shorter, and the hindquarters sloped downward. Its skull was compressed from front to back. As an adaptation to its cold environment, the woolly mammoth evolved small ears, a short goatlike tail, and a coat of dense, furry, short hair overlain by longer, bristly hair. It also had a humplike reserve of fat upon the top of the head and on the shoulders. A subcutaneous layer of fat about eight centimetres thick covered the body. The molars had as many .as 27 lamellae, or plates. The tusks of the male were about 4.8 metres long. Their almost circular curvature and great size suggest that they may have functioned as shovels for exposing vegetation buried under snow. The mammoth is one of the few extinct proboscideans in which the carcass has often been completely preserved. Among the best known are two from Siberia—one discovered near the mouth of the Lena River in 1804; the other in the bank of the Beresovka River in 1900.

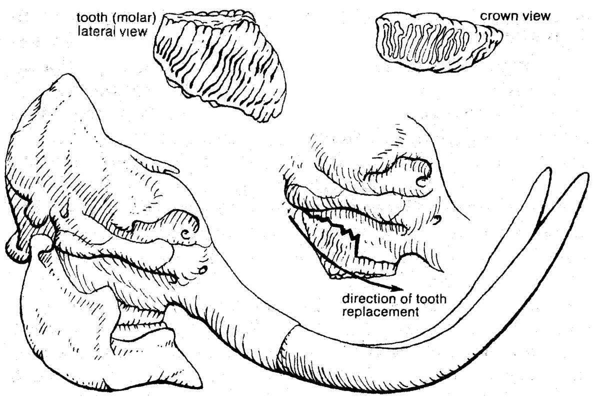


Figure 2: Skull and tooth movement in a mammoth *(Mammuthus primigenius).*

Drawing by Christian D. Olsen; adapted from A.S. Romer, Vertebrate Paleonfoiogy. © 1933. 1945. 1966 by the University of Chicago: all rights reserved

EVOLUTION AND CLASSIFICATION

Origin of the proboscis

Historical development and paleontology. The order Proboscidea has evolved from unknown ancestors that were not much larger than pigs. They flourished during the Paleocene Epoch (65,000,000 to 54,000,000 years ago). During the course of evolution, the lower jaw elongated beyond the upper, and the tusks projected well beyond the upper ones. At this stage the nose, palate, and upper lips developed into an elongated fleshy cover to the projecting lower jaw. It is probable that the nostrils opened well above the extremity of this flap and were near the eyes. A longer lower jaw proved less useful than a shorter one, so the upper flap was converted into a multipurpose tubular proboscis. Because the nostrils shifted to the tip of the proboscis, the animal was able to breathe while submerged in water. When so submerged, the animal had to rely on its sense of smell more than sight to detect approaching predators.

The suborder Deinotherioidea, consisting of one genus, is an early branch of the main proboscidean stock of the Eocene Epoch (54,000,000 to 38,000,000 years ago). They lost their upper tusks and developed a downward-hooked, tusk-tipped mandible. Numerous species of the suborder occurred in Asia, Europe, and Africa and persisted into Pleistocene times (2,500,000 to 10,000 years ago). The largest was the European Deinotherium *gigantissimum*, which reached a height of 3.8 metres (12.5 feet) at the shoulders and existed during the Pliocene Epoch (7,000,000 to 2,500,000 years ago).

In the suborder Mastodontoidea, the family Gomphotheriidae comprises 15 genera, including the earliest members of the order, Phiomia and *Palaeomastodon.* The former were the size of donkeys, but the latter were as large as a modern Asian cow elephant. In this family the skull and neck are elongated, and the teeth low crowned. The second incisors are enlarged; the upper ones are compressed and vertical, and they retain a band of enamel. In the later evolved genera, the lower pair are bent forward, depressed, and expanded into shovellike structures that do not meet the upper tusks. The canines are absent. Among this family are Cordillerion of North America and Cuvieronius of South America. The latter became extinct as recently as AD 200 to 400.

Phiomia, which occurred in Egypt and India, was an archaic shovel-tusked form with an elongated neck. It flourished during the Oligocene (38,000,000 to 26,000,000 years ago). The mandible and its tusks became more shovellike in *Amebelodon* and *Platybelodon* of the Miocene (26,000,000 to 7,000,000 years ago).

In *Palaeomastodon* the skull shortened, and the tusks assumed a cylindrical shape. The genus occurs in middle Oligocene deposits of Egypt. In some later genera, such as *Anancus* and *Stegomastodon,* the jaws shortened and the lower tusks disappeared. In some shovel-tuskers, the mandible remained enlarged and continued to function as a shovel for digging plant bulbs. It was probably protected by a horny pad.

Mastodontidae contains the single genus *Mastodon.* It lacked lower tusks. The tusks were about two metres (seven feet) long. Some species attained a height of about three metres (ten feet), and were covered with hair. Parts of carcasses of the recently extinct American species *Mastodon americanus* have been discovered in peat deposits and swamps.

The earliest proboscideans in Asia are *Phiomia, Gomphotherium, Platybelodon,* and *Serridentinus,* of the family Gomphotheriidae, and *Stegolophodon,* of the family Elephantidae. All occurred in the Miocene of China and some in Burma and India.

In the suborder Elephantoidea, *Stegolophodon* is intermediate between the mastodons and elephants proper. Although this genus first appeared during the Miocene of Europe, Asia, and Africa, it persisted into the upper Pleistocene of these continents.

Stegodon, which occurred during the Pliocene in Asia and during the Pleistocene in Africa and Asia, evolved from *Stegolophodon.* The skull enlarged, the jaws shortened, and the lower tusks disappeared. *Stegodon zhaolongensis* of China was a large species with the most primitive teeth in the genus. The tusks of some species such as *Stegodon magnidens* of India were about 3.3 metres (10.8 feet) in length; in others, however, they were feeble. The molars were usually low crowned, and the depression between each pair of dental folds or plates was Y-shaped, rather than V-shaped or U-shaped as in other elephants.

The genus *Mammuthus* includes all species formerly placed under *Archidiskodon, Metarchidiskodon, Parelephas,* and *Stegoloxodon.* Its species occurred in the Pleistocene of Asia, Europe, America, and Africa. Several species had great bulk and heavy tusk development.

Development of pygmy forms

On various occasions during the Pleistocene Epoch, normal-sized elephants that inhabited a continent were isolated when a part of the landmass was separated into islands by the submergence of low-lying land. As these isolated colonies of elephants multiplied, their fodder supply decreased, and the animals gradually became smaller—an unsuccessful measure against extinction. This process is evident in the East Indies and Philippines, in the islands of the Mediterranean Sea, and in certain islands off the coast of southern California.

Classification. *Distinguishing taxonomic features.* The proboscideans are classified largely according to body size; shape of the skull; dentition; and the shape, size, and degree of reduction of enamel in the tusks. Groups marked with a dagger (†) are extinct and known only from fossils.

Annotated classification.

ORDER PROBOSCIDEA

Oligocene to present; North America, Eurasia, and Africa. Heavy-bodied (graviportal) animals with snout prolonged into a fleshy proboscis (trunk). Tusks developed from upper or lower incisors or both; canines absent; cheek teeth with transverse rows of blunt cones or ridges. Skull short, high; nasal openings at a higher horizontal plane than eyes. Body size medium to large; shoulder height from about 1 m (about 3 ft) to more than 4 m (13 ft). About 300 species, all extinct but two.

†Suborder Deinotherioidea

†Family Deirtotheriidae

Lower Miocene to upper Pleistocene; Europe, Asia, Africa. Upper tusks lacking; lower tusks curving downward from tip of lower jaw. One genus *(Deinotherium);* many species; height to about 3 m.

†Suborder Mastodontoidea

†Family Gomphotheriidae

Lower Oligocene to Recent (AD 200–400); Europe, Asia, North and South America. Skull and neck elongated. Teeth low-crowned; succession vertical. Later genera with protruding, shovellike lower incisors, others with substantial upper tusks and no lowers; premolars with 2 transverse crests, molars with 3. About 15 genera, several dozen species; shoulder height about 1 to 3 m.

†Family Mastodontidae (mastodons)

Lower Miocene to upper Pleistocene (possibly to early historic times); Europe, Asia, Africa, North America. Molars with rounded prominences, but no ridges; lower tusks absent, but upper incisors substantial, reaching over 2 m in length in males of some species. One genus *(Mastodon),* many species; shoulder height to at least 3 m.

Suborder Elephantoidea

Family Elepharztidae (elephants and mammoths)

Lower Miocene to present; fossils from Europe, Asia, East Indies, Africa, and North America; Recent species from Africa *(Loxodonta)* and southern Asia *(Elephas)* which have probably evolved from *Stegolophodon.* Six genera, with several dozen fossil and 2 Recent species; shoulder height from 1 to about 3.5 m. The epiphyses (ends of long bones in limbs) do not fuse until the last molars appear. The molars are replaced at least three times. Marrow disappears from the limb bones early in adulthood.

BIBLIOGRAPHY. L.S. DE CAMP, *Elephant* (1964), suitable for specialist and nonspecialist alike; RICHARD CARRINGTON, *Elephants: A Short Account of Their Natural History, Evolution and Influence on Mankind* (1958), an excellent study of living as well as extinct forms; P.E.P. DERANIYAGALA, *Elephas maximus, the Elephant of Asia* (1951), with much new information on the Asian elephant; and *Some Extinct Elephants, Their Relatives, and the Two Living Species* (1955), a remarkable miscellany of elephant lore and observation, ancient and recent; HENRY F. OSBORN, *Proboscidea* (1939), an extremely detailed and comprehensive pioneer work; A.S. ROMER, *Vertebrute Paleontology,* 3rd ed. (1966), a *sine qua non* for students of proboscidean evolution; IVAN T. SANDERSON, *The Dynasty of Abu: A History and Natural History of the Elephants and Their Relatives, Past and Present* (1962), a good all-around work.

(P.E.P.D.)

Procedural Law

Law, to be effective, must go beyond the determination of the rights and obligations of individuals and collective bodies to an indication of how these rights and obligations can be enforced. It must do this, moreover, in a systematic and formal way. Otherwise, the numerous disputes that arise in a complex society cannot be handled efficiently, fairly, without favouritism, and, equally important for the maintenance of social peace, without the appearance of favouritism. This systematic and formal way is procedural law. Procedural law, then, constitutes the sum total of legal rules designed to insure the enforcement of rights by means of the courts, and thus contrasts with substantive law, the sum total of the rules determining the essence of the rights and obligations.

Since procedural law is only a means for enforcing substantive rules, there are different kinds of procedural law, corresponding to the various kinds of substantive law. Criminal law, for example, is the branch of substantive law dealing with punishment for offenses against the public and has as its corollary criminal procedure, which indicates how the sanctions of criminal law must be applied. In many countries there is an administrative procedure for the enforcement of various rights, obligations, and interests regulated by administrative law. Substantive private law, which deals with the relations between private (that is, nongovernmental) persons, whether individuals or corporate bodies, has as its corollary the rules of civil procedure, and it is civil procedure to which this article is limited. In many countries, private law itself is subdivided into two branches, civil law, the law dealing with nonbusiness relationships, and commercial law, dealing with business relationships. Each often has its own set of courts. In such countries, it is, then, possible to subdivide civil procedure in general into two branches, civil procedure in the strict sense and commercial procedure.

Private law, as opposed to criminal or administrative law, does not usually require the parties to choose the courts to resolve their disputes. They may, in fact, and frequently do submit the disputes to one or more private individuals for resolution. A private individual chosen to resolve a dispute in a binding (rather than merely advisory) fashion is usually referred to as an arbitrator, and the procedure under which he acts, as arbitration. At present, arbitration also is ordinarily clothed with some form of governmental sanction; indeed, in some countries, in particular in England, arbitration and ordinary civil procedure may be quite closely connected.

Since each system of substantive law in the world must obviously be accompanied by a system of procedural law, the different systems of civil procedure existing in the world are too numerous to be discussed within the framework of one article or, indeed, an entire volume. This article deals comparatively with some of those systems of civil procedure that have had more than local or temporary significance. Procedure under U.S. Federal Rules of Civil Procedure, which is a good example of a modern Anglo-American procedure, and those of the French and Austrian codes of civil procedure, which represent different European views of civil procedure, receive the main emphasis.

Historical growth of procedural law

ROMAN LAW

Civil procedure in ancient Rome had a marked influence on later development on the European continent and, to some extent, in England. The procedure of very early Roman law left little permanent impact on the law. Highly formalized, it was based on strict compliance with rules of pleadings and was replaced during the 1st century BC by the more flexible formulary procedure that in some respects bears marked similarity to Anglo-American civil procedure. Law suits were divided into two phases. In the first phase, devoted to defining the issues, the parties presented their claims and defenses orally to a judicial official called a praetor, whose main function was to hear the allegations of the parties and then to frame a formula or instruction applicable to the issue presented by the parties. The praetor did not decide the merits of the case. Instead, with the consent of the parties, he selected from a list of approved individuals a private individual (judex), whose duty it was to hear witnesses, examine the proof, and render a decision in accordance with the applicable law contained in the formula. There was no appeal. The procedure facilitated growth and change in the law: by adapting existing formulas, or modifying them, the praetors were, in effect, able to change substantive rules of law.

The formulary system

The formulary system with its separation of fact finding and determination of the law was not followed in the provinces conquered by the Romans. There, administrative officials rendered justice under general administrative powers. In the late imperial period, the procedure used in the provinces was also introduced in Rome itself. The creative role of the praetor came to an end, the formulas were abolished, and the division of a lawsuit into two phases was also terminated. Lawsuits were now initiated by a written pleading. Appeals from lower to higher judges became possible, and the procedure lent itself to delay. As a result, parties often submitted their disputes to arbitration or to religious leaders for settlement. Consequently, the leaders of various religious communities, including in particular those of the Christian Church, came to exercise judicial functions that in the very late Roman Empire received a degree of state recognition.

MEDIEVAL EUROPE

The Germanic tribes that conquered the Roman Empire in the 5th century carried their own procedure with them into the conquered territories. That procedure was quite formalistic: in court, which often was the assembly of all the freeborn men of the district, the parties had to formulate their allegations in precise, traditional language; the use of improper words could mean the loss of the case. At this point the court determined what method of proof should be used: ordeal, judicial combat between the parties or their champions, or wager of law (whereby each side had to attempt to obtain more persons who were willing to swear on their oaths as to the uprightness of the party they were supporting). Roman law procedure, however, never entirely disappeared from the territories conquered by the Germanic tribes. In addition, a modified form of late Roman procedure was in use in the ecclesiastical courts that applied the still-developing canon law. This late Roman and canonical procedure appears to have been preferable to the Germanic procedure and gradually supplanted it in Italy and France, and somewhat later in Germany, though all elements of the Germanic procedure did not disappear. In Scandinavia, on the other hand, indigenous procedure was able to resist displacement by foreign law.

The Roman-canonical procedure, with its heavy reliance on written, rather than oral, presentations, necessitated representation by learned counsel. The whole procedure was divided into rigid stages. Precise rules governed the presentation of evidence; thus the concordant testimony of two male witnesses usually amounted to "full proof," and one witness was ordinarily insufficient to prove any matter, unless he was a high ecclesiastic. A court order was needed before testimonial evidence could be used; witnesses were ordinarily examined not before the full court but by a judge, with a court clerk or notary committing the witnesses' testimony to writing for later submission to the court. This complex procedure was ill-suited to the day-to-day needs of commerce; as a result, special courts operated by and for businessmen sprang up in important mercantile centres (maritime courts, commercial courts) to deal with matters of inland and maritime commerce.

Early codifications

As the Middle Ages ended, there was an increasing tendency to favour written over oral evidence. At the same time, there was a tendency to "nationalize" the general Roman-canonical procedure prevalent in much of Europe and to create national procedural laws. In 1667 in France, this led to the enactment by Louis XIV of the Ordonnance Civile, also known as Code Louis, a comprehensive code regulating civil procedure in all of France in a uniform manner. The Code Louis continued, with some improvements, many of the basic principles of procedure that had prevailed since the late Middle Ages.

COMMON LAW IN ENGLAND

Originally, procedure in English local and feudal courts resembled quite closely that of other countries with a Germanic legal tradition. But unlike the countries on the continent of Europe, England never romanized its indigenous procedure once the latter had become inadequate but instead developed a procedure of its own capable of substantial growth and adjustment. That England was able to do this seems to have been due to two reiated factors, both the result of the strong monarchy that followed the Norman invasion: the growth of the jury system and the establishment of a centralized royal court system. The former offered a substitute for the antiquated methods of proof of the traditional Germanic law—-ordeal, trial by battle, and wager of law—and the latter led to the creation of a definite legal tradition, the common law, and to the administration of justice through permanent professional judges, and their attendant clerks, instead of the popular assemblies or groups of wise men who rendered justice elsewhere (see also JURY; COMMON LAW).

Development of the writ

Royal courts could be used only if permitted by a special royal writing, or writ, issued in the name of the king. Such writs were at first issued when there was a complaint that local or feudal courts were not rendering justice. Later, they were issued in cases involving land; such a writ might direct the defendant to return the land or explain why he refused to do so or, later on, direct the sheriff to bring the defendant before the court so that he might answer for his conduct. Eventually the writs became standardized. Through ingenious fictions (assumptions, for judicial purposes, of facts that do not exist), substantially all litigation not reserved to the ecclesiastical or other specialized courts could be brought before the royal courts, a situation preferred by suitors, since the royal courts abandoned much of the awkward Germanic law of proof in favour of trial by jury sooner than did local courts.

As the system of royal courts developed, the parties, or rather their counsel, formulated the issues to be settled through their "pleadings" before the court in London; after that the issues would be tried before a jury in the county where the facts arose. The mechanics of pleading gradually became quite complex. Originally, Germanic pleading practices, which involved oral formulation of issues in rather precise words, prevailed. Eventually, the clerks of the court wrote a summary of these oral pleadings and later recorded the entire substance. The plaintiff had to plead facts that came within the writ used to start the action; the defendant could either generally deny the facts asserted by plaintiff or assert specific defenses.

The complexities of the common-law procedure led some parties to request relief directly from the king, who in medieval theory was considered as the ultimate fountainhead of justice. These requests were transferred to the royal chancery—that is, the office of the lord chancellor—which, in this way, developed into another court; it was supposed to deal "equitably" with cases in which the strict rules of the common law failed. In the course of time this function of the chancery developed into a body of well-defined rules known as "equity." Until the 16th century, the chancellors were generally ecclesiastics; hence procedure in chancery to obtain equity was to some extent influenced by canonical procedures. In particular, there was no jury trial, no writ circumscribing a precise cause of action, and so forth.

Movement toward reform

The procedure of the common-law courts and the existence of a separate procedure for equity matters were both taken over in the United States. In the 19th century there developed in both England and the United States movements to simplify procedural complexities. These involved several related approaches: (1) a reform in court organization, doing away with separate courts of equity and, to the extent they existed, with coordinated common-law courts of general jurisdiction and establishing a more rational system of appeals courts; (2) a reform of pleading, abandoning largely the need to plead a specific cause of action based on writs, and giving judges power to promulgate rules of procedure.

The reforms were not entirely successful; early court decisions interpreted the revised pleading rules in a restrictive fashion, and the merger of common-law and equity courts did not result in a complete merger of procedures. U.S. federal and state constitutions, for example, guaranteed a jury trial in all cases at common law, but not in equity.

PERIOD OF NATIONAL CODIFICATIONS IN EUROPE

Dissatisfaction with the system of judicial administration was a major cause of the French Revolution of 1789. One of the earliest actions taken by the newly constituted National Assembly was the creation of a new court system (1790). But no reform of a lasting nature was undertaken in the field of civil procedure. The introduction of a jury system was debated but was adopted for criminal cases only.

Napoleon attempted to restore normality and unity to France after the Revolution through the creation of codes encompassing an entire field of law and containing the best of both the old pre-Revolutionary and the Revolutionary law. His Code of Civil Procedure of 1806, however, relied heavily on the 1667 code but continued certain procedures created during the Revolution.

During the 19th century, codifications of procedural law were enacted in other countries (Italy in 1865 and Germany in 1877). They usually retained large elements of the Roman-canonical or French procedure and were often cumbersome and slow. Austria departed from the Roman-canonical model in 1895 with the adoption of a new Code of Civil Procedure. The new code adopted comprehensively the principle of oral presentation: only matters presented orally in open court were important for a decision of the case; writings could have only a preparatory role; witnesses were no longer heard before a delegated judge who prepared a written record but by the court or judge that actually decided the case; finally, the parties were obligated to present their cases fully and truthfully, and the judge was directed to make certain that all relevant facts were stated. These notions were widely followed by other countries when they amended their codes of civil procedure. Recent changes made in the French Code of Civil Procedure (particularly in 1958 and 1965) were to some extent inspired by the Austrian model. A somewhat contrary trend occurred in Italy, however, where later amendments to the more progressive 1942 Code of Civil Procedure to some extent reemphasized written presentations. A step contrary to some modern European thinking was also taken by the new Belgian Judicial Code of 1967 (effective 1969). It reduces the role of the judge and correspondingly increases that of the parties and their counsel. Even more atypical have been developments in Japan. That country adopted a Code of Civil Procedure, very largely modelled on the German Code of 1877, in the year 1890. In 1926, the code was amended in order to expedite procedures. Austrian ideas about the role of the judge were heavily relied on. But after the defeat of Japan at the end of World War II, an attempt was made to introduce some of the features of the American civil trial, with its heavy reliance on the presentation of facts by the parties' attorneys and the correspondingly less significant role of the judge. For a variety of reasons, the attempt was not entirely successful. Present Japanese law blends a procedure largely based on the German model with some features of Anglo-American origin.

Elements of procedure

CONSIDERATIONS PRIOR TO TRIAL OR MAIN HEARING

Jurisdiction, competence, and venue. The words jurisdiction and competence refer generally to the power of an official body (legislative, judicial, or administrative) to deal with a specific matter. This article is concerned with judicial jurisdiction, the power of a court to act. That power may depend on the relationship of the court to the subject matter of the action; in such an instance one speaks generally of subject matter jurisdiction. The jurisdiction of a court may also depend on the relationship between the court and the defendant in the action. As to that relationship, important conceptual differences exist between the countries of the common-law orbit, which usually refer to this problem as the question of "jurisdiction over the defendant" and countries with a continental European tradition, which are likely to subdivide the problem into questions of "international jurisdiction" (*i.e.*, which country may take the case) and questions of "territorial jurisdiction" (*i.e.*, courts in which part of the country may take the case). In the United States, questions of jurisdiction are complicated by the due process clause of the Constitution, which imposes limits on the states' power to confer jurisdiction on their courts. It has been suggested that the word jurisdiction should be used only when discussing the power of the courts in a state generally to act in a given situation without violating the due process clause, whereas the word competence should be used to refer to the power of a particular court in a state to act pursuant to the laws of that state, but this distinction has not been widely followed; frequently, the terms jurisdiction and competence are used interchangeably. (For a more detailed discussion, especially in relation to matters containing foreign elements, see CONFLICT OF LAWS.)

Venue

For reasons having to do with the historical tradition of the common-law courts—especially with the practice of the royal courts in London to send out judges to conduct trials throughout the country with the help of locally selected juries—in common-law countries the various higher courts existing in a given state are not ordinarily viewed as completely separate tribunals but essentially as parts of one overall court. Hence the question of "venue," which is usually not so problematical as lack of jurisdiction. The most common venue rule is that the action may be initiated where either the plaintiff or defendant resides, where the cause of action arose, or, if real property is involved, where the real property is situated. Even when all formal legal requirements of jurisdiction and venue are fulfilled, American courts are sometimes authorized to dismiss an action on the ground that the choice of court will create serious inconvenience for the parties or the court itself.

Parties. In spite of differences in terminology, rules prevailing in various legal systems concerning the parties to a case show some basic similarities. It is quite generally recognized that in order to participate in a law suit as a plaintiff or as a defendant, a party must have the capacity to sue and must, in addition, be a "proper" party (that is, have standing before the court).

All persons recognized as such by law, including corporations and even groups of individuals without formal corporate status, may, at least in the abstract, assert their rights in court and are liable to suit by others. In practice, however, the law obliges certain persons to act through another person. These persons, such as minors and mental incompetents, are usually said to lack procedural capacity, or to have it only to a limited extent, and must act through parents or guardians. Corporations can frequently sue in their own name, though some countries (such as Sweden) require that actions be brought by or against the board of directors or similar body.

All legal systems limit in some respects the number of individuals who may engage in lawsuits; generally, only persons who have an actual interest in the outcome of the lawsuit may sue or be sued. Furthermore, only a person who owns (or claims to own) the right or obligation under suit can be a party to a suit involving that right. In the United States this rule is frequently called the real party in interest rule, and similar rules are found elsewhere—for example, in Italy and France. Frequently the real party in interest will be the person who will ultimately benefit from any recovery obtained, but this is not true in all cases. In the United States, for instance, the trustee of a trust is deemed the real party in interest in connection with suits involving the trust, though any recovery obtained by him will ultimately benefit the beneficiaries of the trust. Because of the problems inherent in the real parties in interest rule, some modern codifications have omitted any reference to it.

In connection with matters of public law, the ability to sue is sometimes restricted less narrowly than in pure private-law actions. In France, for instance, citizens are able to bring actions in court to attack municipal expenditures (though not expenditures of the national government).

Class suits

Ordinarily, only parties to an action are bound by its outcome. But when a very large group may be affected by a particular controversy, it is frequently impractical for all menibers of the group to join in the litigation. For this reason, the law in the United States sometimes authorizes so-called class actions, in which a limited number of persons sue to vindicate the rights of a much larger class; in the end all members are bound by the outcome of the suit. Class actions are frequently, but by no means exclusively, used in actions involving shareholders of a corporation. Countries with a Roman-law tradition generally do not authorize class actions, though in some limited situations proceedings brought by one person may affect the rights of other persons not party to the suit.

Although a person is ordinarily free to decide for himself whether or not he wants to attempt to enforce his rights by legal proceedings, his refusal to do so may cause harm to others. For this reason, the laws of many countries authorize creditors, for instance, to prosecute actions of their debtors if the debtors fail to do so.

Legal controversies are not necessarily limited to two persons, one plaintiff and one defendant. Sometimes, for instance, in actions involving co-ownership or joint obligations, the rights of several parties may be so inextricably intertwined that, for all practical purposes, it is impossible to adjudicate the rights of one person standing alone. In such cases, the procedural rules of many countries require that all such persons be made parties to the lawsuit. In other cases, however, the presence of several individuals may be merely useful, but not absolutely essential, to a resolution of a dispute. In such cases the law simply "permits" the individuals to join in, or be brought into, the lawsuit. It is also possible that persons not originally participating in a lawsuit may find that their rights are affected in some manner, directly or indirectly, by such suit. To avoid a multiplicity of actions, such persons will often be authorized to intervene in the pending lawsuit, if their own claim has a sufficiently close connection in law or fact with it. In civil-law countries a person wishing to support the claim of some other party must proceed by way of direct intervention. In the United States, an individual who wants to promote the claim of some other party may ask the court for leave to appear as amicus curiae (friend of the court) so that he may present arguments in favour of the person he supports. In certain cases, furthermore, defendants are authorized to bring third parties into an action when, for instance, these third parties are or may be liable to the defendants on account of the claim asserted against the defendants. This is known as impleader.

Amicus curiae

In general, a person's capacity to sue or be sued is not affected by the fact that he is an alien or nonresident, unless a state of war exists between his home country and the country he wishes to sue. Even a state of war generally will not destroy capacity to be sued. But an alien may experience some disadvantages. Many countries, for example, withhold legal aid from aliens, particularly if the alien's home country does not grant reciprocity. More importantly, many European and Latin American countries require alien plaintiffs to post security to guarantee that they will be able to reimburse the defendant for the expenses of the lawsuit, and sometimes even for additional damage, should they lose the case. As a result of the 1954 Hague Convention on Civil Procedure and numerous other treaties, this security for costs has been eliminated between many countries. In the United States, the nationality of a party is not material to the issue of whether security for costs is due; any nonresident of the state where the action is brought is required to post security. The rule in most other countries with an English legal tradition is analogous.

Provisional remedies. Lawsuits frequently take a long time. A judgment in an action concerning whether or not the defendant has the right to cut down certain trees, for instance, will be of little value if, while the suit is pending, the trees have already been cut down; in like manner, a judgment for a sum of money will be quite useless if the defendant is able to conceal or squander his funds while the parties litigate. For these reasons, legal systems quite generally provide so-called provisional remedies that enable the plaintiff to obtain some guarantees that any judgment obtained against the defendant will not be in vain. There appears to be a rather remarkable similarity between remedies in common-law and civil-law countries, although the legal technicalities are often different. The provisional remedies are frequently available even before an action has been initiated; but in such a case, an action must ordinarily be started within a short period of time after the grant of the remedy.

Attachment

Some remedies serve to prevent the disappearance either of funds required for the payment of the eventual judgment or of specific property involved in litigation. This purpose is served by attachment (bringing the property under the custody of the law), replevin (an action to recover property taken unlawfully), or other similar remedies. Usually, the remedy is granted by a judge at the request of the plaintiff, upon a showing that certain facts exist that make it probable that the plaintiff has a good claim and that the payment of the judgment by defendant may be threatened. Attachment ordinarily involves the seizure of the property by an officer of the court, who will hold it pending final disposition of the case, or, occasionally, involves merely an order to the person holding the property not to dispose of it. Attachment is not necessarily limited to tangible property but can also be used in connection with all sorts of intangible property (such as money due, or bank accounts). These remedies are granted in a proceeding in which the defendant is not heard (*i.e.*, ex parte).

Other remedies are intended to stabilize a situation pending the outcome of litigation. In such instances, courts are frequently authorized to issue orders (known in Anglo-American law as temporary injunctions) commanding the parties to do or not to do certain acts that may cause irreparable harm to the other side while the suit is pending. In both civil-law and common-law countries, orders of this nature ordinarily are granted only after a hearing in which both sides appear. Sometimes a court order of an even more temporary and short-lived nature (temporary restraining order) may be obtained without hearing the other side.

In countries with a common-law tradition a person disobeying an injunction issued by a court is guilty of "contempt of court" and can be punished quite severely. In civil-law countries, punishment for contempt is largely unknown, and since broad orders to defendants may therefore be difficult to enforce, such orders are sometimes limited to specific narrow situations. The Supreme Court of the United States has recently cast some doubt on the propriety of granting ex parte remedies.

The commencement of the action. In Anglo-American procedure, finding the facts has traditionally been the function of the jury, which cannot conveniently be kept together except for a brief period of time; proof taking therefore has to be concentrated in one continuous episode, the trial. As a result, a lawsuit is generally divided into two stages, the first, or pleading, stage and the trial stage. At the pleading stage the parties notify each other of their claims and defenses; at the trial stage, they or their counsel prove their factual contentions before the jury primarily through the oral examination of witnesses produced by them. The verdict of the jury and the judgment based on it follow immediately thereafter.

A different general pattern is followed in countries whose procedure is based on the Roman-canonical procedure. Since there is no jury, there is no need for a concentrated trial, and the procedure consists essentially of a series of hearings at which counsel argue their clients' position and submit documentary evidence; any other form of evidence can be utilized only with a special court order definitely describing the type of evidence and the matter to be proved by it. Hearings with argument continue after the evidence has been taken.

The nature of the summons and the requirements of service. In most countries when a civil action is initiated, some form of notice to that effect must be served immediately upon the defendant. This notice may consist merely of a statement to the effect that the plaintiff is suing the defendant and that the defendant must appear in court on a specified day or be in default. Such a notice is commonly referred to as a summons, the successor to the old English "writ" initiating the action. When the notice of the lawsuit consists only of the summons, it is necessary, either at the same or a subsequent time, to supply the defendant with more specific information about the nature of the claim against him. This information is contained in plantiff's first pleading, the complaint.

In common-law countries it was originally necessary to deliver the summons to the defendant in person (personal service). Now, other forms of service to notify the defendant, such as leaving the summons with an agent, employee, or a person of suitable age at his home, are also permissible provided their intent is to apprise the defendant that the suit is pending. Service by publication in a newspaper is generally authorized only when no other form of service is reasonably possible.

Formal rules

In civil-law countries the summons proper is often combined with the statement of plaintiff's claim in a single document (*assignation* in France, *citazione* in Italy). Other detailed formal rules must often be observed, and the documents sometimes must be written on paper bearing tax stamps, a requirement still in force in Italy; in France copies must be presented to a tax office for "registration" for tax purposes. The document need not be served to the individual himself; a member of the household, or even a neighbour or janitor, usually will be an adequate recipient. In Austria and several other countries, service can be effected through the use of the mail.

Pleadings. Pleadings are the formal written documents by which the parties set forth their contentions. They serve several functions including giving notice of the nature of the claim or defense, stating the facts that each party believes to exist, narrowing the number of issues that ultimately must be decided, providing a means to determine whether the party has a valid claim or defense, and serving as a record of what has been actually decided once the suit is ended.

In the English common law the pleadings were primarily designed to state the legal theory relied upon and to narrow the issues to be tried. Accordingly, in common-law proceedings, the plaintiff and defendant alternately submitted documents, each responding to the one that preceded it, and narrowed the field of conflict until there remained only one issue, upon which the trial would be based. Because narrowing the issues was deemed of great importance, the parties were not allowed to plead alternative or contradictory states of fact and the defendant was permitted to rely on only one defense at one time.

In the United States during the 19th century, numerous procedural reforms were instituted. The parties were no longer required to plead on the basis of legal theories but instead were to allege a statement of facts constituting the cause of action or defense; the court could then apply any legal theory that was applicable under the facts alleged and later proved. The insistence upon fact pleading had substantial drawbacks, however, especially since the courts demanded a high degree of specificity, made technical distinctions between fact and evidence (forbidding the insertion of the latter in the pleading), and bound the parties to prove the facts alleged or lose the lawsuit. This last rule was particularly harsh since it forced the party to allege detailed facts early in the proceedings when he frequently was not yet certain precisely what facts had occurred.

Modern reforms have gone a long way toward elimination of the injustices of the former system. U.S. federal rules require only "a short and plain statement of the claim showing that the pleader is entitled to relief"; the defendant "shall state in short and plain terms his defenses." There is no requirement that legal theory be stated in the pleading nor that only facts be alleged. Other rules specifically permit the parties to plead alternative or contradictory claims or defenses and provide that in the usual case, only two pleadings, the complaint and the answer, shall be permitted. The effect of these changes has been to substantially downgrade the importance of the pleading stage of the lawsuit. The primary function of the pleadings is now only to give a general notice of the subject matter of the suit to the opposing party.

European pleadings

Under modem European codes, pleading problems, for a variety of reasons, have not been as pronounced as in Anglo-American law. The narrowing of issues is generally an essentially judicial function, to be achieved either at a special preliminary hearing or even at a plenary hearing before the full court; the creation of a permanent record is a function of the final judgment, which unlike the general—and therefore uninformative—verdict in an American jury trial must ordinarily contain a description of the facts and legal reasons on which it is based. Pleadings therefore serve primarily to inform the court and parties concerning their respective claims, a function of limited importance, since under some codes (such as the Austrian Code of Civil Procedure of 1895) only statements by the parties or their counsel in open court are fully effective for this purpose. In addition, amendments or changes can ordinarily be made without difficulty, though, in order to avoid dilatory tactics and surprise, some limitations exist. In France, for instance, it is not permissible to add new claims unless they are related to the existing claim, but new facts and legal arguments are permissible.

European pleadings tend to be more general than English and American, with fewer distinctions between ultimate facts, evidentiary facts, and matters of law. In some countries, such as France, it is usual to start the action by pleadings of the utmost generality, subject to further elaboration later.

Appearance. The summons or analogous document by which the plaintiff initiates his action quite generally commands the defendant to appear in court a specified number of days after its service. In case of failure to appear, he is threatened with a "default" judgment. In both the Anglo-American and the continental European systems the appearance in court is normally a legal fiction. The defendant "appears" by serving plaintiff with a notice indicating that he will defend the lawsuit and giving the name of the attorney or similar representative who will act for him in this connection. Certain other procedural steps indicating a willingness to defend the lawsuit are sometimes considered the equivalent of such a notice.

The time limits for the appearance vary greatly. European countries frequently provide a great many different time periods varying with the distance between defendant and the court where the action is pending. In France, for instance, a defendant residing in that country must appear within eight days, whereas one residing elsewhere in Europe has eight days and one month; these time periods can be shortened by judicial decision in case of urgency. In some countries the time to appear is fixed by the court. Less attention is usually paid to geography in the United States. In New York, for instance, the defendant must appear in 20 days if the summons was served personally, and 30 days if some other form of service was employed.

In some countries, where appearance involves either actual presence in courts, or at least the delivery of documents to the court (Italy, Sweden), plaintiff and defendant may both be required to appear.

The preparatory stage. As noted above, there is a fundamental difference between Anglo-American procedure and the civil-law procedure, with the Scandinavian countries taking a somewhat intermediate position. In countries whose procedure is based on English common law, the concentrated trial, traditionally before a jury, serves as a climax to earlier procedures. At this time, the parties attempt to prove the facts at issue, primarily through the presentation of oral evidence. The climax of a European proceeding, however, is the hearing before the full bench of judges—a hearing that is essentially devoted to argument of counsel and the presentation of documentary evidence; any other type of evidence usually requires a specific court order for its utilization. In both legal systems there are procedures to prepare for the trial or hearing.

In Anglo-American procedure a preparatory phase can be devoted to numerous purposes. First, since a jury trial is required only when there are disputes as to matters of fact, the court may be asked to make a decision on those cases that can be decided purely on legal matters, without any regard to the facts in dispute. This will be true, for example, when the court lacks jurisdiction or when it is obvious that a dispute between the parties as to the facts is more apparent than real. In these cases the party concerned will address a motion to the court (either a motion to dismiss for lack of jurisdiction or a motion for summary judgment) that can be decided immediately by a judge sitting alone, without waiting for the availability of a trial date.

It should also be noted that there may be a pretrial hearing before a judge, at which the judge will attempt to narrow the issues in controversy and, if possible, try to settle the case, thus making the trial unnecessary.

If the suit has not come to an end as a result of such preliminaries, the parties must prepare for trial. At the trial, evidence is presented in an uninterrupted fashion, without any possibility for additional proof after its close; each side in the end must stand or fall on the testimony presented by it.

European preparatory phase

The European system is in some ways similar to the Anglo-American. Frequently, such questions as jurisdiction can be decided in the preliminary phase, without waiting for the full hearing. The preliminary phase may also serve to narrow issues and produce a settlement. But differences in basic structure in some of the European codes lead to variations in emphasis. The absence of a concentrated trial, for instance, makes it much less important for a party to have detailed knowledge of the facts known by the other side. Further, proof proceedings sometimes occur during the preliminary phases rather than at the main hearing; though in Austria the full court holds hearings devoted to all aspects of the case, without distinguishing between matters considered preliminary and those more pertinent to the main hearing.

Pretrial motions. Because court calendars for jury trials are often extremely crowded, especially in the larger cities, the parties involved in a case often will resort to pretrial motions if there is any remote possibility that such an action would lead to a resolution of the dispute without trial. The party making the motion summons his opponent to appear before a judge designated for that purpose and transmits at that time copies of the papers pertaining to the motion, such as sworn statements (affidavits) of persons having knowledge of the facts or memorandums concerning the applicable law; the other side may submit opposing papers. At the time the judge hears the motion, attorneys for both sides argue briefly concerning the matter in question; no witnesses are heard. In addition to cases in which there may be a lack of jurisdiction, it may also occur that the right asserted by the plaintiff does not exist and that he is not entitled by law to relief; in either case a motion for dismissal would be made.

Summary judgment

On a somewhat different plane stands the motion for summary judgment. Frequently it appears that the issues of fact raised in the pleadings do not really exist. In such a case, since the outcome would not be in any reasonable doubt, a trial would be a mere formality. To avoid the needless expense and delay of a trial, a motion for summary judgment can be made. (The rules relating to this motion are strict so as to abridge neither the right of every man to his day in court nor the constitutionally guaranteed right to a jury trial.) The sole function of the judge is to determine if, from all the available evidence, there exists a material issue of fact that is honestly disputed; he is not to determine what the true facts are. If he finds a material issue of fact to be in dispute, he must deny the motion and set the case down for a future trial. If he finds no such issue, he may grant a final and binding judgment.

In those civil-law countries that have a preparatory phase before a single judge and a final hearing before a three-judge bench, procedural defenses similar to pretrial motions are ordinarily raised before the single judge. Sometimes, however, in cases of lack of jurisdiction or lack of competence a hearing is held before the full court. Where the issue is one of territorial competence the result may be the transfer of the case to the proper court. General summary proceedings have lost considerable importance in France and have been abandoned completely in Italy, but in actions involving claims based on negotiable or other written instruments, for instance, special procedures have been developed that permit a judgment to be obtained with great dispatch, particularly if the defendant has no effective defense on the merits.

Discovery procedures. In general, English common law lacked procedural devices aimed at giving the parties and the court advance notice of the factual contentions of both sides prior to the trial of the action. Whatever information was obtained by a party about the opposing party's case was received from the pleadings. This absence of discovery devices was a reflection of a judicial philosophy that held that surprise was a proper tactical device and that withholding information from one's opponent until trial would prevent an unscrupulous adversary from fabricating evidence. Limited discovery devices were, however, available in the equity courts.

Reforms were instituted in the 19th and 20th centuries. A mid-19th-century New York code, for example, provided that each party could serve written questionnaires on its adversary, could compel the adversary to produce documents prior to the trial, and could, under some circumstances, take the oral deposition of any witness, whether or not a party to the action. Even with these changes, discovery proceedings were limited. In 1938 new U.S. federal rules expanded the discovery process further. It was hoped that more complete disclosure would result in a more thorough preparation and presentation of cases, encourage pretrial settlement by making each party cognizant of the true value of his claim, and expose, at an early stage in the proceedings, insubstantial claims that should not go to trial.

Thus, a party may seek discovery not only of information that would be admissible at trial but also any information that, though not admissible, might lead to the discovery of admissible testimony. Some limitations remain, however; materials prepared in anticipation of the pending litigation by or for a party, for instance, are not discoverable unless the party seeking discovery shows a substantial need for the information and an inability to obtain substantially equivalent information by alternative means. Most discovery devices may be utilized without prior court approval and the procedures take place in lawyers' offices; judicial intervention must ordinarily be sought only when there is a dispute concerning the permissible scope of discovery or when there is a need to impose sanctions for failure to obey a court order compelling discovery.

European procedures to secure information

With the exception of procedures to secure, in advance of lawsuit, evidence that is in danger of being lost (for instance, because a witness may die), there are few procedures in civil-law countries to enable a party to secure information to use later. There are several reasons for this. Sometimes it has been viewed improper to compel a party to disclose information that may help his adversary. The absence of a concentrated trial makes it less important to have all information available at once, even more so because appeals ordinarily involve a rehearing of the whole case. The greater role given the judge in some countries (such as Austria and Germany) in bringing out factual matters further reduces the need to obtain information in anticipation of the hearing.

Consequently, discovery of documents is usually possible only in very limited cases, although a party that actually intends to use a document has to make it available to the other side. In France, for instance, production of documents to the other side is possible in bankruptcy and related commercial matters, and it is required in commercial cases generally that books be produced for inspection by the court. Traditionally, discovery of documents has been unavailable in noncommercial cases; legislation in **1965** did authorize the judge to request the parties to produce any documents, but this is production before the court, not for a party's use as such.

Pretrial conference. The discovery process frequently makes the parties aware of significant issues not previously considered or may make it clear that an issue considered important before discovery is no longer so. In order to provide a means for reflecting these changes and also to assist in simplifying the issues to be tried, shortening the time for trial, and possibly eliminating the need for trial completely, the court may direct the parties to appear before it for a pretrial conference.

At the conference, no testimony of witnesses is heard, and no formal adversary proceeding takes place. The attorneys representing the litigants, with the assistance of the judge, try to reach agreement on amendments to the pleadings, the elimination of issues raised at an earlier stage that are no longer deemed pertinent, and the crystallization of the real, controversial issues that must be determined at the trial.

An indirect benefit of the pretrial conference is the possibility that a settlement of the case will be reached by the parties without the necessity of trial. Although some authorities feel that this should be a primary goal of the pretrial conference, the prevailing view is that "settlements must be a by-product rather than the object of pre-trial, the primary aim being to improve the quality of the expected trial rather than to avoid it." It should be noted, however, that a considerable number of lawsuits, and the vast majority of personal injury cases, are settled before a final verdict.

In civil-law countries, procedures somewhat analogous in purpose to pretrial conferences are fairly prevalent. Since such preliminary hearings are ordinarily held before a single judge, rather than a formal three-judge court, a considerable saving of judicial time may result. In **1965,** France, for instance, reformed its practice in this respect. There, each case is assigned to a special "prehearing" judge, who sets time limits for the exchange of pleadings, decides how many pleadings after the original summons and complaint shall be used and when they shall be submitted, and may penalize dilatory parties by delivering a default judgment or, if both sides are dilatory, by striking the case off the calendar. In addition, he may call in the parties' counsel for a conference and must make sure that all documents that the parties intend to use at the main hearing have been filed. He may also call in the parties themselves for a conference concerning a possible settlement. He must, in short, either settle the case or put it in shape for the formal hearing.

THE TRIAL AND THE MAIN HEARING

The climactic and decisive part of an Anglo-American civil action is the trial, in which the parties present their proof in a concentrated fashion. The climactic event in a lawsuit based on European codes is the hearing before the full court. The differences between these two procedures are so fundamental that discussion of the two will be essentially separate.

The jury system. The following discussion will deal with the jury in terms of specific aspects of trial procedure. For a more detailed presentation of the jury system, see **JURY.**

Many of the procedural rules governing trials in civil actions have been designed to reflect the basic premise that the function of the jury is to determine the facts of the case, whereas the function of the judge is to determine the applicable law and to oversee the parties' presentation of the facts to the court. The consequences of the presence of the jury have been so pervasive that even in cases tried by a judge without a jury, the procedural rules designed to accommodate jury trials remain largely intact, with the important exception, of course, that the judge will determine both the facts and the law.

The order of trial. Although some variations may exist, a trial is conducted most frequently in the following manner. The attorneys for plaintiff and the defendant make opening statements to the jury, outlining what each conceives to be the nature of the case and what each hopes to prove as the trial proceeds. Next, the attorney for the plaintiff presents his case by calling witnesses, questioning them, and permitting them to be cross-examined by the attorney for the defense; when the former has concluded his presentation, the latter frequently will ask for a dismissal of the suit for failure of plaintiff to establish a prima facie case (that is, a case sufficient until contradicted by evidence); if this is unsuccessful, he will call and examine witnesses in order to establish his defenses, and these witnesses are subject to cross-examination by the plaintiff's attorney. The attorneys for each side then make a closing argument to the jury, marshalling the evidence presented in a light most favourable to their respective clients; the judge will instruct the jury on the applicable law; and the jury will retire to deliberate in private until it reaches a verdict, which will then be announced in open court.

The rules of evidence. Although the parties, and not the judge, are charged with the primary obligation to call and question the witnesses, the judge must act as arbiter in all disputes between the parties concerning the admissibility of evidence. When one party objects to the introduction of testimony, the judge will decide whether or not, in accordance with established rules of admissibility, the evidence sought to be introduced is to he heard by the jury. In general, the rules of evidence are designed to screen from the jury evidence that is either deemed not reliable or, if reliable, considered to be capable of confusing the jury in some way. As a consequence, evidence based on hearsay and, to some extent, opinion is prohibited. In keeping with the adversary system, the judge is not entitled to rule that evidence is inadmissible unless a party objects to its introduction. The party objecting to the evidence must state the grounds for his objection and the judge must permit the jury to hear the evidence unless the specified grounds given by the attorney are applicable. Even within this narrow framework, the judge's role is limited, for the rules of evidence leave little room for discretion on the part of the judge.

Directed verdicts. When the party having the burden of proof of an issue has completed its presentation to the jury, the opposing side may ask the court to rule as a matter of law that the evidence presented does not provide sufficient proof for a reasonable jury to find for the party who presented the evidence. When a judge so finds, he may "direct a verdict," thus in effect withholding from the jury the right to rule independently on the issues at all. It has been held that this device, if properly applied, is not a violation of the constitutional right to jury trial because similar devices have historically been available to judges and because a verdict is directed only when there has not been sufficient evidence introduced to create a material issue of disputed fact for the jury to decide. The granting of a directed verdict results in a final judgment, and the termination of the trial.

Instructions to the jury. It is the obligation of the judge, at the conclusion of the trial, to instruct the jury as to the applicable law governing the case in order to guide it in arriving at a just verdict. Although this is solely the judge's obligation, in practice the parties will propose instructions for his consideration. The judge then selects among the proposals that have been submitted and offers the parties the opportunity, out of the hearing of the jury, to object to any proposed instruction that they deem to be incorrect. Failure at this time to object generally precludes a party from arguing later that the instructions given were incorrect.

There has been much debate as to the relevance of jury instructions generally, some commentators urging that the jury seldom understands the instructions given or often ignores them. The charge, however, that the judge gave improper instructions to the jury is one of the most frequent grounds of error offered by parties when appealing an adverse decision.

Comments on evidence

In addition to the judge's obligation to charge the jury on the law, U.S. federal rules and some other procedural codes permit the judge to comment on the evidence. When it is permitted, the judge may give his opinion with regard to the merits of the case so long as he makes clear to the jury that this opinion is not binding and that the jury, not he, is solely responsible for finding the truth as to the facts in dispute.

Types of verdict. Most frequently the jury will be requested to return a general verdict—that is, a decision merely stating in general terms the ultimate conclusion that it has reached (for example, the award of X dollars to plaintiff). This form of verdict gives considerable leeway to the jury and permits, if it does not encourage, some deviation from a strictly logical and technical application of the law to the facts. An alternative that offers greater control over the decision-making process is the special verdict whereby the jury is instructed merely to answer a series of specific factual questions proposed by the judge, who will then himself determine the verdict, based upon the jury's responses to the questions asked. Because of the difficulty in drawing up questions that would cover completely the issues of the case, the special verdict is cumbersome and not frequently used.

New trial and other relief. After the trial is completed, either party may request the trial judge to vacate the verdict and grant a new trial. Innumerable grounds are available for requesting a new trial, including, for example, judicial error, excessiveness of the verdict, and jury misconduct. Considerable discretion is given the judge, and a decision to grant a new trial will seldom be overturned on appeal. The grant of a new trial, unlike the directed verdict, does not result in the judge substituting his opinion for that of the jury but only mandates another jury to hear the case at another trial. But in the very limited cases in which a judge may grant a directed verdict, he can also substitute his decision for that of the jury by a judgment not on the verdict.

The main hearing. In civil-law countries the hearing before the full court is the essential part of a civil action. At that hearing, counsel for both sides present argument as to the law and the facts of the case and submit documentary evidence. The hearing serves several purposes: it informs the court of the contentions of the parties, both legal and factual; it narrows the issues that may have been raised by the original pleadings; and it leads to the submission of at least one type of evidence, namely, documentary evidence. The extent of proof presentation and the narrowing of issues vary from country to country.

In such countries as Italy and France, which divide the lawsuit into a preparatory and a final stage, the judge in charge of the preparatory proceedings attempts to narrow the issues and may, for this purpose, examine the parties. In countries where there is only one stage, this process takes place during the full hearing. In general, in civil-law countries, evidence other than documentary evidence may be introduced only pursuant to a specific court order specifying the matter on which such evidence is to be received and the form that such evidence is to take (witness, experts, etc.). But again two forms are possible. Under the Austrian code of civil procedure, the court that decides the case must hear the witness, expert, or whatever. In such a case, an order will be made at the hearing and will be implemented by the calling of the witness or expert. Subsequently the arguments of counsel may continue, interrupted perhaps by a new proof order, should the court feel this to be necessary. In France and Italy the court or the judge of the prehearing phase will make an order for the hearing of a witness or expert, but the witness or expert will be heard by a single judge not ordinarily part of the court, who will prepare a summary of the testimony. Latei on, that summary will be submitted to the court; there will be additional argument and finally a decision will be made based on the record so made. Because witnesses or experts are always acting pursuant to court order, they are never considered a party's witness.

Types of proof proceedings. Various types of proof proceedings are generally available, including (1) hearing of witnesses who are not themselves parties; (2) the expert's report; (3) the examination of parties, either informally or pursuant to formal interrogatories.

A party wishing a witness to be heard must make an appropriate request, informing the other side of the name of the witness and the subject on which the witness is to be heard; this is to enable the party to prepare its own side of the case. At the examination the judge will ask the witness to state in narrative form what he knows about the precise issue mentioned in the proof order; subsequently, the judge may ask additional, clarifying questions. If counsel for both sides wish to propose questions, they must ordinarily put them to the judge, who presents them to the witness. A more or less extensive summary of the testimony is prepared immediately by a clerk under the direction of the judge and signed by the witness, the judge, and the clerk. In the case of witnesses who live too far away from the court where the action is pending, interrogation sometimes takes place in a local court.

The examination of an expert is obtained in the same manner as that of a witness. Although the parties may suggest an expert to the court, those chosen are ordinarily taken from a list of experts approved by the court. The expert is considered an impartial auxiliary of the court; his use is ordinarily limited to cases involving some technical or scientific problem. The court or judge issuing the proof order may authorize him to make certain scientific investigations (*e.g.,* in an automobile accident case, to examine the car involved) and to report thereon.

Parties not considered witnesses

Parties are not considered witnesses, and different procedures for parties ordinarily exist. **A** court is usually authorized informally to question parties, ordinarily not under oath, either on the court's own motion or on the request of a party. Though this questioning is designed mainly to narrow issues, it does also have a function in terms of evidence. In Austria and some other countries the judge questioning a party may put the party under oath if he feels this to be necessary for an elucidation of the truth. In other countries, a party may be challenged by his adversary to make a statement under oath.

Rules of evidence. In European courts, rules as to the admission of evidence are ordinarily quite liberal since there has been no need to develop complex rules to keep certain evidence from a jury. It is generally required that evidence relate directly to the facts in issue and be neither superfluous nor unduly repetitious. But since judicial review of lower court decisions on the admission of evidence is frequently quite limited, these requirements have never been developed into the detailed rules existing in Anglo-American law.

There are, however, some rules restricting testimonial evidence. In the first place, certain groups of persons, including parties and frequently close relatives of parties, may not be witnesses. Furthermore, confidential information acquired by certain professions (clergymen, doctors, lawyers, public officials, and others) may not ordinarily be divulged. In addition, there are rules requiring written proof in certain cases, such as birth, marriage, death, and some legal transactions.

Hearsay

The hearsay rule and its numerous exceptions are quite unknown in countries whose procedure is based on the European codes. That an individual may have only indirect knowledge of an event will not usually prevent his testimony as to the event, although it will affect the weight given it; at times, however, courts will refuse to hear testimony of a witness whose connection with the events in issue is utterly remote. The absence of a hearsay rule makes it possible to utilize many forms of documentary evidence not available in Anglo-American countries, such as written, unsworn statements of witnesses who do not later testify, written opinions of experts, and so on, though the courts do not necessarily give much credence to such items.

All modern European codes reject the Roman-canonical principle according to which a predetermined weight must be given to the various kinds of evidence; instead the court gives each item of evidence whatever weight seems reasonable under the circumstances of the individual case. There are some exceptions to this, however. Countries with a Latin tradition frequently accord great weight to public and notarial documents.

In general, it can be said that because the free evaluation of evidence is normally possible and because the production of evidence is so largely under judicial control, questions of burden of proof are much less important in countries governed by European procedural codes than in the Anglo-American system.

It is sometimes said that civil procedure in continental European countries is inquisitory in nature and that the strongly accusatory features of Anglo-American civil procedure have not found favour there. European courts often have an affirmative duty to clarify the issue and are frequently authorized to call witnesses or experts on their own motion, though the extent to which they make use of this power may vary. But the basic impetus for the lawsuit always comes from the parties. There are thus inquisitorial elements, but no true inquisitorial procedure.

Judgment and execution. *Drafting and form of judgment.* When proceedings are terminated, the court that has considered the case will render a judgment. In such a case one speaks of a final judgment. Judgments deciding some procedural matter but not terminating the proceedings are known as interlocutory judgments.

In American practice, the judgment of a court after a jury trial is presented in a stylized document that merely recites certain data relating to the lawsuit, such as the names of the parties, the fact that a jury verdict has been rendered, and the disposition to be made. No detailed grounds are given for the decision. If a judge decides a case without a jury, he is often required to indicate the factual and legal bases for his decision in order to facilitate appellate review; in practice, such findings, too, are often of a rather stylized nature. Courts sitting without juries sometimes prepare, in addition, an opinion in which their reasoning is explained in narrative form.

Judgments in civil-law countries quite generally consist of not only statements indicating the names of the parties and the like and the decision of the court but also an opinion in which the court explains its decision. The opinion may vary in style. In Germany and Austria, it is narrative in nature, as in the United States; in France, it is traditionally cast in the form of one long sentence consisting of a syllogism using the facts and the applicable law as premises. When the court consists of several judges, it is frequent practice in Anglo-American countries for judges who disagree with the decision of the majority to prepare and file dissenting opinions, in which they explain the reasons for their disagreements. In civil-law countries, such dissenting opinions are rarely allowed; indeed, the courts are generally forbidden from disclosing the position taken by an individual member.

Filing of judgment

Quite generally, originals of judgments are filed in court clerks' offices; the parties may then procure copies to use as they see fit. In some countries, the rules for the formal preparation, signing, and filing of judgments tend to be quite technical and complex; this is much less so in the United States. Furthermore, judgments must frequently be written on stamped paper or presented to some tax office for the payment of a tax.

The effect of judgments: res judicata; collateral estoppel. Judgments generally have a continuing effect on parties and others long after they are rendered. In some situations the doctrine of res judicata will grant a binding effect on issues determined in the lawsuit. The doctrine is intended to avoid excessive litigation and is known in some form in most countries. Thus, it is uniformly held in the United States that when a valid and final personal judgment in an action for the recovery of money is rendered in favour of the plaintiff, the plaintiff or his legal successors are prevented from instituting an action against the defendant on the *same* cause. In effect, what was considered in the first action, or even that which should have been considered but was not, cannot form the basis of a second action. This does not preclude a second lawsuit based on a *different* cause of action or claim, but the related doctrine of "collateral estoppel" will preclude the parties from relitigating in the second suit based on a different cause of action any issue of fact common to both suits that was actually litigated and necessarily determined in the first suit.

The doctrine of collateral estoppel traditionally had been limited to the parties to the past action. For instance, A, as the driver of B's truck, is involved in an accident with a car driven by C. If A sues C and recovers a judgment because of the negligence of C, the traditional rule has been that in a subsequent suit filed by B against C for damage to the truck, C is not precluded from claiming that he was not negligent since B was not a party to the first suit and would not be bound by the decision in it. Many courts now, however, are holding that even though the same parties are not involved, when the issues are the same and when the defendant has presented a complete and full defense in the first trial, collateral estoppel will now bind him to the finding in the first suit that he was negligent in the occurrence.

Res judicata in civil-law countries

The principle of res judicata is followed in civil-law countries as well, but there are differences. Substantively, res judicata applies generally only in new proceedings between the same parties (or their heirs or successors in interest), and the new proceedings must involve the same type of action (the same bases for the action and the same demand for relief). There is, however, no collateral estoppel, though a judgment that is no longer subject to any form of review (appeal, etc.) is binding as to all procedural rulings. In effect, res judicata becomes procedurally operative only after all normal means of review have been exhausted or the time limit to use them has lapsed.

Enforcement of judgment. All countries have procedures intended to overcome the resistance of a party who fails to comply with the judgment of a court. This is usually known as the enforcement or execution of a judgment. Rules vary greatly, and they are usually highly technical and thus can only be dealt with generally. In the

United States a party who obtains a judgment for a sum of money is entitled normally to avail himself at once of the procedural devices designed to enforce the judgment. The fact that the period for appeal has not yet passed or that an appeal is filed does not, of itself, affect the right to enforce the judgment; the losing party, seeking to postpone enforcement of the judgment pending appeal, must request such relief either from the trial court or the court to which the appeal is taken. Frequently, such a request will be granted if the losing party posts a bond or other security to ensure that the delay in enforcement will not adversely affect the rights of the winning party should the appellate court affirm the judgment of the trial court.

When the judgment results in an order to the losing party to do or refrain from doing some act, the court has the power to enforce the judgment by punishing a party who fails to comply, by a fine or a jail sentence, on the grounds that his disobedience constitutes "contempt of court."

When the judgment results in an award of money damages, the usual procedures for enforcement are the "levy of execution" on property belonging to the defendant or an execution against his income. All property that is not exempt by a specific statute, as well as income earned and debts owed by third persons, are subject to this enforcement process. Exemptions generally are given for such necessities as wearing apparel, tools and implements used in earning a living, and household furniture, and such personal items as wedding rings, family Bibles, and family photographs. The attorney for the party in whose favour the judgment has been rendered or the clerk of the court in which the judgment was obtained issues a command to the sheriff to seize the property, Once the sheriff has taken possession of the property he sells it at public auction and, after deducting his fees, turns over to the judgment creditor only those proceeds of the sale necessary to satisfy the judgment; any excess is returned to the defendant.

Garnishment of wages

The remedy of garnishing the earnings of the defendant, although generally permitted, is accompanied by certain safeguards to prevent oppression. Thus, only if the debtor fails to make payments voluntarily, can his wages be seized, and even then only a limited percentage of the wages.

Rules for the enforcement of judgments in civil-law countries are in some respects similar to those in the United States or other common-law countries, although some differences do exist. Frequently, judgments cannot be enforced by execution or in some other way until all appeals have been heard or until the time for such appeals has run out, but the precise rules differ greatly from country to country and often depend on the subject matter of the action or the court to which an appeal is taken. In Germany, for instance, it is sometimes possible to receive an execution on a judgment still subject to appeal, but the money recovered on execution must be paid into the court clerk's office pending determination of the appeal.

In all countries there are detailed rules exempting certain types of property from seizure, but continental European rules are much less generous toward the debtor than corresponding rules in the United States. In France, for instance, all wages exceeding the equivalent of about $3,000 per year may be seized, whereas in New York no more than 10 percent of wages may ever be taken. If several judgments, which threaten to exceed a debtor's available assets, exist against him, other procedures are available to insure that these assets will be distributed fairly. To some extent, such procedures replace bankruptcy, which in some European countries is available only to businessmen and not to private debtors.

Problems arise in connection with judgments ordering a party to do or not to do a certain act, since contempt procedures, outside of mild fines or jail sentences available to secure the maintenance of order in the courtroom, are generally unknown in Europe. For this reason, Italian judgments will order the performance of a specified act only when, in the case of disobedience by the party, the act can be performed by a substitute appointed by the court. For instance, if the defendant is ordered to tear down a wall and refuses to do so, the court may appoint a contractor to perform this operation. French courts have not limited themselves so narrowly and have developed a kind of civil penalty in order to compel compliance with their judgments.

Costs and disbursements. Generally, the prevailing party recovers not only the amount of the judgment but also the costs and expenses of the suit. These include filing fees, government taxes, witness fees, and the like. but not funds spent in the preparation of the case. In countries like Austria and Germany that regulate the fees of attorneys by an official schedule, such fees are ordinarily recoverable. In countries where such fees are not regulated by schedule, they usually must be borne by the party that has incurred them. In countries such as England or France, where a party is often represented by an agent for litigation (a solicitor, or *avoué*) and a separate attorney to handle oral argument and trial (a barrister, or *avocat*), the fees of the former, but not of the latter, are reimbursable in most situations.

APPEALS AND OTHER METHODS OF REVIEW

A judgment of a court of first instance may be attacked either by appeal to a higher court or by a request for some form of review of the judgment by the court that rendered it. Thus, it is quite generally possible for a defendant who has defaulted to ask a court to reopen the case and hear it on its merits. As noted above, in Anglo-American courts, it is frequently possible to ask for a new trial. In some cases (if, for example, there is newly discovered evidence) procedures analogous to motions for a new trial exist in European countries. In certain countries and in some states of the United States, an appeal of a judgment that is not a final decision can be made in addition to appeals of final decisions.

The appeal process is somewhat different in civil-law and common-law countries. In Europe the appeal from the court of first instance to the intermediate appellate court ordinarily involves a reexamination of the entire case, both the law and the facts, and new evidence frequently can be introduced. An appeal to the supreme or highest court is restricted to matters of law, and the facts found by the lower court are not re-examined. In the Anglo-American system, on the other hand, both the intermediate appellate court and the supreme court examine only the written record created in the court below and do not receive new evidence. Furthermore, review is generally restricted to matters of law, though the scope of review is broader in the intermediate appellate court than the supreme court. Rules of appeal in all systems tend to combine the desire that justice be done and error be corrected with the desire to find some point at which the proceedings will end and judgment will be deemed final.

Common-law appellate procedure. A fundamental principle underlying the function of appellate courts in the United States is the concept that the court serves only to review allegations that errors of law were committed at the trial. In no sense can the appeal be considered a retrial of the entire case. Factual determinations made at a jury trial are not reviewable on appeal except when presented in the context of a legal question. Factual determinations made by a judge in cases tried without a jury are reviewable on appeal, but even in such cases, appellate courts are reluctant to set aside such determinations unless clearly erroneous.

The party appealing the judgment must specify the errors that allegedly occurred at the trial; generally, the appellate court will consider only those points advanced by the appealing party. Moreover, the court will, with few exceptions, refuse to consider an allegation of error, unless the issue had been raised during the initial trial.

Because appellate courts do not hear witnesses or permit the introduction of new evidence on appeal, it is necessary that the record of the trial be made available and include a transcript of the proceedings, original papers, and exhibits. Both parties are required to submit written "briefs" to the court containing legal precedents and the arguments in support of their contentions that error did or did not occur, and each party has an opportunity to present oral legal arguments supporting his position.

Most jurisdictions provide a second appellate court to which a party may appeal from an adverse decision of the first appellate court. The right to such a second appeal is usually limited to certain types of cases raising particularly important issues, and only a small percentage of litigants pursue a second appeal. In the U.S. Supreme Court, a petition to authorize an appeal in certain cases involving the public interest, when it is not available as a matter of right, is known as a petition for a writ of certiorari.

Civil-law appellate procedure. Appeals to intermediate appellate courts from courts of first instance are available quite broadly in Europe, frequently for all judgments exceeding a certain amount (*e.g.*, 2,500 francs, or about $460 U.S., in France) and at times for certain types of judgments, regardless of amount. Since the appeal involves a new hearing of the case, the procedure is essentially similar to that in use in courts of first instance. In the case of a review of a nonfinal judgment, the appellate court frequently limits its review to an examination of the legal correctness of that judgment and then remands the case, so that proceedings in the court below may be completed. Occasionally, appellate courts are authorized to use the occasion of an appeal of a nonfinal judgment in order to decide the entire case themselves. Though an appeal involves a rehearing of the entire case and though parties are, generally speaking, entitled to introduce new evidence, the appeal may not be used to bring forth entirely new claims. The broad availability of a new hearing on appeal encourages appeals to intermediate appellate courts and explains their frequently very heavy case load.

By way of contrast, appeals to the supreme courts of the various countries are generally limited to questions of law. The facts are not ordinarily re-examined, and no new evidence may be introduced. The procedure involves essentially the presentation of written or oral argument by counsel for both sides on the alleged substantive or procedural errors made by the lower court. In several countries, such as France and Italy, the partisan argument by the parties is augmented by independent argument by an officer of the Ministry of Justice representing the law as such. The Court either affirms or reverses the judgment submitted to it for review. If it reverses, it does not, generally, substitute its own judgment for the erroneous judgment below but merely annuls the erroneous judgment and remands the case for new proceedings, frequently to a court different from that from which the case came. Review by supreme courts can usually be sought for all final (and sometimes even nonfinal) decisions of intermediate appellate courts, and frequently also of decisions of courts of first instance if no appeal to an intermediate appellate court is possible. No special permission of the court analogous to the grant of certiorari is ordinarily required. Consequently, case loads are extremely heavy, and to handle them the full court does not usually sit together but instead is divided into panels. In important matters two or more panels may sit together. The court to which the case is remanded is not bound by the view of the law expressed by the appellate court.

BIBLIOGRAPHY

General works: H.J. ABRAHAM, *The Judicial Process,* 2nd ed. rev. and enl. (1968), an introductory analysis of the courts of the United States, England, and France, with emphasis on process rather than procedure; C.E. CLARK, *Procedure: The Handmaid of Justice* (1965), a collection of significant essays; R. MOLEY and S.C. WALLACE (eds.), *The Administratron of Justice* (1933), a study from a political science viewpoint; A.T. VANDERBILT, *Minimum Standards of Judicial Administration* (1949), a classic on the topic.

Historical growth of procedural law: LEOPOLD WENGER, *Institutionen des romischen Zivilprozessrechts* (1925; Eng. trans., *Institutes of the Roman Law of Civil Procedure,* rev. ed., 1940), the classic on the topic; J.P. DAWSON, *A History of Lay Judges* (1960), in disagreement with some of Wenger's conclusions, although not limited to Roman law; A. ENGELMANN *et al., Der Civilprozess, Geschichte und System,* 3 vol. (1889–1901; Eng. trans. of vol. 2, *A History of Continental Civil Procedure,* 1927, reprinted 1968), the classic text in English; W.S. HOLDSWORTH, *History of English Law,* 7th ed. 16 vol. (1956–66), and T.F.T. PLUCKNETT, *A Concise History of the Common Law,* 5th ed. (1956), general treatments of English legal history; F.W. MAITLAND, *The Forms of Actions at Common Law* (1937, reprinted 1962), a minor classic; J.H. KOFFLER and A. REPPY, *Handbook of Common Law Pleading* (1969), a textbook on common-law pleading.

Elements of procedure: Introductory treatments on American law include: W.W. BLUME, *American Civil Procedure* (1955), a discussion of ordinary actions; J.S. BRADWAY, *The History of A Lawsuit* (1958), *a* basic introduction to civil actions for students; HARVARD LAW REVIEW, *Essays on Civil Procedure* (1961), a collection of *a* number of valuable essays in the field. Introductory treatments on other countries include: M. CAPPELLETTI and J.M. PERILLO, *Civil Procedure in Italy* (1965), one of the best available treatments in English; and R.B. GINSBURG and A. BRUZELIUS, *Civil Procedure in Sweden* (1965). On the jury system, see: C.W. JOINER, *Civil Justice and the Jury* (1962), *an* examination of the effectiveness of the jury and related problems; R.W. MILLAR, *Civil Procedure of the Trial Court in Historical Perspective* (1952), a discussion of Anglo-American trial procedure from a historical point of view. For declaratory judgments, see: E.M. BORCHARD, *Declaratory Judgments,* 2nd ed. rev. and enl. (1941), the classic work; and W.H. ANDERSON, C. BOWEN, and G.C. ANDERSON, *Actions for Declaratory Judgments* (1940). On appeals and other methods of review, see: D. KARLEN, *Appellate Courts in the United States and England* (1963), a comparative study.

(P.E.H./Ma.E.O.)

Procellariiformes

The Procellariiformes are a distinct natural order of oceanic birds with about 87 living and 36 fossil species. Of diverse size and range, they are divided into four families: albatrosses, sheanvaters, storm petrels, and diving petrels. All are recognizable by their conspicuous tubular nostrils, which project upon the culmen (upper bill), giving the order its alternative name, Tubinares, "tube nosed." The feet are webbed, and the hind toe is vestigial or missing. All species have a characteristic powerful musky odour caused by the excretion of stomach oil; the oil can be used as a defensive discharge through the mouth when the bird is alarmed.

GENERAL FEATURES

Importance to man. The tube-nosed birds have been of considerable local economic importance as a source of protein food, feathers, and oil wherever man has colonized or has been able to raid the coastal and oceanic islands where they breed; this has resulted in the partial or complete extermination locally of certain species. Man has also been responsible for the introduction of predators, such as rats, pigs, and cats. In regions where bird populations have survived, man has continued to harvest the eggs, the plump young birds (at fledging time), or both. Many thousands of slender-billed, or short-tailed, shearwaters (*Puffinus tenuirostris)* are taken on the Bass Strait islands off Tasmania and sold fresh, salted, or deep-frozen as "muttonbirds." The name muttonbird was most likely derived from the use of the flesh as a supplement for mutton by the early settlers of New South Wales. The numbers of muttonbirds now harvested are regulated so as to preserve a substantial breeding stock.

"Mutton-birds"

In New Zealand the Maori people have harvested young *titi* (sheanvaters of several species) from time immemorial, a right assured them in perpetuity by treaty with Queen Victoria. On the other side of the world, hundreds of Manx shearwaters (*Puffinus puffinus*) were formerly collected for food and as lobster bait on the Welsh islands of Skomer and Skokholm, which are now nature preserves estimated to contain about 200,000 Manx sheanvaters and 2,000 storm petrels *(Hydrobates pelagicus)*. On the Tristan da Cunha Islands in the South Atlantic, resident islanders harvest the eggs and squab (young) of a large, mixed seabird population, which includes more than 6,000,000 greater sheanvaters (*Puffinus gravis*).

The harvesting of northern fulmar petrels (*Fulmarus glacialis*) is an ancient practice among peoples of the cool northern coasts where the birds breed. In Iceland about 50,000 fulmars were taken annually between 1897 and 1925; the occurrence in 1939 of psittacosis (a virulent avian disease) among processors of the birds resulted in prohibition of the use of fulmars for food.

During the early 17th-century colonization of Bermuda, millions of cahows, or Bermuda petrels (Pterodroma *cahow*), were exterminated by the colonists. For nearly 300 years the species was believed extinct, but in 1951 a few pairs were discovered nesting on an offshore islet, where a remnant survives under strict protection. The related black-capped petrel, or diablotin (P. hasitata), of the West Indies was also thought extinct (due to predation by man, rats, and mongooses) until in 1961 a substantial population, estimated to number at least 4,000 birds, was found breeding in the inaccessible forested crags of Hispaniola.

In the 18th and 19th centuries, huge numbers of albatrosses were taken for food (largely by whalers) and for the millinery trade. With the disappearance of sailing vessels, changes in fashions, and the establishment of many nesting grounds as sanctuaries, such predation has virtually disappeared, but albatrosses have not entirely escaped stress at the hands of man. On Sand Island in the Midway Atoll, Laysan albatrosses (Diomedea *immutabilis)* increased from a few pairs in 1900 to about 60,000 pairs in the early 1960s, the increase resulting from shelter provided by introduced vegetation. The use of the island by aircraft after 1935, air raids by the Japanese during World War II, and the loss of 30,000 birds humanely killed by the U.S. Navy (in a control program designed to reduce collisions between birds and aircraft) did not deter the albatrosses from this favoured nesting area. The control program was abandoned after the discovery that levelling certain sand dunes effected a 70 percent reduction in collisions by removing the updrafts near aircraft runways.

Distribution. The majority of procellariiforms breed in the Southern Hemisphere, but several species migrate thousands of miles north across the Equator to winter in the northern summer seas, where they molt, feed, and rest in preparation for the return home in the southern spring. Similarly, species that breed in the Northern Hemisphere also live in perpetual summer by migrating far south to winter in the southern summer. A number are less migratory and do not cross the Equator. Several species are almost sedentary, chiefly smaller petrels breeding in tropical and subtropical latitudes, and the sub-Antarctic prions (Pachyptila) and diving petrels. All latitudes of the unfrozen oceans are thus occupied, but there are fewer living in the calm equatorial region where there is little wind to lift their long wings and where the pelagic (open-ocean) crustacean food on which so many seabirds basically depend is scarce. The zone of upwelling water in the windy latitudes of the Antarctic covergence, between 40" and 60" south latitude, is richest in the shrimplike krill (Euphausia species), attracting surface-feeding tubinares and the diving penguins, prions, and diving petrels. Some feed along the edge of the ice of the Antarctic continent, and four tubinares actually breed on its shores: the Antarctic fulmar, the giant petrel, the snowy petrel, and the tiny but very numerous Wilson's petrel; the nesting burrows of the last may be blocked by snow for days during the protracted breeding season. The only tubinare nesting near the ice limits in the high Arctic is the fulmar, which reaches Franz Josef Land, Greenland, and the Arctic Circle north of the Aleutian Islands.

Drawing by R. Keane

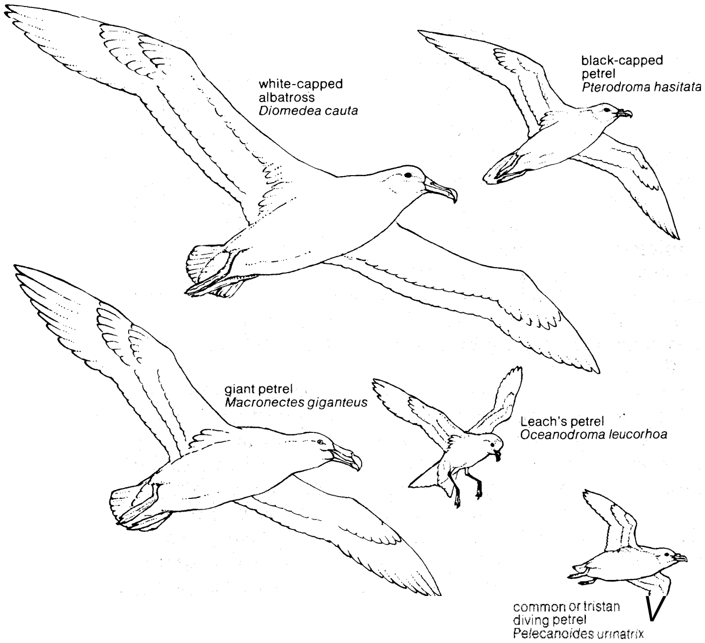


Figure 1: Representative procellartiform birds in flight.

Of the albatrosses (family Diomedeidae), only the two Midway species and the short-tailed albatross (Diomedea *albatrus)* nest well north of the equatorial doldrums. The latter was brought close to extinction by plume hunters and by a volcanic eruption at its nesting island of Torishima. There were enough immature birds at sea at the time to allow a partial recovery, and some 60 individuals were counted there in 1969. Nine albatross species range the Southern Hemisphere, gliding on the eternal winds of the "Roaring Forties" (the region between 40° and 50° latitude) and moving north with the food-rich cold currents along the west coasts of South America, South Africa, Australia, and New Zealand. One species, the waved albatross (Diomedea irrorata), is unique in that it breeds only in the Galápagos Archipelago at the Equator, where probably not more than 3,000 pairs nest on Hood Island.

The family Procellariidae includes the larger petrels, such as the northern and southern fulmars (*Fulmarus* glacialis and F. glacialoides), the gadfly petrels (*Pterodroma)*, several genera of shearwaters, and the prions or whalebirds. Several of the shearwaters and larger petrels breed in burrows far inland on mountain crags in the Andes, West Indies, Madeira, and New Zealand. The largest member of this family is the giant, or stinker, petrel (Macronectes giganteus), an albatross-like scavenger and circumpolar wanderer with a heavy beak and wing span of eight feet. Smallest are the prions, four species of small, stocky, little-studied birds, 22 to 30 centimetres (nine to 12 inches) long, with broad bills and a restricted cold-water range; they breed on sub-Antarctic islands, keeping much to the water, as do the diving petrels.

The storm-petrel family, Hydrobatidae, ranges both hemispheres but is strongest numerically in the Pacific, where Halocyptena microsoma, the least petrel of Baja California, rivals the Atlantic storm petrel as the smallest procellariiform. The word petrel ("little Peter") derives from a habit of the storm petrels of walking on the waves.

The diving petrels form a family (Pelecanoididae) and genus (Pelecanoides) with four species. They are small, rather sedentary, coast-dwelling birds confined to cool southern islands, including Tristan da Cunha, the Falklands, New Zealand, and southeastern Australia. The common diving petrel (P. urinatrix) is circumpolar; the largest species, the Peruvian (P. garnotii), has followed the Humboldt current along the west coast of South America and breeds from Chile to Peru; the Magellanic (P. *magellani*) is confined to the tip of South America; the Georgian (P. georgicus) breeds on South Georgia, Macquarie, and Auckland islands. Diving petrels are specialized birds with short, black-and-white bodies and closely resemble externally and in habits the small auks (family Alcidae, order Charadriiformes) of the Northem Hemisphere.

NATURAL HISTORY

Locomotion. All tube-nosed birds have a protracted life cycle conditioned by their evolution and oceanic environment. Because they spend most of their lives at sea, they are clumsy on land, laboriously using their wings as props to assist locomotion; their legs are too far to the rear to effect a well-balanced bipedal progress. The smaller species nest in burrows and rock crevices and are nocturnal, being helpless and unable to manoeuvre quickly on land when attacked by predators. As a rule, the incubating bird is tame and does not associate the approach of man with danger but will often allow him to stroke and fondle it. Albatrosses are especially docile, hence the name mollymawk (mallemuck), from the Dutch *mollemok* ("stupid gull").

Flight pattern of albatrosses

The long-winged tubinares require a smooth runway for takeoff on a calm day; over rough ground they will utilize the bill to hook along, and either climb a rock or tree to gain a launching height or flop over the edge of the nearest cliff. On the wing, they are perfect aviators, riding out the severest storms of their normally windy oceanic feeding grounds with ease and grace. The great albatrosses can overtake and circle a fast ship at sea, with long glides rarely interrupted by wingbeats. The ability of albatrosses to move upwind without flapping depends on the fact that wind velocity is appreciably lower near the waves than a few metres in the air. The flight pattern is a series of broad ellipses (Figure 2), characteristically including (1) a fast downwind glide beginning at the greater elevation, at which the wind velocity is higher, giving the bird considerable momentum with a small loss of altitude. As it reaches maximum speed, the bird turns crosswind (2), briefly skimming the waves, then moves upwind (3) on a long, level glide through the lower velocity air, maintaining altitude, but losing speed. As soon as its air speed reaches a critical low, the bird climbs steeply cross-wind (4) to complete the ellipse, now at the same altitude as at the start of the previous downwind leg but having gained in position upwind. It then glides downwind again to gather fresh momentum. The same flight pattern may be used, of course, to travel crosswind or downwind. The normal air speed of the royal and wandering albatrosses (*Diomedea* epornophora and *D. exulans),* whose wing spans reach about 3.4 metres (11 feet), is 80 to 110 kilometres (50 to 70 miles) per hour. Although the flight appears effortless, some energy is expended in the muscular action that keeps the long, narrow wings fully extended.

Drawing by R. Keane

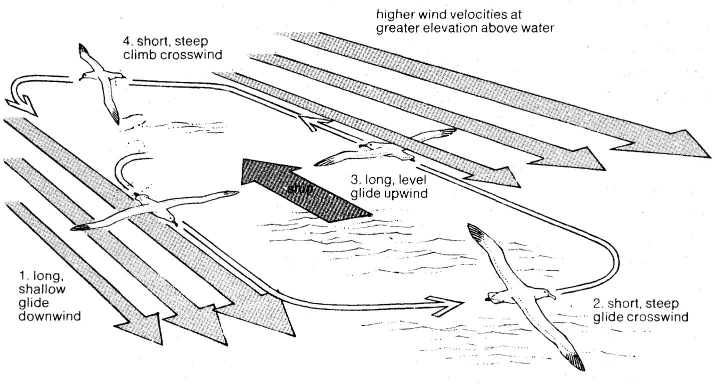

higher wind velocities at greater elevation above water
4. short, steep climb crosswind
ship
3. long, level glide upwind
1. long, shallow glide downwind
2. short, steep glide crosswind

Figure 2: Flight pattern of the albatross.

The medium-sized tubinares (shearwaters and procellariid petrels) have a flight pattern similar to that of albatrosses, but their shorter wings are flapped regularly between the briefer gliding periods. The little storm petrels have an altogether more erratic darting, fluttering, and sometimes hovering flight, their feet hanging down to walk on the water surface.

Feeding habits. Shearwaters, storm petrels, and diving petrels feed by taking small fish and crustaceans close to the surface; they make short dives as necessary. Many of the larger procellariids consume substantial amounts of squid. Albatrosses, giant petrels, and fulmars dive little; they are surface feeders, often settling on the water. At night they devour squid that rise to the surface; during the day they take schooling fish; garbage from ships; wounded, exhausted, or dead birds; and carrion, including the flesh of dead whales and other cetaceans. The giant petrel is probably the only tubinare agile enough on land to kill other birds; at its nesting grounds it will attack young penguins inadequately guarded by their parents.

Reproduction and growth. As a general rule, the mature adults return to the established breeding site many weeks before the single white egg is laid. There is often severe competition for nesting territories in large crowded colonies on small islands. Returning each year to the same nest site, the male and female remain faithful to it and thus to each other for life. It is believed that some albatross pairs also remain together at sea in the nonbreeding season, but many of the burrowing shearwaters and petrels, meeting on land only at night, may never see their mates clearly when ashore (recognition being by voice, touch, and possibly smell) and probably do not deliberately consort in pairs at sea.

The greeting ceremony

At each fresh encounter ashore between breeding birds, there is an elaborate greeting ceremony; the birds clash and fence with the bills, cackling and screaming. These antics occur in both the nocturnal procellariids and the diurnal albatrosses, and in the latter there is also a bowing and dancing display. Such behaviour provides time for mate recognition and relieves and displaces any natural aggression or fear.

The nest type varies somewhat among species. Albatrosses scrape a shallow depression or build a mound of soil and vegetation; the fulmars and other diurnal procellariid petrels nest on ledges or on level ground; most shearwaters, diving petrels, and some storm petrels dig burrows in soft soil; other storm petrels utilize natural crevices.

Once the nest site has been adopted, one member of the pair usually remains on guard against usurpation by other home-hunting birds. The male may remain on guard for several days and nights, while the female feeds at sea to meet the requirements of the developing egg. In some species, the female may depart on a recuperative feeding cruise within a few hours after laying if her mate is there to take over incubation. Mated birds do not feed each other but incubate in spells of several days each, the bird at home fasting and losing weight while the bird at sea is feasting and fattening.

The egg is incubated for a long period, about 80 days in the wandering albatross, 52 days in the Manx shearwater, 40 days in the smallest petrels (the last about equal to the incubation period of the ostrich). For the first week or so after hatching, the helpless, downy chick requires the warmth of the parental body for survival. During this period it is brooded and fed tenderly on an oily broth of semidigested marine organisms pumped from the adult esophagus, which is muscularly constricted to control the flow to the infant's needs. Instinctively, the chick seeks the open, warm, fishy-smelling mouth of the parent, thrusting and groping blindly with its tiny bill crosswise in the open maw of the adult.

The down grows rapidly; a second down sprouts (to which the first remains attached), and the baby is soon homeothermic—*i.e.*, able to keep itself warm while its parents forage far at sea and return with increasingly large crop loads of food. Some of the fishing trips of the parents are real voyages, involving absences of several days; the Manx shearwater may travel nearly 1,000 kilometres (more than 600 miles) from Wales to the

Oiin S. Pettingill, Jr.—National Audubon Society

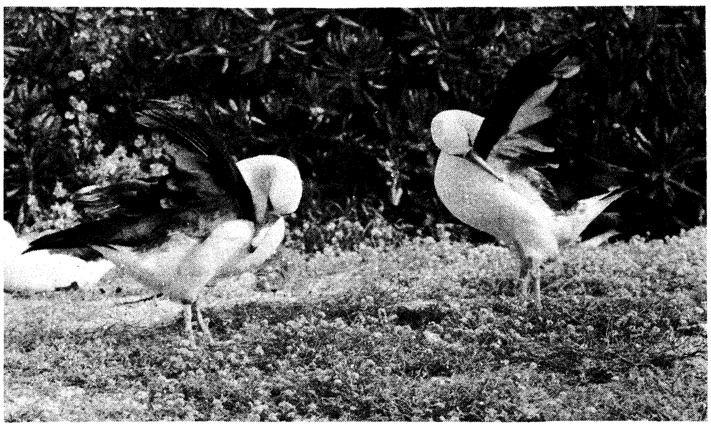


Figure 3: A pair of Laysan albatrosses (*Diomeda immutabilis*) in ritualized preening display, a part of the courtship that precedes nest building.

Period of fledgling starvation

Bay of Biscay and back to load up with its favourite food, sardines. Albatrosses may leave their well-developed nestling for a week or two. If both parents happen to return at the same time, the nestling may ingest food equal to its own weight in one meal. It becomes very fat in the later stages of the long fledgling period, which is not less than two months in the small petrels and reaches nine months in the largest albatrosses. Before it leaves the nest, the chick is deserted by the parents, who retire to molt at sea. This begins a starvation period which may last a week in the smallest petrels, twelve days in the medium-sized shearwaters, and considerably longer in the largest species, before the fledgling goes to sea. When deserted, it is well-feathered and fatter and heavier than the adult; it needs a period of thinning and exercise before it is capable of flight. After days of fasting and wing flapping, it may become airborne one windy night, especially if hatched in a burrow on a gale-swept mountain height from which it can flap and glide to the sea. Calm weather is its enemy; many island-born young tumble down to the sea, too heavy to take off again in still air. They are expert swimmers, however, and can dive deeply to avoid attacks by aerial predators.

Paddling rapidly away from the dangers of land and soon gaining flight, the young procellariiform sets off along the traditional migration route, alone and unguided by the long-departed adults. Driven by an innate impulse to keep flying, it reaches winter quarters that it has never seen before, often at a surprising speed. One Manx shearwater, banded in Wales as a fledgling, travelled 9,900 kilometres (6,200 miles) to southern Brazil in 16% days. Allowing half of each day for resting and feeding, this is equivalent to an average surface speed of 50 kilometres (30 miles) per hour over the period, a remarkable achievement for a bird just out of the nest.

The young albatross remains longest in the nest—so long in the case of the royal and wandering albatrosses that the nestling is overtaken by the Antarctic winter. It endures blizzards and savage winds that force it to grip the nesting mound tightly with its claws, yet it is warm enough under its oily plumage to survive fasting for many wintry days until a parent appears with food. Because of the protracted nesting period, these great albatrosses cannot rear more than one young every other year. To compensate for the slow rate of reproduction, albatrosses are long-lived; life expectancy, once breeding age is reached, appears to be several decades. Marking of Laysan albatrosses has shown that they do not breed successfully until seven years old. In order to maintain their numbers, the wandering and royal albatrosses, breeding for the first time even later, must have the highest average longevity among birds. The Manx and other medium-sized shearwaters lay the first egg when five years old, the least petrel in the third or fourth summer.

There is thus always a large proportion of each tubinare population that is nonbreeding. During its first year at sea, the young bird may not even approach the land. While the mature birds have completed their migration and settled to breed at home, the yearlings may lag far behind on the route, spending the summer at sea. In the next few years adolescents arrive at the breeding islands and shores too late to make more than a preliminary landing and exploration of the ground for a future partnership. At midsummer in the nesting colonies there is a considerable arrival of immature birds familiarizing themselves with prospective breeding territories. Where a colony is already overcrowded, it is always the young, eligible, but inexperienced, birds that leave to form new peripheral colonies in the region of their birthplace.

Ecology and conservation. The tubinares collectively occupy a position midway in the oceanic food chain. Their food is lower on the chain than that of animals which take subsurface fishes (*e.g.*, certain diving birds, seals, and medium-sized predatory fishes) but above that of the minute and larval fishes which feed on plankton. Insofar as man has overfished much of the oceans and removed or reduced competitors, such as the krill-feeding whales, and has taken fish species from the food chain, he might be said to have influenced the success and therefore the numbers of the tubinares, but there is no firm evidence for this. Although the numbers of some species are recorded as increasing, it is likely that the reason is stricter protection and reduction of exploitation at the nesting sites. It is generally accepted that the considerable increase of the fulmar in the North Atlantic and its spread south from the Arctic to new breeding grounds has resulted from the great development of fishing fleets, which fulmars follow in vast numbers to devour the fish offal thrown overboard.

Man's effect on the tubinare population

Only a few tubinares are rare and in danger of extinction. The majority are successful and often numerous. With their pelagic, aerial habits they are less liable to be trapped in oil slicks than are the Northern Hemisphere swimming seabirds (auks, ducks, and loons), which are killed by the thousands by oil. The diving petrels and prions, living in the cleaner Antarctic seas, are less exposed to this danger.

FORM AND FUNCTION

General features. The general body plan of procellariiform birds varies slightly from family to family. In general, they are long-winged, short-necked birds with short to moderate tails and legs. Webbing is present between the front toes, and the hindtoe (hallux) is small or lacking. In contrast to their strong-flying relatives, the diving petrels have short wings. At the other extreme, the aspect ratio (the ratio of wing span to the chord, or width) of the wing may exceed 14:1 in some albatrosses. This long, narrow wing, with a high-lift airfoil, is an extreme adaptation for fixed-wing gliding.

The bill varies from rather short and broad in diving petrels to medium in length (somewhat more than half the total length of the head) in some albatrosses. It is sheathed in horny plates and has a distinct hooked nail at the tip. In albatrosses the two nasal tubes lie separated on the right and left upper lateral surfaces of the bill; in all other procellariiforms, the nostrils are fused into a single tube lying on the dorsal midline of the bill and having a dividing wall or septum, which may end short of the end of the tube, resulting in a single opening.

Procellariiforms are totally lacking in bright plumage colours, being entirely black, white, or shades of brown or gray. Strikingly contrasting patterns of light and dark are often found, however, and the bills or feet of a few species are yellow or pink. A number of shearwaters and procellariid petrels and a few albatrosses are polymorphic; *i.e.*, they occur in light and dark phases (plumage types), some species also having intermediate forms. The polymorphism may be restricted to certain parts of

Drawing by R. Keane

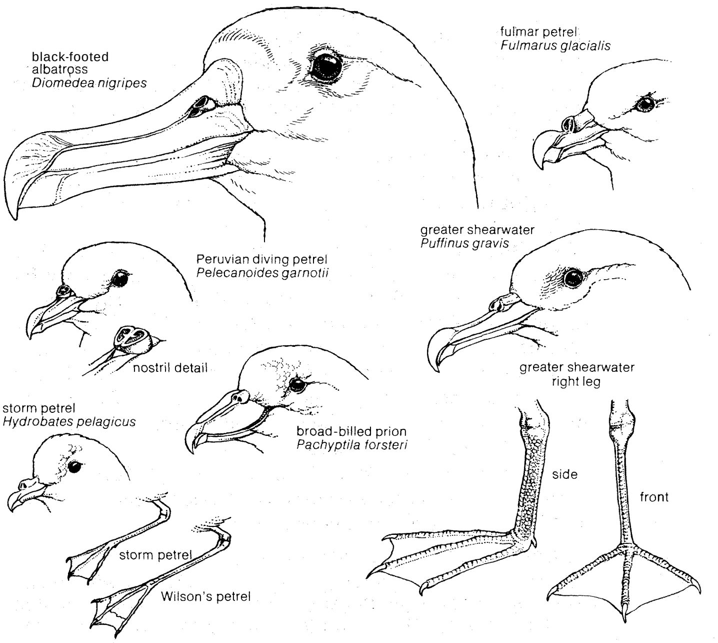


Figure 4: Heads and feet of representative procellariiforms.

the plumage, such as the underparts of the body or the upper surface of the wings.

Squirting oil as a defense mechanism

Stomach oil. Most tubinares, when handled or threatened, eject the oily contents of the stomach with some force. In some species, notably the cliff-nesting fulmars, this habit, a fear reaction that also serves to lighten the bird for flight, has been exploited as a defensive weapon. Facing an intruder, the disturbed bird ejects a spurt of evil-smelling fluid a metre or so in his direction, often with apparently planned accuracy. The habit is instinctive; a baby fulmar, on hatching, has been observed to squirt yellow oil before it is fully out of the shell. Later, the downy chick squirts oil at any visitor, even its parents. Mated fulmars may exchange little squirts of oil during the excitement of bill-fencing ceremonies.

Analysis of this unique oil shows that it is a waxy secretion of the proventriculus (the first chamber of the stomach), rich in vitamins A and D. In most birds, the walls of the proventriculus produce an acid fluid that rapidly breaks down raw food entering from the esophagus. In the tubinares, which feed their young a soup of predigested marine organisms, the proventriculus is much enlarged and internally folded, increasing the surface when dilated and enabling a larger number of glands to function. The latter are groups, or follicles, of oil-producing cells. The colour of the oil varies according to the type of food; it is often reddish from the presence of astacin, a pigment found in crustaceans.

The discharge of stomach oil is partly excretion of surplus fat, which might upset the bird's metabolism if retained in quantity. Ejected through the mouth and nose, it also disposes of excess vitamins and salt in the diet of marine food and seawater. Similar in character to the secretions of the oil glands of other birds, the crop oil may also assist in waterproofing the feathers as the tubinare preens its plumage with its oil-stained bill.

EVOLUTION AND CLASSIFICATION

Relationship to penguins

Evolution. The oldest tubinare fossil is a giant albatross *(Gigantornis)* from the Eocene of Nigeria (about 50,000,000 years ago). It may have had a wing span of six metres (20 feet) and was contemporary with the now extinct giant penguins (order Sphenisciformes). It is generally agreed that the two orders had a common ancestor from which they may have evolved. The penguins occupied the ecological niche of diving and feeding under the surface and became flightless; the tube noses specialized in flight and surface feeding. Support for a common origin (a view contested by some taxonomists) comes from the facts that the oldest fossil penguin had a bill with distinct tube-nosed apertures; the young of the blue penguin *Eudyptula,* considered to be the most primitive of penguins living today, exhibits tubelike openings to its nostrils. Mutual displays of bill fencing and wing movements in courtship, as well as the method of regurgitating digested food, are almost identical in tubinare and penguin. The short-winged diving petrels, which "fly" much underwater but little in air, seem to parallel an early stage in the evolution of penguins, especially when, during a few weeks of the annual molt, they lose their quills and must live in the water.

Classification. *Distinguishing taxonomic features.* The families of the Procellariiformes are separated mainly by the general body plan, the condition of the nostrils, and, in the case of the Procellariidae and Hydrobatidae (long considered one family), the osteology of the skull and sternum. At the genus level, characters used include the shape of the beak, wings, and tail; the degree of flattening of the tarsus (lower leg); size of the hallux; and the relative lengths of the leg bones.

Annotated classification. The following classification, proposed by American ornithologist Alexander Wetmore in 1930, is in nearly universal usage.

ORDER PROCELLARIIFORMES (tubinares)

Oceanic birds with tubular nostrils; bill covered with horny plates and hooked at the tip. Anterior toes webbed; hallux short or lacking. Wing with 11 primary feathers (the outer minute); secondaries short; diastataxic (*i.e.,* with the 5th secondary absent). Two coats of nestling down. Oil gland feathered. Strong musky smell. Single white egg; long incubation and nestling periods. About 87 living and 36 known fossil species; all marine; worldwide.

Family Diomedeidae (albatrosses)

Middle Eocene to present. Extremely long, narrow wings; short tail. Bill longer than remainder of head; nostrils semitubular, small, situated near the base of long groove. Two genera, 13 species; length 50–125 cm (20–50 in.); wing span to 340 cm (11 ft); North Pacific and all southern oceans.

Family Procellariidae
(large petrels, fulmars, prions, shearwaters)

Middle Oligocene to present. Long-winged, short-tailed. Nostrils united on top of bill. Twelve genera, about 50 species; length 22–75 cm (9–30 in.); all oceans, but greatest diversity in southern hemisphere.

Family Hydrobatidae (storm petrels)

Upper Miocene to present. Small black and brown birds, usually with conspicuous white rump; wings rounded; tail square or forked. Often walk on water. Eight genera, 20 species; length 15–20 cm (6–8 in.); all oceans, but more species breeding in Southern Hemisphere.

Family Pelecanoididae (diving petrels)

Upper Pleistocene to present. Small stocky birds, with short wings and tails. Black above; white below. One genus, four species; length 1620 cm (6%–8 in.); cool sub-Antarctic seas.

Critical appraisal. In a system of evolutionary relationships by structural affinities, the tube-nosed birds seem to fit conveniently between the penguins and the pelecaniform birds. But within the order, the arrangement of genera, species, and races of tubinares has been changed all too frequently by taxonomists. This disorder has resulted partly from new information on the structure, habits, and distribution of these birds, hitherto little studied because of their remote oceanic breeding grounds. The ordinal position occupied by the highly specialized prions and diving petrels is still in dispute. The fulmars are sometimes divided into three species because of their distinct differences in bill, body size and shape, and geographical range; but other taxonomists regard them as Arctic, Pacific, and Antarctic races of the species *Fulmarus glacialis.* Several authorities have recently placed several medium-sized shearwaters in one species *Puffinus puffinus,* with eight closely related forms or subspecies. Because there is no evidence of interbreeding between these populations, although they may mingle at sea on migration, and there are no intermediate forms even when they inhabit the same island (in which case, they nest at different times), other workers prefer to treat them as full species of *Puffinus.*

BIBLIOGRAPHY

General works: W.B. ALEXANDER, *Birds of the Ocean,* 3rd ed. (1955), still a standard identification handbook; A.C. BENT, *Life Histories of North American Petrels and Pelicans and Their Allies* (1922); R.A. FALLA *et al., Birds of New Zealand* (1966), deals with some 56 tubinare species; J. FISHER and R.M. LOCKLEY, *Sea-birds* (1954), an introduction to the natural history of the sea-birds of the North Atlantic; R.M. LOCKLEY, *The Island* (1969), includes life-histories of Manx shearwaters and storm petrels, and homing experiments; *Man Against Nature* (1970), descriptions of New Zealand tubinares and muttonbirding; R.C. MURPHY, *Oceanic Birds of South America,* 2 vol. (1936), also many original papers by this noted field worker; R.S. PALMER (ed.), *Handbook of North American Birds,* vol. 1 (1962), contains much information about the species recorded from North America; V. SERVENTY, *A Continent in Danger* (1966), a description of muttonbirding and banding of shearwaters; H.F. WITHERBY (ed.), *The Handbook of British Birds,* vol. 4 (1949), detailed characters and life histories of tubinares recorded in the British Isles.

Monographs: W.B. ALEXANDER *et al.,* "The Families and Genera of the Petrels and Their Names," *Ibis,* 107:401–405 (1965), taxonomical reorganization of the order Procellariiformes; P.E. DAVIS, "The Breeding of the Storm Petrel," *Br. Birds,* 50:85–101, 371–84 (1957), follow-up of an earlier study by Lockley; J. FISHER, *The Fulmar* (1952), a classic monograph; W.S. JAMESON, *The Wandering Albatross* (1958), a readable account of the biology of the largest albatross; R.M. LOCKLEY, *Shearwaters* (1961), an edition of a classic monograph first published in 1942; R.C. MURPHY and L.S. MOWBRAY, "New Light on the Cahow, *Pterodroma cahow,*" *Auk,* 68:266–80 (1951), rediscovery of the cahow at Bermuda; B. NELSON, *Galapagos: Island of Birds* (1968), contains

an extensive section on the waved albatross; M.K. ROWAN, "The Greater Shearwater *Puffinus gravis* at Its Breeding Grounds," *Ibis*, 94:97–121 (1952); J. WARHAM, "The Biology of the Giant Petrel *Macronectes giganteus*," *Auk*, 79:139–60 (1962); D.B. WINGATE, "Discovery of Breeding Black-Capped Petrels on Hispaniola," Auk, 81:147–59 (1964).

(R.M.L.)

Production, Theory of

In economics, the theory of production is an effort to explain the principles by which a business firm decides how much of each commodity that it sells (its "outputs" or "products") it will produce, and how much of each kind of labour, raw material, fixed capital good, etc., that it employs (its "inputs" or "factors of production") it will use. The theory involves some of the most fundamental principles of economics. These include the relationship between the prices of commodities and the prices (or wages or rents) of the productive factors used to produce them and also the relationships between the prices of commodities and productive factors, on the one hand, and the quantities of these commodities and productive factors that are produced or used, on the other.

Three kinds of business decisions

The various decisions a business enterprise makes about its productive activities can be classified into three layers of increasing complexity. The first layer includes decisions about methods of producing a given quantity of the output in a plant of given size and equipment. It involves the problem of what is called short-run cost minimization. The second layer, including the determination of the most profitable quantities of products to produce in any given plant, deals with what is called short-run profit maximization. The third layer, concerning the determination of the most profitable size and equipment of plant, relates to what is called long-run profit maximization.

The article covers the main principles involved in each of these problems in turn. For the sake of clarity, the essential principles will be considered within the context of a firm that produces only a single product or commodity, even though such firms in actuality are rare.

COSTS IN THE SHORT RUN

The relationship between input and output

The production function. However much of a commodity a business firm produces, it endeavours to produce it as cheaply as possible. Taking the quality of the product and the prices of the productive factors as given, which is the usual situation, the firm's task is to determine the cheapest combination of factors of production that can produce the desired output. This task is best understood in terms of what is called the production function; *i.e.*, an equation that expresses the relationship between the quantities of factors employed and the amount of product obtained. It states the amount of product that can be obtained from each and every combination of factors. This relationship can be written mathematically as $y = f(x_1, x_2, \ldots, x_n; k_1, k_2, \ldots, k_m)$. Here, y denotes the quantity of output. The firm is presumed to use n variable factors of production; that is, factors like hourly paid production workers and raw materials, the quantities of which can be increased or decreased. In the formula the quantity of the first variable factor is denoted by x_1 and so on. The firm is also presumed to use m fixed factors, or factors like fixed machinery, salaried staff, etc., the quantities of which cannot be varied readily or habitually. The available quantity of the first fixed factor is indicated in the formal by k_1 and so on. The entire formula expresses the amount of output that results when specified quantities of factors are employed. It must be noted that though the quantities of the factors determine the quantity of output, the reverse is not true, and as a general rule there will be many combinations of productive factors that could be used to produce the same output. Finding the cheapest of these is the problem of cost minimization.

The cost of production is simply the sum of the costs of all of the various factors. It can be written:

$$C = p_1x_1 + \ldots + p_nx_n + r_1k_1 + \ldots + r_nk_n,$$

in which p_1 denotes the price of a unit of the first variable factor, r_1 denotes the annual cost of owning and maintaining the first fixed factor, and so on. Here again one group of terms, the first, covers variable cost (roughly "direct costs" in accounting terminology), which can be changed readily; another group, the second, covers fixed cost (accountants' "overhead costs"), which includes items not easily varied. The discussion will deal first with variable cost.

The principles involved in selecting the cheapest combination of variable factors can be seen in terms of a simple example. If a firm manufactures gold necklace chains in such a way that there are only two variable factors: labour (specifically, goldsmith-hours) and gold wire; the production function for such a firm will be $y = f(x_1, x_2; k)$, in which the symbol k is included simply as a reminder that the number of chains producible by x_1 feet of gold wire and x_2 goldsmith-hours depends on the amount of machinery and other fixed capital available. Since there are only two variable factors, this production function can be portrayed graphically in a figure known as an isoquant diagram (Figure 1). In the graph, goldsmith-hours per month are plotted horizontally and the number of feet of gold wire used per month vertically. Each of the curved lines, called an isoquant, will then represent a certain number of necklace chains produced. The data displayed show that 100 goldsmith-hours plus 900 feet of gold wire can produce 200 necklace chains. But there are other combinations of variable inputs that could also produce 200 necklace chains per month. If the goldsmiths work more carefully and slowly, they can produce 200 chains from 850 feet of wire; but to produce so many chains more goldsmith-hours will be required, perhaps **130.** The isoquant labelled "200" shows all the combinations of the variable inputs that will just suffice to produce 200 chains. The other two isoquants shown are interpreted similarly. It is obvious that many more isoquants, in principle an infinite number, could also be drawn. This diagram is a graphic display of the relationships expressed in the production function.

Substitution of factors. The isoquants also illustrate an important economic phenomenon: that of factor substitution. This means that one variable factor can be substituted for others; as a general rule a more lavish use of one variable factor will permit an unchanged amount of output to be produced with fewer units of some or all of the others. In the example above, labour was literally as good as gold and could be substituted for it. If it were not for factor substitution there would be no room for further decision after y, the number of chains to be produced, had been established.

The shape of the isoquants shown, for which there is a good deal of empirical support, is very important. In moving along any one isoquant, the more of one factor that is employed, the less of the other will be needed to maintain the stated output; this is the graphic representation of factor substitutability. But there is a corollary: the more of one factor that is employed, the less it will be possible to reduce the use of the other by using more of the first. This is the property known as "diminishing marginal rates of substitution." The marginal rate of substitution of factor 1 for factor 2 is the number of units by

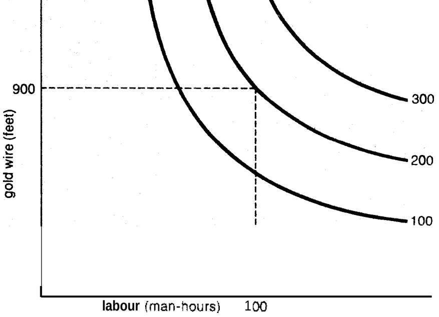


Figure 1: Isoquant diagram of hours of labour and feet of gold wire used per month.

which x_1 can be reduced per unit increase in x_2, output remaining unchanged. In the diagram, if feet of gold wire are indicated by x_1 and goldsmith hours by x_2, then the marginal rate of substitution is shown by the steepness (the negative of the slope) of the isoquant; and it will be seen that it diminishes steadily as x_2 increases because it becomes harder and harder to economize on the use of gold simply by taking more care. The remainder of the analysis rests heavily on the assumption that diminishing marginal rates of substitution are characteristic of the production process generally.

The cost data and the technological data can now be brought together. The variable cost of using x_1, x_2 units of the factors of production is written $p_1x_1 + p_2x_2$, and this information can be added to the isoquant diagram (Figure 2). The straight line labelled v_2, called the v_2-isocost line, shows all the combinations of input that can be purchased for a specified variable cost, v_2. The other two isocost lines shown are interpreted similarly. The general formula for an isocost line is $p_1x_1 + p_2x_2 = \mathrm{v}$, in which v is some particular variable cost. The slope of an isocost line is found by dividing p_2 by p_1 and depends only on the ratio of the prices of the two factors.

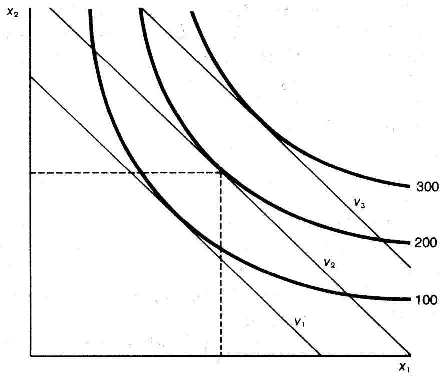


Figure 2: Isoquant diagram for two factors of production, x_1 and x_2 (see text).

The minimization of costs

Three isocost lines are shown, corresponding to variable costs amounting to v_1, v_2, and v_3. If 200 units are to be produced, expenditure of v_1 on variable factors will not suffice since the v_1-isocost line never reaches the isoquant for 200 units. An expenditure of v_3 is more than sufficient; and v_2 is the lowest variable cost for which 200 units can be produced. Thus v_2 is found to be the minimum variable cost of producing 200 units (as v_3 is of 300 units) and the coordinates of the point where the v_2 isocost line touches the 200-unit isoquant are the quantities of the two factors that will be used when 200 units are to be produced and the prices of the two factors are in the ratio p_2/p_1. It may be noted that the cheapest combination for the production of any quantity will be found at the point at which the relevant isoquant is tangent to an isocost line. Thus, since the slope of an isoquant is given by the marginal rate of substitution, any firm trying to produce as cheaply as possible will always purchase or hire factors in quantities such that the marginal rate of substitution will equal the ratio of their prices.

The isoquant-isocost diagram (or the corresponding solution by the alternative means of the calculus) solves the short-run cost minimization problem by determining the least-cost combination of variable factors that can produce a given output in a given plant. The variable cost incurred when the least-cost combination of inputs is used in conjunction with a given outfit of fixed equipment is called the variable cost of that quantity of output and denoted VC (y). The total cost incurred, variable plus fixed, is the short-run cost of that output, denoted $\mathrm{SRC}(y)$. Clearly $\mathrm{SRC}(y) = \mathrm{VC}(y) + \mathrm{R}(K)$, in which the second term symbolizes the sum of the annual costs of the fixed factors available.

Marginal *cost*. Two other concepts now become important. The average variable cost, written $\mathrm{AVC}(y)$, is the variable cost per unit of output. Algebraically, $\mathrm{AVC}(y) = \mathrm{VC}(y)/y$. The marginal variable cost, or simply marginal cost $[\mathrm{MC}(y)]$ is, roughly, the increase in variable cost incurred when output is increased by one unit; *i.e.*, $\mathrm{MC}(y) = \mathrm{VC}\ (y+1) - \mathrm{VC}(y)$. Though for theoretical purposes a more precise definition can be obtained by regarding $\mathrm{VC}(y)$ as a continuous function of output, this is not necessary in the present case.

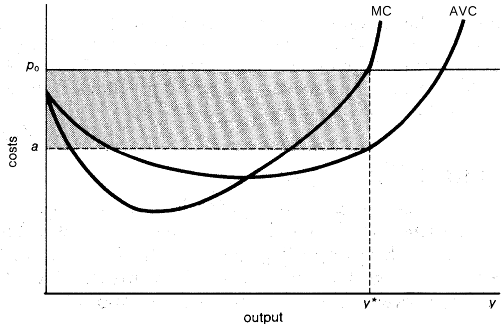


Figure 3: Average variable costs (AVC) and marginal variable costs (MC) in relation to output.

The usual behaviour of average and marginal variable costs in response to changes in the level of output from a given fixed plant is shown in Figure 3. In this figure costs (in dollars per unit) are measured vertically and output (in units per year) is shown horizontally. The figure is drawn for some particular fixed plant, and it can be seen that average costs are fairly high for very low levels of output relative to the size of the plant, largely because there is not enough work to keep a well-balanced work force fully occupied. People are either idle much of the time or shifting, expensively, from job to job. As output increases from a low level, average costs decline to a low plateau. But as the capacity of the plant is approached, the inefficiencies incident on plant congestion force average costs up quite rapidly. Overtime may be incurred, outmoded equipment and inexperienced hands may be called into use, there may not be time to take machinery off the line for routine maintenance; or minor breakdowns and delays may disrupt schedules seriously because of inadequate slack and reserves. Thus the AVC curve has the flat-bottomed U-shape shown. The MC curve, as might be expected, falls faster and rises more rapidly than the AVC curve.

OUTPUT IN THE SHORT RUN

The average and marginal cost curves just deduced are the keys to the solution of the second-level problem, the determination of the most profitable level of output to produce in a given plant. The only additional datum needed is the price of the product, say p_0.

How the firm maximizes its profit

The most profitable amount of output may be found by using these data. If the marginal cost of any given output (y) is less than the price, sales revenues will increase more than costs if output is increased by one unit (or even a few more); and profits will rise. Contrariwise, if the marginal cost is greater than the price, profits will be increased by cutting back output by at least one unit. It then follows that the output that maximizes profits is the one for which $\mathrm{MC}(y) = p_0$. This is the second basic finding: in response to any price the profit-maximizing firm will produce and offer the quantity for which the marginal cost equals that price.

Such a conclusion is shown in Figure 3. In response to the price, p_0, shown, the firm will offer the quantity y^* given by the value of y for which the ordinate of the MC curve equals the price. If a denotes the corresponding average variable cost, net revenue per unit will be equal to $p_0 - a$, and the total excess of revenues over variable costs will be $y^*(p_0 - a)$, which is represented graphically by the shaded rectangle in the figure.

Marginal cost and price. The conclusion that marginal cost tends to equal price is important in that it shows how the quantity of output produced by a firm is influenced by the market price. If the market price is lower than the lowest point on the average variable cost curve, the firm will "cut its losses" by not producing anything. At any higher market price, the firm will produce the quantity for which marginal cost equals that price. Thus the quantity that the firm will produce in response to any price can be found in Figure 3 by reading the marginal cost curve, and for this reason the marginal cost curve is said to be the short-run supply curve for the firm.

The short-run supply curve for a product—that is, the total amount that all the firms producing it will produce in response to any market price—follows immediately, and is seen to be the sum of the short-run supply curves (or marginal cost curves, except when the price is below the bottoms of the average variable cost curves for some firms) of all the firms in the industry. This curve is of fundamental importance for economic analysis, for together with the demand curve for the product it determines the market price of the commodity and the amount that will be produced and purchased.

One pitfall must, however, be ncted. In the demonstration of the supply curves for the firms, and hence of the industry, it was assumed that factor prices were fixed. Though this is fair enough for a single firm, the fact is that if all firms together attempt to increase their outputs in response to an increase in the price of the product, they are likely to bid up the prices of some or all of the factors of production that they use. In that event the product supply curve as calculated will overstate the increase in output that will be elicited by an increase in price. A more sophisticated type of supply curve, incorporating induced changes in factor prices, is therefore necessary. Such curves are discussed in the standard literature of this subject.

The returns to the factors of production

Marginal product. It is now possible to derive the relationship between product prices and factor prices, which is the basis of the theory of income distribution. To this end, the marginal product of a factor is defined as the amount that output would be increased if one more unit of the factor were employed. all other circumstances remaining the same. Algebraically, it may be expressed as the difference between the product of a given amount of the factor and the product when that factor is increased by an additional unit. Thus if $MP_1(x_1)$ denotes the marginal product of factor 1 when x_1 units are employed, then $MP_1(x_1) = f(x_1 + 1, x_2, \dots, x_n; k) - f(x_1, x_2 \dots, x_n;$ k). The marginal products are closely related to the marginal rates of substitution previously defined. If an additional unit of factor 1 will increase output by f_1 units, for example, then one more unit of output can be obtained by employing $1/f_1$ more units of factor 1. Similarly, if the marginal product of factor 2 is f_2, then output will fall by one unit if the use of factor 2 is reduced by $1/f_2$ units. Thus output will remain unchanged, to a good approximation, if $1/f_1$ units of factor 1 are used to replace $1/f_2$ units of factor 2. The marginal rate of substitution is therefore f_2/f_1, or the ratio of the marginal products of the two factors. It has already been shown that the marginal rate of substitution also equals the ratio of the prices of the factors, and it therefore follows that the prices (or wages) of the factors are proportional to their marginal products.

This is one of the most significant theoretical findings in economics. To restate it briefly: factors of production are paid in proportion to their marginal products. This is not a question of social equity but merely a consequence of the efforts of businessmen to produce as cheaply as possible.

Further, the marginal products of the factors are closely related to marginal costs and, therefore, to product prices. For if one more unit of factor 1 is employed, output will be increased by $MP_1(x_1)$ units and variable cost by p_1; so the marginal cost of additional units produced will be $p_1/MP_1(x_1)$. Similarly, if additional output is obtained by employing an additional unit of factor 2, the marginal cost will be $p_2/MP_2(x_2)$. But, as shown above, these two numbers are the same; whichever factor i is used to increase output, the marginal cost will be p_i/MP_i (x,) and, furthermore, the firm will choose its output level so that the marginal cost will be equal to the price, p_0.

Thus it has been found that $p_1 = p_0 MP_1(x_1)$, $p_2 = p_0 MP_2(x_2), \dots$, or the price of each factor is the price of the product multiplied by its marginal product, which is the value of its marginal product. This, also, is a fundamental theorem of income distribution and one of the most significant theorems in economics. Its logic can be perceived directly. If the equality is violated for any factor, the businessman can increase his profits either by hiring units of the factor or by laying them off until the equality is satisfied, and presumably the businessman will do so.

The theory of production decisions in the short run, as just outlined, leads to two conclusions (of fundamental importance throughout the field of economics) about the responses of business firms to the market prices of the commodities they produce and the factors of production they buy or hire: (1) the firm will produce the quantity of its product for which the marginal cost is equal to the market price and (2) it will purchase or hire factors of production in such quantities that the price of the commodity produced multiplied by the marginal product of the factor will be equal to the cost of a unit of the factor. The first of these conclusions explains the supply curves of the commodities produced in an economy. Though the conclusions were deduced within the context of a firm that uses two factors of production, they are clearly applicable in general.

LONG-RUN ADJUSTMENTS

The relationship between the short run and the long run

The theory of long-run profit-maximizing behaviour rests on the short-run theory that has just been presented but is considerably more complex because of two features: (1) long-run cost curves, to be defined below, are more varied in shape than the corresponding short-run cost curves, and (2) the long-run behaviour of an industry cannot be deduced simply from the long-run behaviour of the films in it because the roster of firms is subject to change. It is of the essence of long-run adjustments that they take place by the addition or dismantling of fixed productive capacity by both established firms and new or recently created firms.

At any one time an established firm with an existing plant will make its short-run decisions by comparing the ruling price of its commodity with cost curves corresponding to that plant. If the price is so high that the firm is operating on the rising leg of its short-run cost curve, its marginal costs will be high—higher than its average costs—and it will be enjoying operating profits, as shown in Figure 3. The firm will then consider whether it could increase its profits by enlarging its plant. The effect of plant enlargement is to reduce the variable cost of producing high levels of output by reducing the strain on limited production facilities, at the expense of increasing the level of fixed costs.

In response to any level of output that it expects to continue for some time, the firm will desire and eventually acquire the fixed plant for which the short-run costs of that level of output are as low as possible. This leads to the concept of the long-run cost curve: the long-run costs of any level of output are the short-run costs of producing that output in the plant that makes those short-run costs as low as possible. These result from balancing the fixed costs entailed by any plant against the shod-run costs of producing in that plant. 'The long-run costs of producing y are denoted by $LRC(y)$. The average long-run cost of y is the long-run cost per unit of y (algebraically $LAC(y) = LRC(y)/y$). The marginal long-run cost is the increase in long-run cost resulting from an increase of one unit in the level of output. It represents a combination of short-run and long-run adjustments to a slight increase in the rate of output. It can be shown that the long-run marginal cost equals the marginal cost as previously defined when the cost-minimizing fixed plant is used.

Costs in the long run. Cost curves appropriate for long-run analysis are more varied in shape than short-run cost curves and fall into three broad classes. In constant-cost industries, average cost is about the same at all levels of output except the very lowest. Constant costs prevail in manufacturing industries in which capacity is expanded by replicating facilities without changing the technique of production, as a cotton mill expands by increasing the number of spindles. In decreasing-cost industries, average cost declines as the rate of output grows, at least until the plant is large enough to supply an appreciable fraction of its market. Decreasing costs are characteristic of manufacturing in which heavy, automated machinery is economical for large volumes of output. Automobile and steel manufacturing are leading examples. Decreasing costs are inconsistent with competitive conditions, since they permit a few large firms to drive all smaller competitors out of business. Finally, in increasing-cost industries average costs rise with the volume of output generally because the firm cannot obtain additional fixed capacity that is as efficient as the plant it already has. The most important examples are agriculture and extractive industries.

CRITICISMS OF THE THEORY

The theory of production has been subject to much criticism. One objection is that the concept of the production function is not derived from observation or practice. Even the most sophisticated firms do not know the direct functional relationship between their basic raw inputs and their ultimate outputs. This objection can be got around by applying the recently developed techniques of linear programming, which employ observable data without recourse to the production function and lead to practically the same conclusions.

On another level the theory has been charged with excessive simplification. It assumes that there are no changes in the rest of the economy while individual firms and industries are making the adjustments described in the theory; it neglects changes in the technique of production; and it pays no attention to the risks and uncertainties that becloud all business decisions. These criticisms are especially damaging to the theory of long-run profit maximization. On still another level, critics of the theory maintain that businessmen are not always concerned with maximizing profits or minimizing costs.

Though all of the criticisms have merit, the simplified theory of production does nevertheless indicate some basic forces and tendencies operating in the economy. The theorems should be understood as conditions that the economy tends toward, rather than conditions that are always and instantaneously achieved. It is rare for them to be attained exactly, but it is just as rare for substantial violations of the theorems to endure.

Only the simplest aspects of the theory were described above. Without much difficulty it could be extended to cover firms that produce more than one product, as almost all firms do. With more difficulty it could be applied to firms whose decisions affect the prices at which they sell and buy (monopoly, monopolistic competition, monopsony). The behaviour of other firms that recognize the possibility that their competitors may retaliate (oligopoly) is still a theory of production subject to controversy and research.

BIBLIOGRAPHY. P.A. SAMUELSON, *Economics,* 8th ed., ch. 23–28 (1970), is an excellent introductory treatment of the subject. Authoritative, intermediate-level discussions may be found in W.J. BAUMOL, *Economic Theory and Operations Analysis,* 2nd ed., ch. 3, 11–13 (1965); and in G.J. STIGLER, *The Theory of Price,* 3rd ed. (1966). V.L. SMITH, *Investment and Production* (1961), emphasizes especially the relationship between long-run costs and investment. Probably the best technical presentation is in P.A. SAMUELSON, *Foundations of Economic Analysis,* ch. 4 (1947). For a discussion of the evolution of the theory, see G.J. STIGLER, *Production and Distribution Theories* (1968); on the relationship between theory and empirical data, P.J.D. WILES, *Price, Cost and Output* (1956); and CONFERENCE ON PRICE RESEARCH, COMMITTEE ON PRICE DETERMINATION, *Cost Behaviour and Price Policy* (1943). J.R. HICKS, *Value and Capital,* ch. *6–8* (1939); and J. VINER, "Cost Curves and Supply Curves," *Zeitschrift für Nutionalokonornie,* 3:23–46 (1931), remain classic works; for an excellent survey of recent work with a full bibliography, see A.A. WALTERS, "Production and Cost Functions," in *Econometric~*,31:1–66 (1963).

(R.D.)

Production and Consumption, Government's Role in

A significant development in the intellectual history of the 20th century has been the explicit recognition by economists, politicians, and the public at large of the importance of government in the operation of the economy. An older view, popular in the 18th and 19th centuries, held that "that government is best which governs least." This usually was interpreted to mean that government should restrict its activities to defense, to fire and police protection, to the performance of various administrative tasks, and to certain statistical and other services that do not lend themselves well to the efforts of firms and private individuals. Indeed, most of these tasks were performed by private armies, volunteer fire departments, vigilantes, etc. A more common view now is that all levels of government (central, regional, local) are, and must be, important participants in the production and consumption decisions of a modern economy.

THE PUBLIC AND PRIVATE SECTORS

The idea that government, especially the central government, must participate in basic economic decisions is not in itself entirely new. Some governments have completely dominated the lives of their people. In the past, this was frequently associated with a marriage between secular and religious power, the outstanding illustration being the civilization of the Incas, in which the emperor was at once the divinity, the lawgiver, and the law in a society planned down to the last detail. Remote from the concerns of modern industrial economies as such societies seem, they do show the possible social costs of excessive centralization. The technical aspects of excessive centralization are, however, more germane to present-day problems, and these will be discussed below.

The modern conception of the economic role of the public (government, as distinct from private) sector is that the various levels of government must be recognized as major links in the economic process. Their contributions to political and economic welfare must, however, be evaluated, not merely in terms of technical efficiency but also in the light of acceptability to a particular society at a particular state of political and economic development. Even in a dictatorship, this principle is formally observed, although the authorities usually destroy the substance by presuming to interpret to the public its collective desires.

The question of efficiency

The relative efficiency of the private and public sectors in allocating economic resources (capital, labour, materials) must be subject to continuous appraisal by an estimate of the comparative costs and benefits of resource-using projects entrusted by society to the public or private sectors. Indirect costs and benefits, which are often crucial to the decision whether a given project should be undertaken by private industry or by government, are difficult to estimate. For example, a new public highway may contribute to the enjoyment of the automobile and may reduce transport costs, but it also may destroy the ecology of a wilderness area. In the private sector, production of a cheap and safe airplane may enhance leisure and increase the mobility of business executives, but it also may add seriously to the noise level. Budgeting, in the broadest sense, then, involves the weighing of alternative costs; and these costs are not merely the obvious, tangible ones that are taken into account in much cost-benefit analysis but include such fundamental, yet intangible, consequences as the impact on the future of decisions made in the present.

Modern societies recognize, therefore, that one of their basic concerns is continuously to weigh alternatives with respect to the optimum allocation of resources between and within the public and private sectors. Once the concept of social welfare that is acceptable to a national ethos has been established, it must be recognized that there is no presumption in favour of either public or private responsibility for economic decisions. The palm is awarded on the basis of comparative efficiency.

Two subjects will occupy the remainder of this article.

The first is a consideration of the logical basis for arriving at the optimal ratio of public to private spending; the second, an examination of the significance of trends in public and private spending for the centralization of economic power.

Public and private goods

A clarification of terms is necessary at the outset. A strictly private good is one that can be consumed by only one purchaser: if X buys and consumes an orange, it is not available to Y. A strictly public good, on the other hand, is equally enjoyable by everyone. A park is accessible to all, even though wealthy taxpayers may contribute a relatively high proportion of its cost. (There are, obviously, mixed cases in between these two extremes.) Two points need to be observed with respect to this distinction. First, although governments usually produce the bulk of public goods, they also produce toll roads, subways, electricity, and air-travel services that are sold in markets just as are goods produced by private firms. The widespread use of subsidies to hold down the prices of these services confuses matters, however, for subsidies involve a transfer of purchasing power from taxpayers to the users of the government-produced output. Conversely. business firms may produce goods that are not sold in private markets. An example is military aircraft. Second, the distinction between public and private goods, though it is a handy one for logical thinking, is not absolutely clear-cut. Many public, or collective, goods could be sold in private markets. Flood control, for example, may involve the creation of artificial lakes that can be enjoyed by all without charge. On the other hand, the government could sell these services in the same way that the owner of a private amusement park does. Again, flood-control installations may lead to a rise in the value of surrounding land, and the government conceivably could charge rent to the beneficiaries.

THE ALLOCATION OF RESOURCES

A principle is needed to determine the amount of resources that should be devoted to the production of public goods. Such a principle exists in the private sector, where the allocation of resources is determined by (1) consumers, when they attempt to derive the most satisfaction in distributing their incomes between consumption and saving and between different types of consumers' goods and services, and *(2)* producers, in maximizing profit plus some combination of other possible business objectives (the firm's desired position in the industry, for example, or its anticipated size at some future date). Clearly, such a principle also exists with respect to public goods, except that here a complex political process aims at interpreting the desires of the public with respect to what public goods should be produced and how they should be financed. (It should be noted that consumer choice in any literal sense is often rather remote from reality in the provision of both private and public goods and services. Just as firms successfully influence consumers by advertising campaigns, so in the public sector a variety of pressures are brought to bear on the voter.) Public discussions of government activity often revolve around the issue of taxation. But it is important not to overestimate the part played by the tax decision in the budgetary process. The taxes that pay for the production and distribution of public goods do, of course, reduce the ability of individuals and firms to purchase private goods. Taxes, along with government borrowing, are one of the financial instruments utilized for transferring resources from the private to the public sector. But the decision to tax usually plays a subsidiary role in the decision to transfer resources to the government. Taxation is especially important when a subordinate level of government elects to restrict its spending to receipts from a given tax system, and there have been many cases in which an inefficient central-government tax structure has discouraged extensive spending programs. It is much more usual, however, for the government to decide first on the program, even when spending and tax proposals are presented simultaneously to the legislature in an integrated budget. In a progressive economy, tax receipts rise automatically with rising national income and spending. To the extent that no tax rate increases are necessary to finance expanded public-spending programs, the tax decision is likely to be subordinate to the spending decision. In an era of sharply rising public demand for collective goods and services, the automatic rise in tax receipts from a given tax structure will be inadequate; and distaste for higher tax rates is very likely to constrain decisions with respect to government resource taking.

The preceding remarks can be summarized as follows: the decision on the proportion of available resources to be devoted over a given period of time to public and private goods is a budgetary matter. But the budgetary decision is not merely a choice between the anticipated benefits of the public goods and the additional taxes required to finance them. (This is, of course, an important consideration in the political implementation of any decision to change the level of public spending, for taxes are likely to be prominent in the methods of financing that have to be considered.) The fundamental decision is, rather, the choice between devoting a certain amount of physical resources to the production of commodities for purchase in private markets and allocating the same resources to the output of public goods.

In a small country with a homogeneous population and a stable political system, this decision is likely to be made rather smoothly. Indeed, the very smoothness may be a source of concern if the favourable effects of a reasonable amount of conflict on economic and social progress are considered. It does not necessarily follow that in a large country, beset by an array of complex issues, there may not be periods during which the weight of public opinion unambiguously favours a rise or a fall in the proportion of public to private goods. Usually, however, there is much division of opinion.

The concept of balance

The answer to what it is that determines whether a nation feels comfortable with a particular distribution of resources between public and private use can be given most broadly by reference to the concept of balance, but it is necessary to know what it is that has to be balanced. The term balance may itself be somewhat misleading; sometimes a country progresses more rapidly with imbalance, in the usual sense of the term, than with balance. There is considerable debate over the relative merits of balanced and unbalanced growth and therefore of balance and imbalance in the constituents of growth, including public and private spending. Imbalance would be justified, for example, if there were reason to believe that public spending would, in the foreseeable future, be more effective than would private investment in opening up new markets. One example is education, which might be expected to have the incidental effect of creating more sophisticated wants on the part of the general public. Another is massive government spending on the conservation of natural resources, pure air and water, and open spaces — necessary ingredients of a viable society in future generations.

It is possible, nevertheless, to accept the view that, with a sufficiently broad horizon, efficiency calls for balance in the use of resources by the public and private sectors. The question then becomes one of whether there are any mechanisms that tend to keep society on the desired path. The idea that such a mechanism exists in the "way of growth" of the social organism is deep-seated and of ancient lineage. Heracleitus thought that deviations from the inherent growth process of the organism are punished. The frequent recurrence of this idea in the history of thought probably owes something to the feeling that man usually manages to "muddle through." With respect to public expenditure, even a persistent tendency for a rise in the ratio of public to private spending is not necessarily inconsistent with the idea of a self-steering mechanism. One might share the view of the 19th-century German economist Adolph Wagner that not only will the state inevitably increase the scope of its activities so tha‘ its share in national output continuously rises but thai this is socially desirable. On this view, balance refers to the change in the ratio of public to private spending, not to the ratio itself.

Wagner couched his argument in terms of per capita government spending and national output. It is not difficult to understand that public expenditures tend to rise proportionately more than the rise in population. So do private expenditures. Rising trends in productivity, due largely to technological progress, imply rising per capita real income. This "social dividend" is at the disposal of society and is hardly likely to be divided between the public and private sectors in such a way that only one of them benefits from it. Thus, in an absolute sense, both public and private resource use can rise, and the "law" of increasing per capita public expenditure is really only a manifestation of economic progress.

Such a tendency does not imply any inevitable increase in governmental functions. Even though there may be no theoretical ceiling on the proportion of public to private spending short of the government's responsibility for all of it, political and economic pressures build up against an indefinite increase in the ratio. Even in national states subject to highly centralized planning, these pressures become important when the proportion of public to total spending enters the range of about 35 to 40 percent.

Economic budgeting

The pressures that ultimately work against an indefinite increase in the ratio of public to total spending may be summed up in the concept of balance, and this, in turn, is the essence of economic budgeting. Economic budgeting is a form of planning wherein the objectives of government spending programs are carefully weighed against the sacrifice entailed to private spending programs if the former are accepted. Part of government spending is allocated to projects that cannot efficiently be carried out by the private sector. In this case, balance must be sought in terms of the contribution of these projects to social welfare if they are accepted. The rest of government spending is on projects that might be carried out by either government or the private sector, and here the choice should be determined by comparative efficiency.

In the absence of formal central planning, society depends on political forces to keep the balance of public and private spending at a point that represents the optimum technical relationship between the two, but it may be difficult—short of a high degree of formal planning—to determine whether this relationship is in fact the criterion accepted in governmental decisions. It is to be noted that the decision is not solely in the hands of the government. When firms and individuals are determined to increase or maintain their resource taking—if necessary by borrowing from the banks—prices may be driven up; and governmental appropriations made in money terms will be found to command fewer real resources than expected. If government then backtracks from its original plans (in the interest of inflation control), the private sector will have succeeded in tilting the balance in resource distribution between it and the public sector.

No distinction has so far been made between investment and consumption spending. With respect to the former, the comparative roles of the public and private sectors will depend on which category of investment is expected to make the greater contribution to economic growth. The ratio between public and private consumption spending is decided by the legislative and executive branches of government, responding to the interaction of group pressures. The special interests include not only private lobbies but also the various government agencies charged with administering programs concerned with public consumption. Their conflicting aims are resolved periodically in each budget. In wartime the criteria are different. Military expenditures, though necessary to achieve national objectives, do not contribute significantly to the accumulation of productive capital and must therefore be treated as public consumption. Within limits imposed by the necessity of feeding and clothing the population, government consumption takes precedence over private consumption in time of war. The public-consumption target is simply taken to be the maximum that is possible without significantly injuring the will or the capacity to fight. The magnitudes involved may be illustrated by the experience of the United States in World War II. In 1944, out of a national income of $183,000,000,000, some $103,000,000,000, or *56.5* percent, was devoted by all levels of government to public spending. In 1940 the percentage had been 22.6. Large as the increase was, it did not represent the maximum possible; to a surprising extent, the United States was able to have both guns and butter in World War II.

HISTORICAL TRENDS

The historical rise in government spending in nearly all countries in the 19th and 20th centuries cannot be taken as a measure of either the relative importance of government as a whole in economic decision making or the comparative roles of the central and lower levels of government. Inflation aside, in most countries the major reasons for the persistent rise in public spending since the middle of the 19th century have been war and the preparation for war, the rise in the cost of pensions for soldiers of former wars, the great increase of the administrative role of government in response to expanded and urbanized populations, and the marked rise in the demand for a varied list of public services as the vote was gradually extended to the lower income classes.

Writing in 1890, the Irish economist Charles Bastable observed that "in nearly all modern States outlay is steadily increasing," and "the older doctrines of economy and frugality have disappeared." He was referring to doctrines that had developed in the latter part of the 18th century, particularly in connection with the Industrial Revolution. He did not mean that there had been a "golden age" in which governments entirely refrained from interfering in the private sector. As Bastable himself pointed out, even the strictures of Anne-Robert-Jacques Turgot and Adam Smith on "excessive" government intervention did not preclude the encouragement of new industries.

The tradition of government action

In western Europe there was a long tradition of government influence on private economic decisions. The interventionist policies in the England of Henry VIII, Elizabeth I, and Oliver Cromwell, the France of Louis XIV and Colbert, and the Russia of Peter the Great are examples of such influence. But the sense of confidence conferred on the industrial class by the industrial and transportation revolutions of the 19th century, especially in Britain and the United States, produced an atmosphere that was unfavourable to government intervention. This did not, however, prevent rising pressure for government spending on economic resources, together with a secular rise in the magnitude and variety of the output of public goods that is still in evidence at the present time.

The complaints of the 19th-century exponents of laissez-faire are echoed today by the 20th-century "neoliberals," who deplore "governmental" interference with the free-market mechanism in the form of subsidies and tax incentives to business firms and even to individuals (for example, incentives to save rather than consume or to invest savings in government securities). Opponents of government intervention are not, however, concerned as much with the increased role of government in economic affairs as they are with the social cost of interference in the market mechanism. This cost takes two forms: (1) decreased efficiency of the market mechanism as an instrument for the efficient allocation of resources and (2) the arrogation of power by the central government.

As long as the ratio of public to private spending is not extremely high and is not rising rapidly, government intervention in the market mechanism carries more of a threat than does the rise in government output; however, the two phenomena are closely related. For example, despite a stated policy aimed at restraining federal expenditures in the U.S. during the first 30 years of the present century, these expenditures nevertheless rose greatly. The reason was a tremendous expansion in services (*i.e.*, public goods) to agriculture, commerce, and industry. On the one hand, this diversion of resources from the private to the public sector was designed to encourage economic growth by adding to the efficiency of the private sector; on the other hand, in providing these services, the federal government enhanced its own role in the economic process.

The tendency is not inevitable. The costs of centralization may become so burdensome that they produce a countertendency toward decentralization. In some industrialized countries, the public goods demanded by an increasingly sophisticated and urbanized public are felt to be more efficiently provided by secondary political units and by localities rather than by the central government; there is also the feeling that conscious efforts ought to be made to avoid excessive concentration of power in a central government that is remote from many local problems. It is not possible to predict how the contest between centralizing and decentralizing forces will be resolved. But it is of interest that national plans are devoting more and more attention to the gains in efficiency to be derived from a measure of decentralization of authority.

BIBLIOGRAPHY. A useful introduction is J.M. BUCHANAN, *The Public Finances* (1965), especially ch. 5, "Reasons for Growth of the Public Sector." On a more theoretical level are L. JOHANSEN, *Public Economics* (Eng. trans. 1965); and RA. MUSGRAVE, *The Theory of Public Finance* (1959). The question of whether the public sector of the economy ought to be expanded is dealt with in F.M. BATOR, *The Question of Government Spending* (1960); and J.K. GALBRAITH, *The Affluent Society* (1958). For historical trends in public expenditure, see C.F. BASTABLE, *Public Finance* (1892); H.F. WILLIAMSON (ed.), *The Growth of the American Economy*, 2nd ed. (1958); and A.T. PEACOCK *et* al., *The Growth of Public Expenditure in the United Kingdom* (1961).

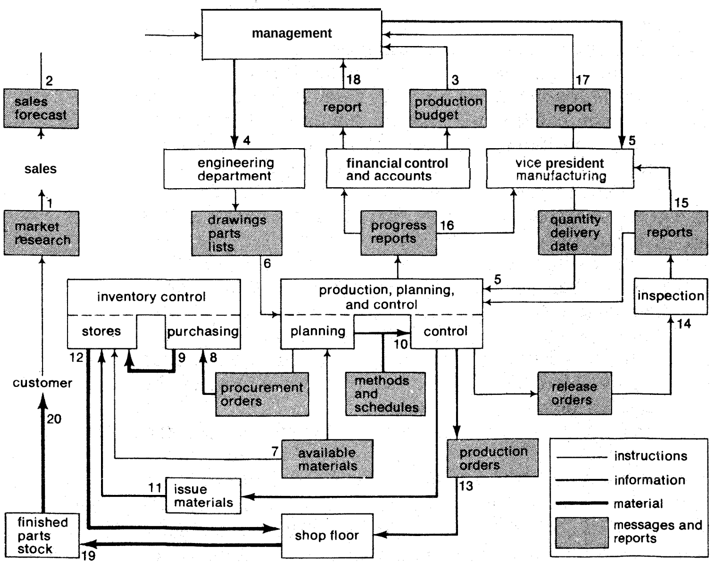
(K.E.P.)

Production Management

Production, though it is often thought of as the processes involved in the operation of a manufacturing plant, is not restricted to the making of material objects; it may also include the performance of services, such as maintenance or transportation. Production management is concerned with planning and controlling the process of production so that it moves smoothly at the required level.

From customer to drawing board

Flow channels. In a large enterprise, a complex managerial system may be needed to do the work of advising, coordinating, controlling, and providing services to the production departments. In the last analysis, production management consists of making choices about the use of men, money, materials, and time. The way this is done in one fairly typical manufacturing firm is shown in the Figure, in which the main flow channels of instructions, information, and materials, are outlined. The sales department (number 1 in the Figure) analyzes the market for the firm's products. It also carries out market research to estimate the selling potential of new products. It makes a sales forecast (2) and submits it to top management. The financial department, in consultation with the production department, draws up a production budget (3). The proposed budget and the sales forecasts are closely scrutinized by management, and decisions as to the annual or semi-annual quantity of a given item to be produced are reached. The engineering department (4) is then instructed to prepare drawings, parts lists, and specifications or to check and modify existing ones. The manufacturing budget is adjusted correspondingly. Instructions are issued to the production planning and control department (5) specifying the quantities to be made and their delivery schedules. The necessary technical information (6) is obtained from the engineering department (including drawings, parts lists, specifications, standards, and so forth) and passed on to the planning section.



Production–consumption cycle.

From drawing board to customer

One of the first functions of the production planning and control department is to be well informed about the availability of materials and the expected delivery dates of materials already ordered (7). Detailed schedules are prepared. Inventory levels are checked to determine the orders that have to be issued (8) for procurement of materials and standard parts. Parts and assemblies that are subcontracted are also ordered by the purchasing department. When these purchased materials arrive, they are inspected and stored (9) until instructions to release them to the shops are received. The production-planning section supplies all the necessary data on methods of production, the loading of machines, and machine utilization, as well as production schedules, to the control section (10) for dispatching. The control section ieleases orders (11) for materials, tools, fixtures, and so on. Supplies (12) are released to the shop. Detailed production orders (13) are dispatched to the shop by the production-control section, specifying what, how, when, and where operations should be performed.

The control functions continue throughout the manufacturing period, and progress is constantly compared with the schedules so that suitable modifications may be considered and made when required. Inspection orders (14) are released. This is the function known as quality control, and its purpose is to ensure that the product conforms to specifications. Final inspection is carried out before the product leaves the shop. Evaluation of the production operations is the main part of the control function and has to be carried on both during and after these operations. Inspection reports (15) are one element of evaluation; they form the basis for corrective changes in the processes or methods of production, and sometimes even for changes in the specifications of raw materials. The production planning and control department reports on the progress of the work (16) to the management official responsible for manufacturing. These reports also go to the financial control department. The control section also evaluates data obtained from the shops about operation times, the idle time of men and machines, the causes and effects of breakdowns, fluctuations in output, and the like. The official responsible for manufacturing in turn reports (17) to management, which also receives a report (18) from the financial department. The finished product is transferred to stock (19). Finally, the product is delivered to the customer (20), who, after comparing the characteristics of the product with those of its competitors and with his expectations, is ready to contribute his views and reactions to market researchers.

This general description of one manufacturing firm's operations is sufficient to indicate the various activities and functions that comprise production management. Some of these may be examined in more detail.

The stages of control

The control function. Much of production management is concerned with control. Control has two purposes: first, to ensure that operations are performed according to plan and, second, to evaluate the production plan and to see whether a better one could not be devised in the light of the experience gained. Production control may be divided into four stages: (1) the observation and recording of progress, (2) the analysis of factual data in relation to plans and objectives, (3) the taking of corrective action to modify plans and redirect activities, and (4) evaluation. In each of these stages the production controller has to take certain steps or perform certain operations. Examples of these, as they relate to the control of processes, to the control of inventory, to inspection, and to cost control, appear in the Table.

Production-control Activities

	processes	inventory	inspection	costs
Observation	measuring rate of output; recording idle time or downtime	recording stock levels	inspecting materials and parts	collecting cost data
Analysis	comparing progress with the plan	analyzing demand for stocks in different uses and at different times	estimating process capabilities	computing costs in relation to estimates
Corrective action	expediting	issuing production and procurement orders	initiating full inspection; adjusting processes	adjusting selling price of product
Evaluation	estimating production capacity and maintenance schedules	drawing up replenishment policies and inventory systems	reassessing specifications; improving processes and procedures	evaluating production economics; improving data

Control of variety

One of the important problems in production management is control of the variety of products, materials, and methods. Variety, from the standpoint of management, is a kind of disorder; it has a tendency to grow, unless a conscious effort is made to limit it. Variety is encouraged by the efforts of salesmen to meet the individual tastes of their customers. From the standpoint of the firm there are advantages in having a variety of products: it can satisfy a wider range of demand, it can hold customers who might be inclined to take their business elsewhere, it can adjust more readily to changes in the market, and it may thus be enabled to expand its business into new lines or to acquire new categories of customers. But there are also advantages in limiting the variety of products: the firm can then operate with smaller stocks of materials and finished goods, and with less plant and equipment; it can simplify production planning and control procedures; it can reduce the range of skills required and simplify its training methods; and it can get higher output per man-hour.

A balance must therefore be sought. There are not many products for which demand is high enough to justify the existence of specialized production lines working on a continuous schedule. The desire to keep production facilities busy, coupled with the pressure from customers who increasingly prefer variations of the basic product instead of the standard model, lead inevitably to a diversification of products. It is not uncommon, among firms with a diversified product range, to find that one-quarter of the products account for three-quarters of the total sales receipts and there have been cases in which 10 or 20 percent of the products were responsible for more than 80 percent of the receipts.

Scheduling. Production scheduling involves sequencing the activities of the plant so that the product will emerge from the assembly line at the time set for it. This task becomes complex when the same facilities are used for several products, or when some jobs can be run concurrently while others must be done in sequence. In some cases the problem is to steer the job through the system, whenever alternative routes exist, and to determine the queuing discipline at each queue—that is, to order the jobs in the sequence in which they are to be processed by each machine. In other cases there is the problem of determining production levels for each product in order to meet various levels of demand.

Project scheduling. When a project consists of numerous individual jobs, only some of which have to be carried out in a particular sequence, it is necessary to draw up a schedule showing which jobs control the completion of the project. An example of such a project is the construction of an office building, in which some operations (for example, the steelwork) are prior to the performance of others, such as brickwork, although both may take place concurrently in different locations. One approach to this is network, or critical-path, analysis. The project plan is drawn up in the form of a network diagram, consisting of arrows which each represent a particular job; the points where the arrows join (called nodes) denote the starting or finishing of jobs. The network will show which jobs are concurrent and which follow a sequence. The diagram makes it possible to determine the "critical path" that consists of all the jobs whose actual duration must not exceed a given minimum if the project is to be completed on schedule. The use of network analysis in project scheduling allows the reallocation of resources and the rearrangement of the sequence of jobs at the planning stage and also provides an effective means of control during the execution of the project.

Job-production scheduling. In some cases several orders have to be processed on common facilities or production centres, each job having its own unique specifications and requirements in terms of production resources. A "job" may be defined as an order consisting of a single item or a batch of identical items. The scheduling problem is to determine the sequence in which jobs should be processed at each centre—that is, the scheduling rules or the "queuing disciplines" that should be adopted.

A distinction can be made between the static case, in which a fixed number of available jobs needs to be scheduled, and the dynamic case, in which jobs arrive in sequence, sometimes at random, and in which scheduling is a continuous activity. Another distinction is usually made between jobs in which processing times are known in advance or at the time of scheduling, and jobs in which processing times may deviate from given estimated times.

The objectives of job scheduling are the minimization of total time for a given number of jobs, the minimization of the average (or the variance or both) of time (or delay time) per job, the meeting of due dates (when jobs are due to leave the production system) or the minimizing of deviations from given due dates. One method of comparing the relative merits of alternative scheduling rules is to use simulation. This consists of generating, with the aid of a computer, simulated jobs with an arrival pattern and work content similar to the situation under stndy, and experimenting with various control procedures.

Batch-production scheduling. In cases in which the rate of production exceeds the rate of demand, an item may be produced in batches in order to avoid excessive stock building. Between batches of one product the manufacturing facilities are available for other products. The scheduler has to determine the batch sizes of the various products and the order in which they should be produced. Two methods have been used to solve this problem. One employs mathematical programming; the other uses a mathematical model that attempts to maximize the ratio of profit to production costs for the whole production cycle (in which all the products are produced at least once in turn) by calculating batch sizes of the products in the ratio of their demand.

Continuous-production scheduling. When the demand for a product is high enough to justify production on a continuous basis, the production level may need to be adjusted from period to period because of fluctuating demand. This is called the production-smoothing problem. When more than one product is involved, it may be treated either as an aggregate-production-smoothing problem

The production-smoothing problem

(in which the production level for the system as a whole is determined and then broken down into schedules for each product), or as a variant of the batch-production problem.

Two principal methods of solving the production-smoothing problem are: (1) the use of linear programming to compare the costs of using alternative production resources (such as regular time, overtime, or subcontracting) in relation to the constraints on the availability of resources and on minimum-demand requirements that must be met; and (2) production-decision rules that take account of future demand in the light of the present level of production (and, when appropriate, employment). Both methods are chiefly concerned with minimizing costs over a given period of time.

Inventory control. Inventories include raw materials, component parts, work in process, finished goods, packing and packaging materials, and general supplies. The control of inventories, vital to the financial strength of the firm, in general involves deciding at what points in the production system stocks shall be held and what their form and size are to be. As some unit costs increase with inventory size—including storage, obsolescence, deterioration, insurance, investment—and other unit costs decrease with inventory size—including setup or preparation costs, delays because of shortages, and so forth—a good part of inventory management consists of determining optimal purchase or production lot sizes and base stock levels that will balance the opposing cost influences. Another part of the general inventory problem is deciding the levels (reorder points) at which orders for replenishment of inventories are to be initiated.

Inventory control is concerned with two questions: when to replenish the store and by how much. There are two main control systems. The two-bin system (sometimes called the min-max system) involves the use of two bins, either physically or on paper. The first bin is intended for supplying current demand and the second for satisfying demand during the replenishment period. When the stock in the first bin is depleted, an order for a given quantity is generated. The reorder-cycle system, or cyclical-review system, consists of ordering at fixed regular intervals. Various combinations of these systems can be used in the construction of an inventory-control procedure. A pure two-bin system, for example, can be modified to require cyclical instead of continuous review of stock, with orders being generated only when the stock falls below a specific level. Similarly, a pure reorder-cycle system can be modified to allow orders to be generated if the stock falls below the reorder level between the cyclical reviews. In yet another variation, the reorder quantity in the reorder-cycle system is made to depend on the stock level at the review period or the need to order other products or materials at the same time or both.

Other aspects of production management. On a broader scale the concerns of production management include the effective use of all resources available to the enterprise. These resources include manpower, materials, machines, money, and methods.

The effective use of manpower has long been the object of work-study techniques first developed by an American industrial engineer, Frederick W. Taylor (1856-1915), who introduced time-and-motion studies, as well as systems of incentive pay for labour. The measurement of repetitive work has come to be widely accepted as a way of increasing labour productivity and reducing labour costs.

The effective use of materials often involves investigations of the causes of scrap and waste, and the study of alternative materials that may be used in the same production process. Likewise, the effectiveness of machines depends on their suitability for specific tasks, the degree of their utilization, the extent to which they are kept in optimum running condition, and the extent to which they can be mechanically or automatically controlled.

The effective use of financiai resources is generally regarded as being outside the responsibility of production management. Rut many problems of production management have a considerable effect on the finances of the enterprise: the selection of processes, the installation and replacement of equipment, and decisions involving inventory and purchasing policies.

Operations research

Production management has always made use of quantitative analysis. Recent developments in the field of operations research or, more broadly, in the management sciences, have increased the mathematical emphasis in production management (see OPERATIONS RESEARCH). Mathematical programming techniques have been applied to a variety of production allocation and scheduling problems. Queuing theory is used in the analysis of production lines and maintenance problems. System simulation, with electronic computers, is used to study the characteristics of complex production systems. From this point of view the production system consists of a sequence of operations that transform a set of physical inputs (raw materials, components, customers) into a specified set of physical outputs (finished products, serviced customers). The transformation takes one or more of the following forms: (1) *disintegration*, transforming a single input into several outputs, as in producing steel bars and sheets from ingots, or a variety of chemical products from crude oil; (2) *integration* or *assembly*, combining many inputs to produce a relatively small number of finished products, as in the manufacture of machines, automobiles, and various chemical products that are made by blending; (3) *service*, performing an operation that changes certain characteristics of an object, as in maintenance, transportation, or the process of aging.

There is even a possibility that production management and other specialized managerial functions may eventually disappear and be replaced by a general science of management. Even if this happens, the actual management of production is not likely to become a mere package of techniques or management by formula. Production decisions are made by people in human organizations existing in a dynamic environment, in which only the fact of change is constant.

BIBLIOGRAPHY. A practical handbook is G B CARSON (ed.), *Production Handbook,* 2nd ed. (1958). An analytical approach to production control is S. EILON, *Elements of Production Planning and Control* (1962). The use of operational-research models ıs covered in E H BOWMAN and R.B. FETTER, *Analysis for Productron and Operations Management,* 3rd ed. (1967); E S. BUFFA, *Production-inventory Systems* (1968); and M K. STARR, *Pi oduction Management: Systems and Synthesis* (1964). More specialized works are M.K. STARR and D.W. MILLER, *Inventory Control: Theory and Practzce* (1962); and J F. MUTH and G.L. THOMPSON (eds.), *Industrial Scheduling* (1963).

(S.E.)

Productivity, Economic

Productivity in economics is the ratio of what is produced to what is required to produce it. Usually this ratio is in the form of an average, expressing the total output of some category of goods divided by the total input of, say, labour or raw materials. In principle, any input can be used in the denominator of the productivity ratio. Thus, one can speak of the productivity of land, labour, capital, or subcategories of any of these factors of production. One may also speak of the productivity of a certain type of fuel or raw material or may combine inputs to determine the productivity of labour and capital together or of all factors combined.

The productivity of labour, land, and capital

Labour is by far the commonest of the factors used in measuring productivity. One reason for this is, of course. the relatively large share of labour costs in the value of most products. A second reason is that labour inputs are measured more easily than certain others, such as capital. This is especially true if by measurement one means simply counting heads and neglecting differences among workers in levels of skill and intensity of work. In addition, statistics of employment and man-hours are often readily available, while information on other productive factors may be difficult to obtain. Not least important in explaining the emphasis on labour as an input is the fact that, historically, technological advance has made itself felt through the displacement of labour—that is, through increases in labour productivity—rather than through the displacement of other factors. Some kinds of labour productivity measures are thus valuable as indicators of this process and of the resulting improvements in man's material well-being.

The productivity of land, though it receives considerably less attention than the productivity of labour, has been of historical interest. In ancient and preindustrial times the products of the soil comprised the bulk of total output, and land productivity thus constituted the major ingredient in a people's standard of living. Soil of low productivity could, and over much of the Earth still does, mean poverty for a region's inhabitants. It is, however, no longer generally believed, as it was in past centuries, that a country's economic well-being is inevitably tied to the productive powers of the land, and the productive potential of the land itself has proved to be not fixed but greatly expandable through the use of modern agricultural methods. Moreover, industrialization, where it has taken place, has greatly reduced man's dependence on agriculture. These circumstances, together with expanding opportunities for trade, have enabled some countries to overcome in substantial degree the handicaps of a meagre agricultural endowment.

Attempts to measure the productivity of capital

The productivity of capital—plant, equipment, tools, and other physical aids—is a subject of long-standing interest to economists, though concern with its empirical aspects is of more recent origin. Improved statistical reporting and the availability of data in a few industrially advanced countries, notably since World War II, have encouraged systematic efforts to measure the productivity of this factor. Compared with achievements in measuring labour productivity, however, the progress realized has been quite limited. There are considerable theoretical and practical difficulties to be overcome.

USES OF PRODUCTIVITY MEASUREMENT

As an index of growth. A nation or an industry advances by using less to make more. Labour productivity is an especially sensitive indicator of this economizing process and is one of the major measures used to chart a nation's or an industry's economic advance. An overall rise in a nation's labour productivity signifies the potential availability of a larger quantity of goods and services per worker than before and, accordingly, a potential for higher real income per worker. Countries with high real wages are usually also those with high labour productivity, while those with low real wages are generally low in productivity. If, for the moment, other productive factors are neglected, one can see that the wage level will then be equal to the total national product divided by the number of workers; that is, it will be equal to the level of labour productivity.

The change in a nation's overall labour productivity during any given interval represents the sum of changes in the major economic sectors and industries. Some sectors and industries move ahead more rapidly than the overall average while others may gain more slowly or even decline. In the movement of a country from a level of low productivity and low income to one of high productivity and high income a strategic role is played by the industrial, rather than by the agricultural and other sectors. In the late 18th and early 19th centuries the effect of the Industrial Revolution was felt first in the manufacture of woollen and cotton textiles, power generation, the metal trades, and machine-making industries. Along with the development of new processes came the development of new products and services that formed the basis for new industries. An outstanding feature of these changes was an increased labour productivity that in turn laid the foundations for an enormous expansion of output. Technological change exerted its influence irregularly and unevenly and continues to do so.

In the compilation of overall averages this diversity is concealed because high rates in some industries offset low rates in others. Thus, the rate of increase of productivity for the economy as a whole varies within narrower limits than the spread of rates among individual industries would suggest. Aside from erratic short-term movements, the rate of growth of productivity may appear to be fairly stable over extended periods. A surge of labour-saving innovations would cause the overall average rate to move higher, while a technological lull would depress the average rate. History suggests that the surges tend to be associated with basic technological changes such as, for example, the steam engine, the gasoline engine, the electric motor, and the concept of the standardization of parts. Once introduced, such inventions or developments are used in many different industries. These surges tend also to be associated with such developments as, for instance, employment of the open-hearth furnace in steel manufacture or the introduction of the steam railroad.

Productivity is valuable also as an indicator of comparative rates of change among industries and products. Growth in general can be better understood if the relative contributions of individual industries and the circumstances underlying productivity changes in each of these industries are understood.

As a measure of efficiency. Productivity is also used to measure efficiency, as an aid in economic planning and forecasting, and as a means of assessing the uses to which resources are being put. As to the first of these, the efficiency of industrial operations, for instance, may be evaluated by the yardstick of output per man or machine, and such a yardstick may also provide the basis for supplemental or premium payments for workers. When pay is based on piecework alone, labour productivity becomes the sole determinant. Productivity may also serve as a standard for grading and evaluating any group of workers performing common tasks, distinguishing the more from the less productive. And applied to equipment, productivity standards can indicate when a machine is performing poorly and is in need of service. In forecasting, productivity estimates are useful when it is necessary to be able to project the performance of the economy at some future date, given the probable size of the working force. A variant of this is common in planning for underdeveloped countries that want to increase their productivity; information about target levels of productivity, together with expectations as to the growth of the labour force and some undeistanding of the relation between capital per worker and output per worker, helps in estimating the amount of capital investment needed to reach the target. Again, estimates of the probable annual gain in labour productivity together with estimates of the probable annual increase in output allow one to estimate how many jobs will become available at some time in the future. Finally, productivity is a helpful analytical tool in studying the possible allocation of resources among different uses. The extent to which resources flow to various uses depends, among other things, on their productivity in each of those uses. Changes in productivity in the course of time alter the pattern of use and cause the quantities of resources required in particular uses to change. The resulting trends depend on several things. On the one hand, an increase in the productivity of, for instance, labour, since it means a decrease in labour requirements per unit of output, will tend to reduce the demand for labour. But it will also imply a cheapening of labour relative to the cost of other competing factors of production. Hence there will be a tendency to substitute labour for other factors. When labour cost represents a large fraction of total cost, a productivity increase will contribute toward a reduction in the price of the product, thereby expanding sales and with them the demand for labour. The net result will depend upon the sum total of all of these separate effects. It is by no means uncommon to find that the expansionary effects predominate, and many economists consider this to be the normal outcome. In any event, the productivity concept and data on productivity trends can contribute to an understanding of resource and output flows.

As an economic standard. Productivity is an important factor in determining prices and wages. Economists are far from a full understanding of the relations among the variables, but there is substantial agreement on the following points:

Variables in price and wage standards

1. The large increases in real wages that have come about over the long term in many countries are closely associated with large increases in labour productivity in these countries.

2. In the absence of increases in labour productivity, a stable price level is inconsistent with persistent increases in money wages.

3. An increase in labour productivity or in the productivity of other factors usually brings with it a reduction in costs and hence tends to result in price reductions, wage increases, or both.

4. In industries in which sales of products are comparatively insensitive to price changes, increases in labour productivity will tend to reduce employment and possibly also reduce wages.

5. Wage increases in individual plants and industries may induce productivity increases by encouraging the substitution of capital and other factors for labour.

This last statement forms the basis for one explanation of the apparently high rate of mechanization in the United States: abundant economic opportunities for labour on the land and in the frontier west, it is said, resulted in a scarcity of labour in industry, and hence in high wages and intensive mechanization. The same logic, with a reverse twist, is sometimes used to account for low labour productivity in underdeveloped countries. In this instance, the argument is that very low wage rates make it economical to rely heavily on labour and to use little capital. While these explanations leave many things out of account, they appear to be valid within limits.

FACTORS THAT DETERMINE PRODUCTIVITY LEVELS

The level of productivity in a country, industry, or enterprise is determined by a number of factors. These include the available supplies of labour, land, raw materials, capital facilities, and mechanical aids of various kinds. Included also are the education and skills of the labour force; the level of technology; methods of organizing production; the energy and enterprise of managers and workers; and a range of social, psychological, and cultural factors that underlie and condition economic attitudes and behaviour.

These variables interact and mutually condition one another in determining productivity levels and their changes. Thus, in any country one expects the level of technology, the skills of the work force, the quantity of capital, and the capacity for rational economic organization to be positively correlated. A country with low productivity is likely to have deficiencies on all counts; a country with high productivity is likely to score high on all. To put it differently, the numerous productivity-determining factors behave as variables in a system of simultaneous equations, with all acting concurrently to shape the outcome. Within this system, there are no grounds for assigning causal priority to one or a few variables. All interact mutually to determine the outcome. Within certain problem frameworks, however, it may be entirely appropriate and indeed essential for explanatory purposes to emphasize certain variables over others.

Two broad problem frameworks may be distinguished, both of them of perennial concern to students of productivity and growth. One of these involves changes in productivity over time, the other involves differences in productivity levels among enterprises, industries, and countries at a given time. Within these frameworks are countless problems and subproblems, each of which may lead to a somewhat different selection and emphasis of variables.

Changes in long-term productivity

Explanations of long-term productivity changes in a country, region, or industry usually stress technological change and, as an adjunct, changes in the quality and quantity of capital. Other variables, though not ignored, are regarded as playing a passive role and are thus given a subordinate position. The justification for this is that change in technological knowledge and the capital embodying it is not only essential to substantial gains in productivity but is the factor most immediately associated with those gains. It ordinarily is perceived as the leading and moving force in the process. When technological change occurs, the quaiiry of capital improves and the amount available to aid each worker usually increases. The kinds of raw materials used may change, with better grades being required or the use of lower grades becoming possible. Changes occur in the way productive factors are organized and production is carried on. Although in some periods and in some circumstances work may have become harder and more tedious following technological advance and although the transition from land to factory has often entailed special hardships, the dominant trend has been toward shorter hours and a diminution of the arduousness of labour,

Emphasis on technological change and capital accumulation as primary forces arises also from a recognition that they are essential and unique to large and systematic advances in productivity. Those gains that can be obtained solely through a reorganization of work or the use of better raw materials or the breakdown of restraining attitudes or practices may occasionally be dramatic, but they are always limited. By contrast, very substantial gains can follow in the wake of growing technological knowledge and increasing supplies of capital. If allowance is made simply for adaptive changes in other factors, the prospects for advance become almost unlimited. Only these two factors can fairly be singled out as constituting the engines of productivity growth.

It has been noted that both the quantity of capital and its quality change as productivity increases, and it is not possible adequately to separate the two in terms of their effects. Increases in capital per worker through the accumulation of more and more of the same kinds of equipment and tools would not lead continuously to proportionate or more than proportionate increases in output per worker. They would, after a point, ·lead to diminishing increases and eventually even to a decline in output per worker. The onset of a decline would be far distant in an industry or economy possessed of a high level of technical knowledge but starting near the bottom of the accumulation ladder and affected by an acute scarcity of capital instruments. But an ultimate decline would be expected.

Qualitative changes in capital, reflecting advances in knowledge and skill and leading to the design and construction of improved capital instruments, offer an escape from this principle. If capital can be steadily improved over time, its expansion need not entail diminishing returns. In countries for which data from broad sectors and many individual industries are available, there is a rough correlation between growth in the quantity of capital per worker and increases in labour productivity.

HISTORICAL TRENDS

Man's use of capital as an aid in production is, strictly speaking, as old as his use of primitive tools of wood and stone. But the introduction of power-driven machinery, its systematic improvement, and its progressive substitution for labour are a set of much more recent phenomena.

Growth of industry. The influences of technological change and industrial development have touched virtually all regions of the world, though in different ways in widely differing degrees. Some areas were involved only through trade, receiving manufactured goods in exchange for raw-material exports. Others developed industry to a limited extent that served principally as an adjunct to foreign trade and did not penetrate deeply into the domestic economy. Still other countries, like Great Britain and the United States, were fundamentally affected as their economies underwent progressive advance and transformation. At the close of the 19th century, France, Italy, Germany, Russia, Japan and Canada, among others, also possessed substantial manufacturing capability. The transition was most evident in western and northern Europe and in North America and least evident in the Middle East, Africa, Asia, and South America.

The countries that stood in the industrial forefront by the first decade of the 20th century remained there up to World War II, though their relative positions shifted. Most conspicuous was the forward surge of the Soviet Union, following the inauguration of the five-year plans in 1928. After World War II large strides were made by other countries, particularly those of eastern and southern Europe. China and India also moved ahead, as did several South American countries. Japan's advance in the decades since 1950 has been outstanding.

The changes that accompany increasing productivity

The consequences for labour productivity of the changes under discussion — industrialization, mechanization, and capital accumulation — are summarized in Table 1 for the United States, Great Britain, Germany, and Japan. Because of the difficulties inhering in productivity measurement, the figures are only rough approximations. Nonetheless, they suffice to distinguish the rates of change among the four countries. It should be noted that in 1890 each of the countries occupied a different rung on the ladder of industrial development. By certain yardsticks, Britain was already a mature country; the United States and Germany were less advanced, while Japan stood in a comparatively early stage of industrialization. Equally important, each country was differently endowed in terms of skills and resources, possessed different social and economic institutions, and was differently situated geographically. As a result, the potentials for further growth differed among the four countries.

Table 1: Long-Term Trends in National Output per Worker
(index numbers, 1890 = 100)

year	country			
	United States	Great Britain	Germany*	Japan
1890	100	100	100	100
1900	122	107	100	144
1910	138	110	107	166
1920	142	100	—	228
1929	172	116	90	366
1938	182	132	127	547
1948	223	132	—	314
1960	295	161	166	747

*For 1948 and later, German Federal Republic (West Germany) only.
Sources: Figures adapted and assembled by the author from data in numerous sources, including the following: Solomon Fabricant, ***Basic Facts on Productivity Change,*** National Bureau of Economic Research, Occasional Paper **63,** Table **A** (1959); Joint Economic Committee, 85th Congress, 1st session, ***Productivity, Prices, and Incomes,*** Tables 1 and 2; Colin Clark, ***The Conditions of Economic Progress,*** 3rd ed., ch. 3 (1957); Kazushi Ohkawa, ***The Growth Rate of the Japanese Economy Since 1878,*** Appendix Table ***6*** (1957); M. Frankel, "Some Implications of International Postwar Productivity Trends," ***The*** 1958 ***Proceedings of the Business and Economic Statistics Section of the American Statistical Association,*** Table ***2*** (1959).

The figures show great unevenness in the rate of gain for each country from period to period; differences among the countries are marked, both for individual subperiods and for the entire 70-year span. These differences reflect the variety of forces that contribute to productivity change, their varying importance at different times and in different places, and the complex ways in which they interact. Britain experienced no perceptible gain from 1890 to 1920 but moved ahead significantly during the next 18 years. The United States gained during the 1938–60 interval at a rate much above those that had prevailed during the preceding 50 years. Japan's rate during the 18 years following 1920 was about 80 percent higher than its rate for the preceding 30 years. The very high rate in Japan between 1948 and 1960 was the result of recovery and restoration following World War II. A similar phenomenon is observable in a number of other countries, including Germany, though the figures in Table 1 do not show it. Britain and Germany are noteworthy for their low average rates over the long term. Japan, in contrast, is unique for the high average rate attained over a very long period.

Changes in labour input. An index of output per worker, like any other single productivity index, provides only limited information on the process of productivity change. Long-term increases in labour productivity have usually been accompanied by reductions in the number of hours worked per day and per week. In Britain or the United States, for example, a 60-hour week was not uncommon in 1850, but a century later 40 hours was a familiar figure. Hence, if the data in Table 1 were put on a man-hour basis, the rates of increase would, in general, be much higher.

A reduction in hours worked is simply one of the ways in which a people may benefit from productivity increases. Such reductions as have taken place probably have not entailed proportionate sacrifices in the quantities of goods and services that, with constant hours, might otherwise have been obtained. For when hours of work are very long, worker efficiency tends to be low, and within limits, as hours are reduced efficiency rises. There is substantial evidence that the decline of the working day from very high levels has brought compensatory gains of this kind.

Systematic long-term increases in labour productivity have also been accompanied by changes in the quality of labour and in the quality and quantity of other inputs. Generalizations about these other changes are, however, difficult to make. In many fields of production and for many classes of work, the introduction of machines has entailed a downgrading of skills. The loss of many traditional handicraft skills during the early stages of mechanization is a case in point. This type of loss has been repetitive throughout the history of mechanization. A machine replaces one class of skills and at the same time generates requirements for other skills, usually of a lower order.

But the sequence is by no means one-directional. Machines are complex, and over the decades their complexity has increased. This growing complexity has created a need for skills of a high order for machine development and for the installation, servicing, and operation of machines. Organizational requirements for the effective use of machines have also become more complex, thereby increasing the need for a large array of highly trained specialists. As a result, the downgrading and displacement of some skills have been accompanied by the upgrading of others and the emergence of new ones. There is little doubt that the latter tendency has predominated over the long term. The general educational and technical qualifications of an industrial, highly productive labour force are very much higher than those of a low productivity labour force engaged largely in agricultural and handicraft pursuits.

Increases in the productivity of capital

Land and capital. Changes in the productivity of other factors have accompanied the increases in labour productivity. Unfortunately the record as to the nature of these changes is far less comprehensive and in some respects less clear than in the case of labour productivity. By introducing constant improvements in farming methods, farm machinery, and fertilizers, technology led to large increases in the productivity of land as measured by the output of foods and fibres per acre. But it should be noted that, unlike output per worker, output per acre does not correlate well with living standards because of the varying intensity with which cultivation is carried on in individual countries. This circumstance accounts, for example, for the fact that output per acre for a particular commodity may sometimes be higher in a poor country than in a rich one. Technology has also made it possible to economize in the use of some raw materials in production and thus raise output per unit of raw material consumed. But in practice, higher productivity, while possible, may not result, since with a cheapening of this factor it often proves economical to use it more liberally than before. Sometimes, also, more mechanized methods of production are inherently more wasteful of raw materials than less mechanized ones.

In some countries, at least for some periods, the productivity of capital has risen. In the United States, from 1900 to 1960, output per unit of tangible capital increased about 80 percent, or at an average annual rate of just over 1 percent, with the bulk of the increase occurring after 1929. On the other hand, it had barely changed between 1890 and 1920. As between the two factors, labour and capital, it is clear that over the long term the major savings have been in labour. Both theoretical considerations and statistical data suggest that continued technological change will bring further large gains in labour productivity, but the probable trend with respect to capital is uncertain.

From the standpoint of the individual enterprise, any reduction per unit in inputs of labour and capital is important. The economic benefit is the same whether a given net reduction in cost results from the saving of labour or of capital or of raw material. But from a worldwide or economy-wide point of view, labour savings occupy a unique place because labour, interpreted in the broadest sense, is not only an input in the productivity process but also represents the goal of that process. Ultimately, goods and services are consumed by people. An increase in output per worker or per capita signals the availability or potential availability of increased quantities of goods and services for each individual. An increase in output per unit of capital or raw material does not necessarily carry this connotation. If the objective is to attain ever-rising living standards, then it is labour productivity that, over the long term, must be increased.

THE COMPARISON OF PRODUCTIVITY TRENDS

Trends in overall productivity are composites of trends in the several sectors and subsectors of the economy and, in the last analysis, of trends in individual industries and enterprises. At these subordinate levels trends often differ widely, both from one another and from the broader economy-wide averages.

Farm and **nonfarm.** Some of this diversity may be seen by comparing the trend in output per man-hour in the total United States private economy with similar trends in farm and nonfarm subsectors. For the entire period from 1909 to 1956 production per man-hour in both subsectors moved ahead by about the same amount and roughly in line with the overall movement. Between 1909 and 1929, however, the increase in the farm sector was far below that in the nonfarm sector, while for the period from 1938 to 1948 the reverse was true. The periods of 1929 to 1938 and 1948 to 1956 show a higher trend in the farm sector than in the nonfarm sector.

The figures serve to qualify the popular impression that agriculture is a persistent laggard in the technological march. In the United States, during the 19th and early 20th centuries, the nonfarm sectors—including mining, manufacturing, communications, and utilities—were in the vanguard. But in later decades, productivity in agriculture rose. The farm surpluses of the 1950s and 1960s were the result.

These productivity movements do not exhibit a close relation to movements in output. For all periods, increases in farm output, though relatively low, averaged close to 1 percent per year. At the same time productivity increases varied widely, ranging from under .5 percent to over 3 percent. For the nonfarm sector an annual 2 percent productivity increase was registered between 1929 and 1938, a period when output declined by 1 percent a year. By contrast, much the same productivity increase occurred from 1938 to 1948 when output advanced at 6 percent per year. This outcome is perhaps not surprising in view of the fact that the sources of output and productivity increases are not identical. Productivity increases will result in output increases provided that the quantity of inputs does not decline. But output increases may also result independently of productivity increases, from an expansion of inputs or from an increase in the intensity with which resources are used.

To gauge the contribution of any sector to the production total it is necessary to consider the size of the sector and its productivity level. Because agriculture in the United States contributes less than 5 percent of the national product and employs less than 9 percent of the labour force, even a large change in output per worker would have but a nominal effect on the total. The outcome may be very different if the sector is large and if its productivity level, measured by value of output per worker, is high relative to other sectors. In that case and depending on the method of measurement, a redistribution of labour in favour of the high productivity sector can produce large productivity changes in the total, even though sector productivity changes little. This source of productivity gain, while important for all dynamic economies, is especially so for one that is in the early stages of industrialization. In such economies, low-productivity agricultural pursuits absorb the bulk of the labour force, and hence the potential gain from their transfer to more productive occupations is very large.

Among countries. Changes in productivity and changes in total output do not always move in pace. When they do, it means that the gains in output have come exclusively from increased productivity and not from changes in the working force. (Strictly speaking, this is true only if the productivity figures relate to output per man-hour; if they represent output per man-year, for example, they may reflect an increase in the working force.) By contrast, an increase in output accompanied by no change in productivity would indicate that thc advance was attributable solely to the use of greater manpower or other resources. The data for the years following World War II show that both rising productivity and greater manpower contributed to the advances in output in all countries but that the influence of the former was generally greater. In Japan, Italy, France, and Austria, for example, productivity was more important, while in Poland, West Germany, the Soviet Union, and Canada manpower increases played a larger role.

Long-run and short-run growth rates

It is ordinarily preferable that increases in output be obtained through productivity growth rather than through a larger labour force, for this implies not only more product but also more leisure, or at least a potential for more leisure. A tacit assumption here is that the leisure is voluntary. If it is nor, then the problem of unemployment must be reckoned in the balance. Italy is a case in point. Between 1948 and 1957 industrial productivity in Italy rose at a rapid rate and accounted for almost the whole increase in output, but industrial employment expanded only slightly and did not relieve serious unemployment.

If average rates of growth over a long period, say 1938–60 for example, are examined, it is found that the figures differ considerably from those for the shorter postwar intervals. Indeed, one is left by the long-term data with a quite conservative impression of the progress made by some countries, both separately and as a group.

The contrast between growth rates over the long term and the short term reflects the impact of war and the efforts to recover from it. Japan and Austria are excellent illustrations. The high rates of increase in productivity realized by Japan and Austria in the postwar period did little more, by 1960, than restore those economies to their prewar levels. Much the same could be said of West Germany and of Britain, though in the latter case the gap between long-term and short-term growth rates is more modest. Other west European countries—France and Italy, for example—had by 1960 surpassed their 1938 levels by more substantial margins, though these economies also were obliged to devote a major fraction of their energies to overcoming wartime arrears. This also applies to the east European countries, the Soviet Union, Poland, and Czechoslovakia. The U.S. registered an average advance in excess of 2 percent from 1938 to 1960. A notable feature of the U.S. trend, in contrast with trends for most other countries, was its stability. Gains in the 1948–60 period were appreciably greater than those for the longer interval, but, comparatively speaking, the gap was small. The comparative closeness of long-term and short-term rates for the U.S. reflects its good fortune in escaping the physical destruction and economic disorganization of war suffered by other countries.

NATIONAL LEVELS OF PRODUCTIVITY

It is sometimes useful to compare the levels of productivity of different countries at the same period of time. Table 2 makes such a comparison, based on estimates of output per worker and per capita for 16 countries. The fact that some of the figures are for 1950 rather than a more recent year is of minor importance because overall data of the kind given tend to change only slowly. Neglecting small intercountry differences, the basic impressions conveyed by the table were doubtless still valid in the early 1960s. Yet it must be stressed that the figures are to be treated as only rough approximations.

Various countries compared

The general level of productivity in the United States is by a wide margin the highest of those shown, and a more comprehensive listing of countries would not alter its position. New Zealand and Canada are not far behind the United States, but the gap between them and the others is fairly wide. Australia, Sweden, and Denmark bracket the high side of what might be called a middle productivity

Table 2: Comparison **of Levels** of Output, About 1960*
(in both columns the figure 100 represents U.S. output)

country	real output per worker	real output per capita
Argentina	35†	27
Australia	62†	72
Brazil	21†	20
Canada	92†	74
Chile	29†	24
Denmark	62	62
France	52	52
Germany, West	46	51
Great Britain	50	58
Ireland	36	39
Italy	37	29
Japan	31	27
New Zealand	88†	90†
Soviet Union	25†	25†
Sweden	57†	65
United States	100	100

*A country's output per capita, relative to that in the U.S., exceeds its output per worker, relative to that in the U.S., if the fraction of its population in the employed labour force is greater than the corresponding fraction in the U.S. † About 1950 rather than 1960.

Sources: Data developed from Colin Clark, ***The Conditions of Economic Progress,*** 3rd ed. (1957); ***Statistical Yearbook*** (United Nations); ***Yearbook of Labour Statistics***, International Labour Organization; ***Economic Survey of Denmark,*** Secretariat of the Government of Denmark; ***Survey of Current Business,*** United States Department of Commerce; ***United States Income and Output***, United States Department of Commerce.

range, while Britain, France, and—with allowance for incomplete postwar recovery—West Germany fall on the lower half of this range. The remaining countries, with productivity levels of between one-third and one-fifth that in the U.S., are located still lower on the scale. A large group of excluded countries, belonging in the underdeveloped class, would occupy still lower positions. This group would embrace the bulk of the peoples in Africa, Asia, the Middle East, and South America. It is well to keep in mind that these people comprise a majority of the world's total population and, accordingly, that the average productivity level for the world as a whole is, by Western standards, very low.

Differences among countries in the relation between the output per worker and output per capita figures are attributable to differences among countries in the percentage of the total population at work. If the proportion of those seeking and finding gainful employment was the same in all countries, the relation between each pair of figures in Table 2 would be the same. One might expect that with higher output and income levels there would be a tendency for the percentage of the population seeking work to decline. Entrance into the labour force would be delayed by longer schooling, and fewer family members would be obliged to work in order to maintain any given income level. Table 2 suggests a relationship of this sort, though it is not a strong one. It may be that the tendency in this direction is stronger than appears but that it is offset by the more abundant and attractive occupational opportunities existing in countries with high incomes. In any case, the differences that exist between the two series are not large and do not alter the overall intercountry pattern. The close correspondence between them affirms the overriding importance of productivity levels for living standards.

The reasons for differences. At least two broad approaches can be taken in seeking to explain such productivity differences as are found in Table 2. One of these may be termed cross-sectional. It seeks to identify the chief factors associated with the observed differences and to ascertain their relative importance. Measurement and analysis of the degree of association between intercountry differences in output per worker and corresponding differences in, say, size of plant is an example of this approach. It does not of itself tell the direction of cause and effect between two variables nor even that a causal relationship is present. Neither does it say anything about the time trend of the relationship—whether it has always existed, whether its form has changed over the years, or why it may have changed. Nonetheless, the approach is helpful in suggesting which factors may contribute significantly to productivity differences.

The other approach might be called developmental, since it involves an assessment of the process of economic change as it affects productivity levels. It goes beyond the cross-sectional approach in that it endeavours to explain how certain causal factors came to be important and how they have interacted with other factors to cause shifts in labour productivity over time. Necessarily in this approach there arises a question as to where and how to apportion responsibility for change. The answers supplied by economists have varied. The English economist T.R. Malthus (1803) selected population growth as his key variable. The Austrian J.A. Schumpeter (1912) emphasized the role of entrepreneurship in a country's development. In much contemporary literature, whether concerned with industrialized or underdeveloped countries, the process of capital accumulation is regarded as of overriding importance. Despite their differences, all of these treatments may be characterized as developmental since they give attention to the movement and influence of particular factors over time.

The two approaches are complementary. While useful hypotheses might be built wholly from one or the other, the value of each is enhanced by their joint use. The cross-sectional approach, in identifying factors of probable significance, facilitates the application of the developmental or historical approach. At the same time, the latter helps to amplify the meaning and significance of the factors found to be associated with productivity differences.

Needless to say, each country's current circumstances and development, as these bear on productivity levels, possess unique characteristics. For this reason no single, simple hypothesis, whether rooted in one approach or in a combination of approaches, is likely to go very far in accounting for the wide range of differences in Table 2. One probably valid generalization does however seem worth making. Differences in growth rates are essential to an understanding of most large productivity differences among countries. Almost without exception, the countries with high productivity are those that have succeeded in mastering improved technology; with the aid of capital accumulation they have diffused this mastery through the several sectors of the economy and have continuously improved it. Only by continued productivity advances over a fairly long period is it possible for a country to achieve such levels as prevail among the high-income countries.

PROBLEMS OF PRODUCTIVITY MEASUREMENT

Complexi of the dat:

Many difficulties, both theoretical and practical, attend the development of productivity statistics.

Heterogeneity. A wide variety of inputs and outputs go to make up the productivity ratio. Ordinarily such ratios are computed not for a single, homogeneous type of output but for product categories that embrace a host of subproducts. Thus the output of the steel industry consists of sheets, wire, rods, structural shapes of various kinds, and a great many other items. Computation of satisfactory productivity data for an industry usually requires, therefore, that account be taken of the diversity of output and its changing composition. This can and often is done by weighing the subproducts in accord with their relative values and then adding them together. Unfortunately, however, relative values change over time, raising the question of what weights to use. Should they be initial year weights or terminal year weights or some combination of the two? The question is not academic, since different weights sometimes produce significantly different amounts of change between the initial and the terminal years. A parallel problem arises on the input side. Strictly speaking, the labour force is not homogeneous. It is made up of a range of skills and specialties that embody varying amounts of education and training, and it is reasonable to suppose that their contributions to the production process vary accordingly.

Incommensurability. Labour inputs are relatively easy to measure, particularly if one is content simply to count heads. But if one wishes to take account of differences in the quality and intensity of labour inputs, the question of measurement immediately arises. At what rate should one type of labour be converted into another? Wage rates often are used for the purpose, not because of a belief that they are correct but because they are the most readily available yardstick.

Measurement problems are equally difficult in connection with capital inputs. Plant and equipment lose their value gradually over time, and in any one year contribute but a part of this value in production. There is no universally acceptable procedure for measuring this contribution. There exist a few recognized accounting practices, but no one of them is recognized as superior to the others. It probably is true that no one procedure is best in all situations but that each has unique advantages for certain applications. Measurement is further complicated by the diverse composition of capital that rules out the sort of headcounting that can be done with labour inputs.

Moreover, the composition of capital changes over time with the result that the kinds of machines and facilities used in one year may differ markedly from those employed a generation earlier. Any one of these circumstances would alone suffice to make measurement difficult.

For some factors of production, such as organization or managerial ability, there seems to be no acceptable method of measurement. The quality and very likely the quantity of each has changed greatly over the years, and their importance relative to other inputs has changed. But no reasonably rigorous means for evaluating these changes is available.

On the output side, measurement difficulties are especially apparent with regard to services. In contrast to commodity outputs, which are tangible and divisible into identifiable units, some service outputs are very difficult to define and measure satisfactorily. In concrete terms, what is the output of an enterprise engaged in business consulting? Of an organization dispensing banking services? Of a government agency exercising regulatory functions? More is involved here than the fact that each of these activities embraces a mixture of distinguishable services. The main trouble is the lack of any acceptable standard by which to measure such outputs.

This problem is not unique to service outputs, but it shows itself most clearly in that sphere. Lacking a suitable direct measure of service outputs the student will often use inputs (numbers of workers) as a substitute measure, the assumption being that inputs and outputs are equal or will move together over time. Whatever its merits for gauging output, this method is unsuited to measuring productivity. For the validity of a measure of productivity depends on a clear separation of outputs from inputs. Considering the nature of the problems, it is not surprising that the bulk of available productivity information relates to commodity production.

Lack of information. Censuses of production and population, national income statistics and their components; series regularly published on industrial and farm output, on wages, hours, and the labour force; and data provided by regulatory and other government agencies are among the basic source materials. The scope and quality of these materials vary widely among countries and tend to be more comprehensive and reliable in the more advanced, industrialized countries. With few exceptions, unfortunately, such data are not collected for the purpose of providing information on productivity. Special studies often must be undertaken either to supplement what information is available or, for certain input–output areas, to obtain any data at all. Even where information for recent periods of time is extensive the researcher may find it difficult to prepare productivity tables that extend back very many years, and for no country can knowledge of productivity trends in the 19th century be described as better than meagre. The situation for many underdeveloped countries is even less satisfactory.

A miscellany of other difficulties affect productivity measurement. One is the problem of choosing a time period over which to measure changes in productivity. Another involves the inputs of the productivity ratio. If the input is labour, should it cover all employees or only production workers? Should it be on a man-hour or a man-year basis? Yet another problem is the content of the output category and the breadth or narrowness with which it is defined. Decisions on these problems, which are both theoretically sound and practical, are often very hard to reach.

BIBLIOGRAPHY. General works dealing with productivity and its measurement include: SOLOMON FABRICANT, ***Basic Facts on Productivity Change*** (1959); JEAN FOURASTIE, ***La Producfivith,*** 5th ed. (1962); GERHART E. REUSS, ***Produktivitatsanalyse: Ökonomische Grundlagen und statistische Methodik (1960);*** UNITED STATES BUREAU OF LABOR STATISTICS, ***Productivity: A Bibliography*** (1966); and the UNITED STATES CONGRESS, JOINT ECONOMIC COMMITTEE, ***Productivity, Prices, and Incomes*** (1967). Some of the earliest work in the measurement of productivity was done in the United States. A pioneering study was SOLOMON FABRICANT, ***Employment*** in ***Manufacturing, 1899–1939: An Analysis of Its Relation to the Volzime of Production*** (1942). Other studies using U.S. data are EDWARD F. DENISON, ***The Sources of Economic Growth in the United States and the Alternatives Before Us*** (1962); and JOHN W. KENDRICK, ***Productivity Trends in the United States*** (1961). For productivity trends in various countries, see ABRAM BERGSON and SIMON KUZNETS (eds.), ***Economic Trends in the Soviet Union*** (1963); COLIN CLARK, The ***Conditions of Economic Progress,*** 3rd ed. (1957); EVSEY D. DOMAR ***et al.,*** "Economic Growth and Productivity in the United States, Canada, United Kingdom, Germany and Japan in the Post-War Period," ***Review of Economics and Statistics,*** 46:33–40 (1964); ANGUS MADDISON, ***Economic Growth in the West: Comparative Experience in Europe and North America*** (1964). The relation of productivity to technoloey is examined in W.E.G. SALTER, ***Productivity and Technical Change,*** 2nd ed. (1969).

(Ma.F.)

Prokofiev, Sergey

The range and imagination of the huge musical legacy of the Russian composer Sergey Prokofiev make clear his central role in enriching the structure and the expressiveness of Western music in the first half of the 20th century. His more than 135 works include works of almost every musical genre, from children's songs to oratorios, with contributions that were especially important in the development of opera and ballet, in which he was an audacious reformer, and nearly as important in symphonic and piano music. These works reflect not only the distinct musical traditions of the 19th-century Russian masters and of the innovators of 20th-century music but also the social upheavals of a half-century that embraced cataclysmic revolutions and two world wars.

Sovfoto

Prokofiev.

Pre-Revolutionary period. Sergey Sergeyevich Prokofiev (Prokofjev in the transliteration system of the Akademiya Nauk of the Soviet Union) was born on April 23, 1891, in the village of Sontsovka in Ekaterinoslav province (now Donets region) into a family of agriculturalists. Village life, with its peasant songs, left a permanent imprint on him. His mother, a good pianist, became the highly gifted child's first mentor in music; and at the age of five he composed his first little piece for piano, "Indian Gallop," which was carefully written down by his mother. Later, there appeared a whole series of "little pieces" for piano and the first naïve attempts at composing an opera, under the influence of those he had seen on a trip to Moscow. A high evaluation was put upon the boy's talent by a Moscow composer and teacher, Sergey Taneyev, on whose recommendation the Russian composer Reinhold Glier twice went to Sontsovka in the summer months to become young Sergey's first teacher in theory and composition and to prepare him for entrance into the conservatory at St. Petersburg (later Petrograd, now Leningrad), Russia's foremost conservatory. The years Prokofiev spent there—1904 to 1914—were a period of swift creative growth. His teachers, who were struck by the originality of his musical thinking, included such composers as Nikolay Rimski-Korsakov, Anatoly Lyadov, and Nikolay Tcherepnin. When he graduated in 1914 he was awarded the Anton Rubinstein Prize in piano for a brilliant performance of his own first large-scale work—the *Piano Concerto No. I in D Flat Major,* Opus 10.

The conservatory gave Prokofiev a firm foundation in the academic fundamentals of music, but he avidly sought musical novelty, studying the works of such innovative confemporaries as the French composers Claude Debussy and Maurice Ravel, the Germans Richard Strauss and Max Reger, and a fellow Russian, Aleksandr Scriabin. His enthusiasms were supported by progressive circles advocating musical renewal. Prokofiev's first public appearance as a pianist took place before such a group in St. Petersburg in 1908. A little later he met with friendly sympathy in a similar circle in Moscow, which helped him make his first appearances as a composer, at the Moscow summer symphony seasons of 1911 and 1912. The Moscow journal *Muzyka* expressed immediate enthusiasm for his work.

Debuts as pianist and composer

Prokofiev's talent developed rapidly as he applied many new musical ideas. Unlike others who were similarly disenchanted with late Romanticism and Russian academicism, he did not become a follower of either the ecstatic symbolism of Scriabin or the Impressionism of Debussy. Instead, he studied the compositions of Igor Stravinsky, particularly the early ballets, but maintained a critical attitude toward his countryman's brilliant innovations.

Contacts with the then new currents in theatre, poetry, and painting also played an important role in Prokofiev's development. He was attracted by the work of modernist Russian poets; by the painting of the Russian followers of Cézanne and Picasso; and by the theatrical ideas of Vsevolod Meyerhold, whose experimental productions were directed against an obsolescent naturalism. In 1914 Prokofiev became acquainted with the great ballet impressario Sergey Diaghilev, who commissioned a ballet from him and became one of his most influential advisors for the next decade and a half.

After the death of his father in 1910, Prokofiev lived under more straitened material conditions, though his mother provided for his continuing studies. On the eve of World War I, he visited London and Paris to acquaint himself with the newest in art. The tense pre-storm atmosphere that pervaded Russia sharpened in him a feeling of skepticism, of disbelief in romantic ideals, but did not shake his essentially healthy outlook on life. Exempt from war mobilization as the only son of a widow, Prokofiev continued to perfect his musicianship on the organ and appeared in concerts in the capital and elsewhere. His youthful and mischievous music disturbed academicians and aesthetes but won the support of his most perspicacious contemporaries, including the authors Maksim Gorky and Vladimir Mayakovsky. Concert agencies began to organize his sensational composer's recitals. The pre-Revolutionary period of Prokofiev's work was marked by intense exploration. The harmonic thought and design of his work grew more and more complicated. Prokofiev wrote the ballet *Ala and Lolli* (1914), on themes of ancient Slav mythology, for Diaghilev, who rejected it. Thereupon, Prokofiev reworked the music into the *Scytlzian Suite,* Opus 20, for orchestra. Its premiere in 1916 caused a scandal but was the culmination of his career in Petrograd. The ballet *The Tale of the Buffoon Who Outjested Seven Buffoons* (1915; *The Buffoon,* 1915), also commissioned by Diaghilev, was based on a folktale; it served as a stimulus for Prokofiev's searching experiments in the renewal of Russian music. Despite Diaghilev's assertion of the priority of ballet over opera, which he considered a dying genre, Prokofiev was active in the field of opera. Following the immature *Maddalena,* which he wrote in 1911–13, he composed in 1915–16 *The Gambler,* a brilliant and dynamic adaptation of the story by Dostoyevsky. Continuing the operatic tradition of Mussorgsky, Prokofiev skillfully combined subtle lyricism, satiric malice, narrative precision, and dramatic impact. During this period, Prokofiev achieved great recognition for his first two piano concertos—the first the one-movement *Concerto in D Flat Major* (1911), and the second the dramatic four-movement *Concerto in G Minor* (1913).

Ballets for Diaghilev

The year 1917—the year of two Russian revolutions—was astonishingly productive for Prokofiev. When the Tsar was overthrown in February 1917 he was in the streets of Petrograd, expressing the joy of victory. Perhaps as if inspired by feelings of social renewal, he wrote within one year an immense quantity of new music: two sonatas, the *Violin Concerto No. I in D Major,* the *Classical Symphony,* and the choral work *Seven, They are Seven;* he began the magnificent *Piano Concerto No. 3 in C Major,* Opus 26, and planned a new opera—*The Love for Three Oranges,* after a comedy tale by the Italian dramatist Carlo Gozzi. In the summer of 1917 Prokofiev was included in the Council of Workers in the Arts, which led Russia's left wing of artistic activity; but for almost nine months he was stranded in the Caucasus, cut off from Petrograd by the civil war. Only in the spring of 1918 did he succeed in returning there. In the difficult circumstances of these years, however, he concluded that music had no place and decided to leave Russia temporarily to undertake a concert tour abroad. Prokofiev travelled, with official sanction, over the difficult route through Siberia, where civil strife was raging.

Effects of the 1917 revolutions

Foreign period. The next decade and a half are commonly called the "foreign period" of Prokofiev's work. For a number of reasons, chiefly the continued blockade of the Soviet Union, he could not return at once to his homeland. Nevertheless, he did not lose touch with Russia.

The first five years of Prokofiev's life abroad are usually characterized as the "years of wandering." On the way from Vladivostok to San Francisco, in the summer of 1918, he gave several concerts in Tokyo and Yokohama. In New York City the sensational piano recitals of the "Bolshevik Pianist" evoked both delight and denunciation. The composer had entrée to the Chicago Opera Association, where he was given a commission for a comic opera. The conductor and the producer of the opera, both Italian, gladly backed the idea of an opera on the Gozzi plot. Accordingly, *The Love for Three Oranges* was completed in 1919, though it was not produced until 1921. Within a few years the opera was also produced with immense success on the stages of the U.S.S.R. as well as in western Europe.

In America, Prokofiev met a young singer of Spanish extraction, Lina Llubera, who eventually became his first wife and the mother of two of his sons, Svyatoslav and Oleg.

Not meeting with continuing support in the United States, the composer set out in the spring of 1920 for Paris for meetings with Diaghilev and the conductor Serge Koussevitzky. They soon secured for him wide recognition in the most important west European musical centres. The production of *The Buffoon* by Diaghilev's ballet troupe in Paris and London in 1921 and the Paris premiere of the *Scythian Suite* in 1921 and that of *Seven, They Are Seven* in 1924 evoked enormous interest, consolidating his reputation as a brilliant innovator. The successful performance of the *Third Piano Concerto* (1921), completed in France, also marked one of the peaks of Prokofiev's dynamic national style.

Triumph in Paris

Prokofiev spent more than a year and a half in 1922–23 in southern Germany, in the town of Ettal in 'the

Bavarian Alps. Resting after fatiguing premieres and reviewing the course of his creative path, he also prepared many of his compositions for the printer. The attention of the composer at this period turned to the basic conception of the opera *The Flaming Angel,* after a story by the contemporary Russian author Valery Bryusov, depicting the savage "witch trials" of 16th-century Germany. In complete contrast to the witty eccentricity of the Gozzi story, he now selected a story of great human passions expressed in realistic drama. The opera, which required many years of work (1919–27), did not find a producer within Prokofiev's lifetime.

Meanwhile, Prokofiev, finding himself not interested in the musical activity in Germany, settled in Paris in the autumn of 1923. There he was in close touch with progressive French musical figures, such as the composers Francis Poulenc and Arthur Honegger, while continuing his own intensive creative activity. Vexed by criticisms of his melodically lucid *First Violin Concerto,* which had its premiere in Paris in 1923, he addressed himself to a search for more avant-garde style. These tendencies appeared in such compositions of the early 1920s as the epic *Symphony No. 2 in D Minor,* Opus 40, commissioned by Koussevitzky. Its intense dramatic quality and its striking sense of proportion are also found in the *Symphony No. 3 in C Minor* (1928), based on thematic material from the opera *The Flaming Angel,* reworked by the composer. In close collaboration with Diaghilev, Prokofiev created new one-act ballets, *Le Pas d'acier* (performed in 1927) and *The Prodigal Son* (performed in 1929). *Le Pas d'acier* had a sensational success in Paris and London, thanks to its original staging and bold evocation of images of Soviet Russia at the beginning of the 1920s—with its economic dislocation and the beginnings of industrialization. *The Prodigal Son* had a lofty biblical theme and music that was exquisitely lyrical. It reflects a striving toward emotional relaxation and toward clarification of style that are also seen in the *String Quartet No. I in B Minor,* Opus 50, in the *Sonata for Two Violins in C Major,* Opus 56 (1932), and in the ballet *On the Dnieper.*

In 1927 Prokofiev toured the U.S.S.R. and was rapturously received bv the Soviet public as a world-renowned Russian musician-revolutionary. While there, he strengthened his old associations with the innovative theatrical producer Vsevolod Meyerhold, who helped him in a basic revision of the opera *The Gambler,* produced in 1929 in Brussels.

During the 1920s and early 1930s, Prokofiev toured with immense success as a pianist in the great musical centres of western Europe and the United States. His U.S. tours in 1925, 1930, and 1933 were attended with tumultuous success and brought him new commissions, such as the *Symphony No.* 4 *in C Major,* Opus 47 (1930), incorporating musical material of *The Prodigal* Son, for the 50th anniversary of the Boston Symphony, and the *First String Quartet,* commissioned for the Library of Congress. His new piano concertos—No. 4, Opus 53 (1931), for the left hand, and *No.* 5 *in* G *Major,* Opus 55 (1932)—demonstrated anew his bent for impulsiveness and virtuoso brilliance.

Soviet period. Although he enjoyed material well-being, success with the public, and contact with outstanding figures of Western culture, such as Stravinsky, Chaplin, and Picasso, Prokofiev increasingly missed his homeland. Visits to the U.S.S.R. in 1927, 1929, and 1932 led him to the decision to conclude his foreign obligations and return to Moscow once and for all. From 1933 to 1935 the composer gradually accustomed himself to the new conditions and became one of the leading figures of Soviet culture. He established close connections with theatres and cinema studios and received interesting commissions, while not abandoning concert tours abroad until 1938. He valued the opportunity for contact and collaboration with eminent figures in Soviet art—the film maker Sergey Eisenstein, the ballerina Galina Ulanova, the pianist Svyatoslav Richter, the writers Ilya Ehrenburg and Aleksey Tolstoy, and others.

'ollab-
ration
'ith
ther
rtists

In the two decades constituting the "Soviet period" of Prokofiev's work—1933 to 1953—the realistic and epical traits of his art became more clearly defined. The synthesis of traditional tonal and melodic means with the stylistic innovations of 20th-century music was more fully worked out.

In the years preceding World War II, Prokofiev created a number of classical masterpieces. These included his *Violin Concerto No. 2 in* G *Minor,* Opus 63 (1935), the ballet *Romeo and Juliet* (1935–36), and the music for Eisenstein's film *Alexander Nevsky* (1938). His work in theatre and the cinema gave rise to a number of charming programmatic suites, such as the *Lieutenant Kije* suite (1934), the *Egyptian Nights* suite (1934), and the symphonic children's tale *Peter and the Wolf* (1936). Turning to opera, he cast in the form of a contemporary drama of folk life his *Semyon Kotko,* depicting events of the civil war in the Ukraine (1939). The basis of the brilliantly modernized opéra bouffe the *Betrothal in a Monastery* (composed in 1940, produced in 1946) was the play *The Duenna,* by the 18th-century British dramatist Richard Brinsley Sheridan. Testing his powers in other genres, he composed the monumental *Cantata for the 20th Anniversary of the October Revolution* (1937), on texts by Lenin, and the epic cantata *The Toast* (1939).

On his last trip abroad, Prokofiev visited Hollywood, where he studied the technical problems of the sound film; the experience thus gained he applied brilliantly in the striking national music for Eisenstein's film *Alexander Nevsky* (1938), depicting the heroic Russian struggle against the Teutonic Knights of the 13th century. The cantata *Alexander Nevsky* was based on the music of the film. One of the summits of Prokofiev's art was the production of his ballet *Romeo and Juliet* in Leningrad, with Ulanova in the leading role. Throughout the 1930s Prokofiev took part in the organizational work of the Composers' Union, made appearances as conductor and as pianist, and travelled much throughout the country.

The war
years

On the eve of World War II, a change occurred in his personal life: leaving his first family, he linked his destiny with that of the poet Mira Mendelssohn, who became his second wife.

The war sharpened Prokofiev's national and patriotic feelings. Regardless of the difficulties of the war years, he composed with remarkable assiduity, even when the evacuation of Moscow in 1941 made it necessary for him and his wife to move from one place to another until they were able to return in 1944.

From the first days of the war, the composer's attention was centred on a very large-scale operatic project: an opera based on Tolstoy's novel *War and Peace.* He was fascinated by the parallels between 1812, when Russia crushed Napoleon's invasion, and the then current situation. The first version of the opera was completed by the summer of 1942, but subsequently the work was fundamentally revised, a task that occupied more than ten years of intensive work. Those who heard it were struck both by the immense scale of the opera (13 scenes, more than 60 characters) and by its unique blend of epic narrative with lyrical scenes depicting the personal destinies of the major characters. An increasing predilection for national-epical imagery is manifested in the heroic majesty of the *Symphony No.* 5 *in B Flat Major,* Opus 100 (1944), and in the music (composed, 1942–45) for Eisenstein's two-part film *Ivan the Terrible* (Part I, 1944; Part II, 1948). Living in the Caucasus, in Central Asia, in the Urals, the composer was everywhere interested in the folklore, an interest reflected in the *String Quartet No. 2 in F Major,* Opus 92, on Kabardinian and Balkarian themes (1941); and in the comic opera *Khan Buzai,* on themes of Kazakh folktales. Documents of those troubled days are three piano sonatas, *No. 6,* Opus 82 (1940), *No.* 7, Opus 83 (1942), and *No.* 8, Opus 84 (1944), which are striking in the dramatic conflict of their images and their irrepressible dynamism. Critics saw in the intense climaxes of the *Seventh Sonata* and the *Fifth Symphony* the spirit of Russia courageously repelling the Nazi assault. Prokofiev spent the summers of 1944 and 1945 near Ivanovo, at a former state poultry farm placed at the disposal of composers during the war; there he worked in close association with Nikolay Myaskovsky, Dmitry Shostakovich, Aram Khachaturian, Dmitry Kabalevsky, and others.

Illness and reproach

Overwork was fatal to the composer's health: in 1945 he became afflicted with a severe form of high blood pressure. In the following years, the illness continually worsened. His doctors strictly limited his work schedule. They even forbade him his favourite diversions—playing chess and driving a car. Nevertheless, the last eight years of his life were, like the preceding ones, filled with work, which continued despite the strong criticism that was directed at some of his compositions in the late 1940s. Unfounded accusations appeared in the press against Prokofiev's work; they were instigated by supporters of formalism and decadent complexity. A decade later, however, these charges were officially repudiated by the Central Committee of the Communist Party.

During the last years of his life, Prokofiev seldom left his villa in a suburb of Moscow. His propensity for innovation, however, is still evident in such important works as the *Symphony No. 6 in E Flat Minor,* Opus 111 (1945–47), which is laden with reminiscence of the tragedies of the war just past; the *Sinfonia Concerto for Cello and Orchestra in E Minor,* Opus 125 (1950–52), composed with consultation from Mstislav Rostropovich; and the *Violin Sonata in F Minor,* Opus 80 (1938–46), dedicated to David Oistrakh, which is in Russian folk imagery. Just as in earlier years, the composer devoted the lion's share of his energy to musical theatre: cases in point were the opera *The Story of a Real Man* (1947–48), the ballet *The Stone Flower* (1948–50), and the oratorio *On Guard for Peace* (1950), which was performed by a mixed chorus and soloists. In these years Prokofiev reworked some of his earlier compositions and created fundamentally new versions of the *Fourth Symphony* (1947) and the *Piano Sonata No. 5 in C Major,* Opus 135 (1952–53). The lyrical *Symphony No. 7 in C Sharp Minor,* Opus 131 (1951–52) was the composer's swan song.

Works left incomplete

On March 5, 1953, Prokofiev died suddenly of cerebral hemorrhage. On his worktable there remained a pile of unfinished compositions, including sketches for a 6th concerto for two pianos, a 10th and an 11th piano sonata, a Kazakh comic opera, and a solo violoncello sonata. The subsequent years saw a rapid growth of his popularity in the Soviet Union and abroad. In 1957 he was posthumously awarded the Soviet Union's highest honour, the Lenin Prize, for the *Seventh Symphony.*

MAJOR WORKS

Orchestral music

SYMPHONIES: *No. 1 in D Major (Classical Symphony),* op. 25 (1916–17); *No. 2 in D Minor,* op. 40 (1924); *No. 3 in C Minor,* op. 44 (1928); *No. 4 in C Major,* op. 47 (1930); *No. 4 in C Major,* op. 112 (1947, revision); *No. 5 in B Flat Major,* op. 100 (1944); *No. 6 in E Flat Minor,* op. 111 (1945–47); *No. 7 in C Sharp Minor,* op. 131 (1951–52).

CONCERTOS: Nine concertos, including *First Piano Concerto in D Flat Major,* op. 10 (1911); *Third Piano Concerto in C Major,* op. 26 (1917–21); *First Violin Concerto in D Major,* op. 19 (1916–17); *Second Violin Concerto in G Minor,* op. 63 (1935); *Cello Concerto in E Minor,* op. 58 (1935–38); *Sinfonia Concerto for Cello and Orchestra in E Minor,* op. 125 (1950–52).

BALLETS: *The Buffoon,*op. 21 (1915, ballet; 1922, symphonic suite); *Le Pas d'acier,* op. 41 (1925); *The Prodigal Son,* op. 46 (1928, ballet; 1929, symphonic suite); *Romeo and Juliet,* op. 64 (1935–36, ballet and suites); *Cinderella,* op. 87 (1940–44); *The Stone Flower* (1948–50, ballet and suites).

OTHER ORCHESTRAL WORKS: *Scythian Suite,* op. 20 (1914–15); *Lieutenant Kije,* op. 60 (1934, suite); *Egyptian Nights,* op. 61 (1934, suite); *Russian Overture,* op. 72 (1936).

Chamber music

Two quartets, a quintet for wind instruments, an Overture-sextet on Jewish themes, a violin sonata, and two sonatas for violin and piano.

Piano works

Numerous compositions for piano, including ten sonatas; *Ten Pieces for Piano,* op. 12 (1906–13); *Sarcasms,* op. 17 (1912–14, five pieces for piano); *Visions Fugitive,* op. 22 (1915–17, 20 pieces for piano); *Tales of the Old Grandmother,* op. 31 (1918, four pieces for piano).

Vocal works

OPERAS: *The Gambler,* op. 24 (1915–16, rev. 1927); *The Love for Three Oranges,* op. 33 (1919); *The Flaming Angel,* op. 37 (1919–27); *Semyon Kotko,* op. 81 (1939); *War and Peace,* op. 91 (1941–52); *Betrothal in a Monastery (The Duenna),* op. 86 (1940); *The Story of a Real Man* (1947–48).

CANTATAS: *Seven, They Are Seven,* op. 30 (1917–18); *Cantata for the 20th Anniversary of the October Revolution,* op. 74 (1937); *Alexander Nevsky,* op. 78 (1938–39); *Zdravitsa* (earlier called *Toast*), op. 85 (1939).

OTHER WORKS FOR VOICE AND ORCHESTRA: *Peter and the Wolf,* op. 67 (1936, symphonic tale); *Winter Bonfire,* op. 122 (1949, suite for narrator, children's choir, and orchestra): *On Guard for Peace,* op. 124 (1950, oratorio).

SONGS: About 60 songs and romances, including "The Ugly Duckling," op. 18 (1914); 12 arrangements of Russian folk songs for voice and piano, 2 vol., op. 104 (1944).

BIBLIOGRAPHY. Two of the most important sources of information are the documentary collections S. *Prokofiev: Autobiography, Articles, Reminiscences,* ed. by S. SHLIFSTEIN (1960; orig. pub. in Russian, 1956; 2nd ed., 1961), and a larger collection of the composer's memoirs, *Prokofiev by Prokofiev* (1979; edited by David H. Appel from the original compilation edited by M. Kozlova). Although in the years since the death of the composer an extensive musicological literature concerning him has appeared, very little has been published in English. ISRAEL NESTYEV, *Prokofiev* (1960; orig. pub, in Russian, 1957; new ed., 1973), is the first serious work to systematically study the man and his music. A popular biographical and critical study is R. HOFMANN. *Serge Prokofiev: l'homme et son oeuvre* (1964).

In Russian, the following analytical studies are recommended: СЕРГЕЙ М. СЛОНИМСКИЙ, *Симфонии Прокофьева. Опыт 'исследования* (1964), on the symphonies; МАРИНА Д. САБИНИНА, *«Семен Котко» и проблемы оперной драматургии Прокофьева* (1963), on the dramaturgy of the operas; and В.Н. ХОЛОПОВА and Ю.Н. ХОЛОПОВ, *Фортепьянные сонаты С.С. Прокофьева* (1961), on the contemporary characteristics of Prokofiev's harmony.

(I.V.N.)

Propaganda

Propaganda is the more or less systematic effort to manipulate other people's beliefs, attitudes, or actions by means of symbols (words, gestures, banners, monuments, music, clothing, insignia, hairstyles, designs on coins and postage stamps, and so forth). Deliberateness and a relatively heavy emphasis on manipulation distinguish propaganda from casual conversation or the free and easy exchange of ideas. The propagandist has a specified goal or set of goals. To achieve these he deliberately selects facts, arguments, and displays of symbols and presents them in ways he thinks will have the most effect. To maximize effect, he may omit pertinent facts, and he may try to divert the attention of the reactors (the people whom he is trying to sway) from everything but his own propaganda.

Propaganda distinguished from education

Comparatively deliberate selectivity and manipulation also distinguish propaganda from education. The educator tries to present various sides of an issue—the grounds for doubting as well as the grounds for believing the statements he makes, and the disadvantages as well as the advantages of every conceivable course of action. Education aims to induce the reactor to collect and evaluate evidence for himself and assists him in learning the techniques for doing so. It must be noted, however, that a given propagandist may look upon himself as an educator, may believe that he is uttering the purest truth, that he is emphasizing or distorting certain aspects of the truth only to make a valid message more persuasive, and that the courses of action that he recommends are in fact the best actions that the reactor could take. By the same token, the reactor who regards the propagandist's message as self-evident truth may think of it as educational; this often seems to be the case with "true believers"—dogmatic reactors to dogmatic religious or social propaganda. "Education" for one person may be "propaganda" for another.

PROPAGANDA AND RELATED CONCEPTS

Connotations of the term propaganda. The word propaganda itself, as used in recent centuries, apparently derives from the title and work of the Congregatio de Propaganda Fide (Congregation for Propagation of the Faith), an organization of Roman Catholic cardinals founded in 1622 to carry on missionary work. To many Roman Catholics the word may therefore have, at least in missionary or ecclesiastical terms, a highly respectable connotation. But even to these persons. and certainly to many others. the term is often a dirty one tending to connote such things as the discredited atrocity stories and deceptively stated war aims of World Wars I and II, the operations of the Nazis' Ministry of Public Enlightenment and Propaganda, and the broken campaign promises of a thousand politicians. Also, it is reminiscent of countless instances of false and misleading advertising (especially in countries using Latin languages, in which *piopagande commerciale* or some equivalent is a common term for commercial advertising).

Agitprop: agitation and propaganda

To informed students of Communism, the term propaganda has yet another connotation, associated with the term agitation. The two terms were first used by the Marxist Ceotgy Plekhanov and later elaborated upon by Lenin in a pamphlet *What* Is *To Be Done?* (1902), in which he defined "propaganda" as the reasoned use of historical and scientific arguments to indoctrinate the educated and enlightened (the attentive and informed publics, in the language of today's social sciences); he defined "agitation" as the use of slogans, parables, and half-truths to exploit the grievances of the uneducated and the unreasonable. Since he regarded both strategies as absolutely essential to political victory, he twinned them in the term agitprop. Today every unit of a Communist party must have an agitprop section, and to the Communist, the use of propaganda in Lenin's sense is commendable and honest. Thus, a standard Soviet manual for teachers of social sciences is entitled *Propagandistu politekonoinii (For the Propagandist of Political Economy)*, and a pocket-sized booklet issued weekly to suggest timely slogans and brief arguments to be used in speeches and conversations among the masses is called *Bloknot agitatora* (*The Agitator's Notebook).*

Related terms. Related to the word propaganda, as it is more generally used, is the concept of "propaganda of the deed." This denotes taking nonsymbolic action (such as economic or coercive action), not for its economic or coercive effects but for its possible propagandistic effects. Examples of propaganda of the deed would include staging an atomic "test" or the public torture of a criminal for its presumable deterrent effect on others, or giving foreign "economic aid" primarily to influence the recipient's opinions or actions and without much intention of building up the recipient's economy.

Distinctions are sometimes made between overt propaganda, in which the propagandist and perhaps his backers are made known to the reactor, and covert propaganda, in which the source is secret or disguised—a "hidden persuader." Covert propaganda might include such things as unsigned political advertisements, clandestine radio stations using false names, and statements by editors, politicians, or others who have been secretly bribed by governments, political backers, or business firms. Sophisticated diplomatic negotiation, legal argument, collective bargaining, commercial advertising, and political campaigns are of course quite likely to include considerable amounts of both overt and covert propaganda and to be accompanied by propaganda of the deed.

Another term related to propaganda is psychological warfare (sometimes abbreviated to "psychwar"), which is the prewar or wartime use of propaganda directed primarily at confusing or demoralizing enemy populations or troops, putting them off guard in the face of coming attacks, or inducing them to surrender.

Still another related concept is that of brainwashing. This term usually means intensive political indoctrination. It may involve long political lectures or discussions, long compulsory reading assignments, and so forth, sometimes in conjunction with efforts to reduce the reactor's resistance by exhausting him either physically through torture, overwork, or denial of sleep or psychologically through solitary confinement, threats, emotionally disturbing confrontations with interrogators or defected comrades, humiliation in front of fellow citizens, and the like. The term brainwashing has been widely used in sensational journalism to refer to Such activities (and to many other activities) when they have allegedly been conducted by Maoists in China and elsewhere.

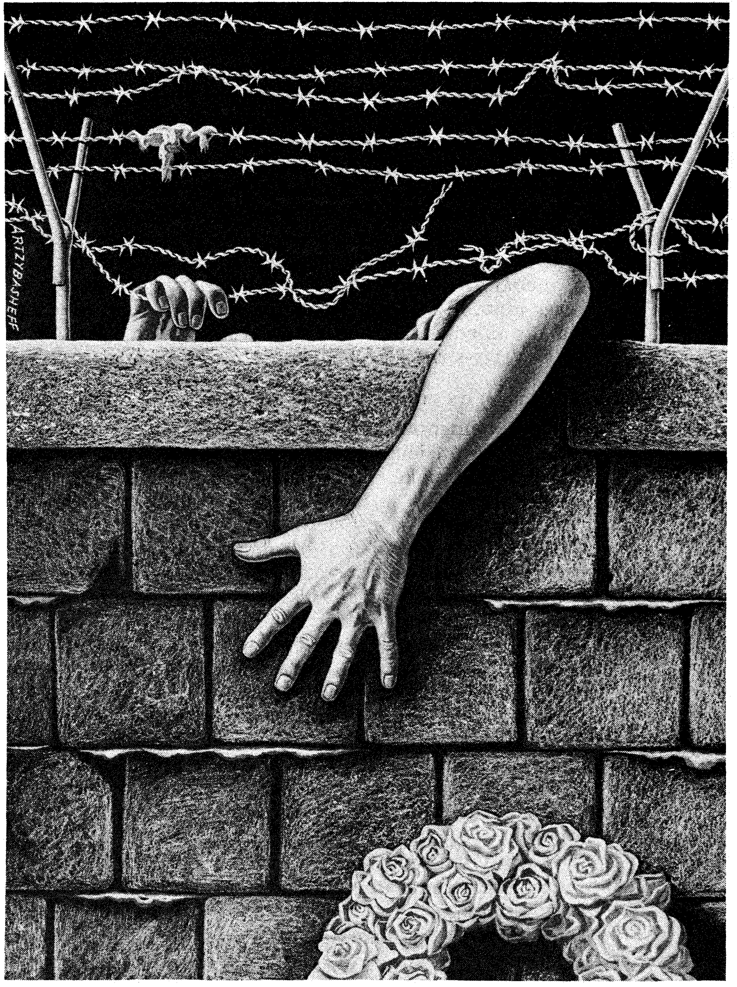
Another related word, advertising, has mainly commercial connotations, though it need not be restricted to this; political candidates, party programs, and positions on political issues may be "packaged" and "marketed" by advertising firms. The words promotion and public relations have wider, vaguer connotations and are often used to avoid the implications of "advertising" or "propaganda." "Publicity" and "publicism" often imply merely making a subject known to a public, without educational, propagandistic, or commercial intent.

Signs, symbols, and media used in contemporary propaganda. The 20th-century propagandist with money and imagination can use a very wide range of signs, symbols, and media to convey his message. Signs are simply stimuli—"information bits" capable of stimulating, in some way, the human organism. These include sounds, such as words. music, or a 21-gun salvo; gestures (a military salute, a thumbed nose); postures (a weary slump, folded arms, a sit-down, an aristocratic bearing); structures (a monument, a building); items of clothing (a uniform, a civilian suit); visual signs (a poster, a flag, a picket sign, a badge, a printed page, a commemorative postage stamp, a swastika scrawled on a wall); and so on and on.

The nature of symbols

A symbol is a sign having a particular meaning for a given reactor. Two or more reactors may of course attach quite different meanings to the same symbol. Thus, to Nazis the swastika was a symbol of racial superiority and the crushing military might of the German folk; to some

Time cover by Artzybasheff, © 1962, Time Inc.



The cover of Time, August 31, 1962, exemplifies propaganda against the construction of the Berlin wall by East Germany.

Asiatic and North American peoples it is a symbol of universal peace and happiness. Some Christians who find a cross reassuring may find a hammer and sickle displeasing and may derive no religious satisfaction at all from a Muslim crescent, a Hindu cow, or a Buddhist lotus.

The contemporary propagandist can employ elaborate social-scientific research facilities, unknown in previous epochs, to conduct opinion surveys and psychological interviews in efforts to learn the symbolic meanings of given signs for given reactors around the world and to discover what signs leave given reactors indifferent because, to them, these signs are without meaning.

Media are the means—the channels—used to convey signs and symbols to the intended reactor or reactors. A comprehensive inventory of media used in 20th-century propaganda could cover many pages. Written media include letters, handbills, posters, billboards, newspapers, magazines, books, and handwriting on walls and streets. Among audiovisual media, television may be the most powerful for many purposes. Television can convey a great many types of signs simultaneously; it can gain heavy impact from mutually reinforcing gestures, words, postures, and sounds and a background of symbolically significant leaders, celebrities, historic settings, architectures, flags, music, placards, maps, uniforms, insignia, cheering or jeering mobs or studio audiences, and staged assemblies of prestigious or powerful people. Other audiovisual media include public speakers, motion pictures, theatres, marching bands, mass demonstrations, picketing, face-to-face conversations between individuals, and "talking" exhibits at fairs, expositions, and art shows.

The nature of organizational media

The larger the propaganda enterprise, the more important are such mass media as television and the press and also the organizational media—that is, pressure groups set up under leaders and technicians who are skilled in using many sorts of signs and media to convey messages to particular reactors. Vast systems of diverse organizations can be established in the hope of reaching leaders and followers of all groups (organized and unorganized) in a given area, such as a city, region, nation or coalition of nations, or the entire world. Pressure organizations are especially necessary, for example, in closely fought sales campaigns or political elections, especially in socially heterogeneous areas that have extremely divergent regional traditions, ethnic and linguistic backgrounds, and educational levels and very unequal income distributions. Diversities of these sorts make it necessary for products to be marketed in local terms and for political candidates to appear to be friends of each of perhaps a dozen or more mutually hostile ethnic groups, of the educated and the uneducated, and of the very wealthy as well as the poverty-stricken.

EVOLUTION OF THE THEORY OF PROPAGANDA

Early commentators and theories. The archaeological remains of ancient civilizations indicate that dazzling clothing and palaces, impressive statues and temples, magic tokens and insignia, and elaborate legal and religious arguments have been used for thousands of years, presumably to convince the common people of the purported greatness and supernatural prowess of kings and priests. Instructive legends and parables, easily memorized proverbs and lists of commandments (such as the Analects of Confucius, the Judaic Ten Commandments, the Hindu Laws of Manu, the Buddhists' Eightfold Noble Path), and highly selective chronicles of rulers' achievements have been used to enlist mass support for particular social and religious systems. Very probably, much of what was said in antiquity was sincere, in the sense that the underlying religious and social assumptions were so fully accepted that the warlords' spokesmen, the pharaohs' priests, and their audiences believed all or most of what was communicated and hence did not deliberate or theorize very much about alternative arguments or means of persuasion.

The systematic, detached, and deliberate analysis of propaganda, in the West, at least, may have begun in Athens about 500 BC, as the study of rhetoric (Greek: "the technique of orators"). The tricks of using sonorous and solemn language, carefully gauged humour, artful congeniality, appropriate mixtures of logical and illogical argument, and flattery of a jury or a mob were formulated from the actual practices of successful lawyers, demagogues, and politicians. Relatively ethical teachers such as Isocrates, Plato, and Aristotle compiled rules of rhetoric (1) to make their own arguments and those of their students more persuasive and (2) to design counterpropaganda against opponents and also (3) to teach their students how to detect the logical fallacies and emotional appeals of demagogues.

The Greco-Roman tradition of rhetoric

Early students of rhetoric also examined what today's analysts would call the problem of source credibility—what a speaker can say or do to convince his hearers that he is telling the truth, is well intentioned, is public-spirited, and so forth. For example, an Athenian lawyer defending an undersized man on trial for murder might instruct him to say to a jury: "Is it likely that an undersized man like me, so often ridiculed for being clumsy with a sword, would have attacked and killed this very tall war veteran who is famous everywhere for his swordsmanship?" But a tall and strong defendant might be told to invert the plea: "Would any man of my unusual height, who is rather well known to have slain 300 Persians in sword fights, have allowed himself to be drawn into a quarrel with this puny man—knowing full well that a jury of reasonable Athenians would be inclined from the start to hold me guilty if someone killed him?" So well did Greek rhetoricians analyze the arts of legal sophistry and political demagoguery that their efforts were imitated and further developed in Rome by such figures as Cicero and Quintilian. Aristotle's Rhetoric and similar works by others have, indeed, served as model texts for Western scholars and students until this day.

Hindu and Chinese traditions

There have been similar lines of thought in other major civilizations. In ancient India, the Buddha, and in ancient China, Confucius, both advocated, much as Plato had, the use of truthfulness, "good" rhetoric, and "proper" forms of speech and writing as means of persuading men, by both precept and example, to live the good life. Toward 400 BC in India, Kauṭilya, a Brahmin believed to have been chief minister to the emperor Candragupta Maurya, reputedly wrote the *Arthaśāstra* (Principles of Politics), a book of advice for rulers that has often been cempared with Plato's Republic and Machiavelli's much later work The Prince. Kauṭilya discussed, in some detail, the use of psychological warfare, both overt and clandestine, in efforts to disrupt an enemy's army and capture his capital. Overtly, he said, the propagandists of a king should proclaim that he can do magic, that God and the wisest men are on his side, and that all who support his war aims will reap benefits. Covertly, his agents should infiltrate his enemies' and potential enemies' kingdoms, spreading defeatism and misleading news among their people, especially in capital cities, among leaders, and among the armed forces. In particular, a king should employ only Brahmins, unquestionably the holiest and wisest of men, as propagandists and diplomatic negotiators. These morally irreproachable experts should cultivate the goodwill of their king's friends, and of friends of his friends, and also should woo the enemies of his enemies. A king should not hesitate, however, to break any friendships or alliances that are later found to be disadvantageous.

Similar advice is found in The Art of War by the Chinese theorist Sun Tzu, who wrote at about the same time. "All warfare," he said, "is based on deception. Hence, when able to attack, we must seem unable; when using our forces, we must seem inactive; when we are near, we must make the enemy believe that we are far away; when far away, we must make him believe we are near. Hold out baits to entice the enemy. Feign disorder, and crush him."

The spread of all complex political systems and religions probably has been due very largely to a combination of earnest conviction and the deliberate use of propaganda. This mixture can be detected in the recasting in various times and places of the legends of the Judaeo-Christian messiah, of heroes of the Hindu Mahdbhdrata.

of the Buddha, of the ancestral Japanese Sun Goddess, of the lives of Muhammad and his relatives, of the Christian saints, of such Marxist heroes as Marx, Engels, Lenin, and Stalin, and even in the story of George Washington and the cherry tree.

Scattered and sometimes enlightening comment on political and religious propaganda has occurred in all major civilizations. In ancient Greece and Rome there was much writing on election tactics. In 16th-century Italy, Machiavelli discussed, very much like Kauṭilya and Sun Tzu, the uses of calculated piety and duplicity in peace and war. In Shakespeare's plays, Mark Antony and the Duke of Buckingham display the principles of propaganda and discuss them in words and concepts that anticipate the present-day behavioral scientist (see *Julius Caesar*, Act III and *Richard III*, Act III). They refer to such propaganda stratagems as the seizure and monopolization of propaganda initiatives, the displacement of guilt onto others (scapegoating), the presentation of oneself as morally superior, and the coordination of propaganda with violence and bribery.

Modern research and the evolution of current theories. After the decline of the ancient world, no elaborate systematic study of propaganda appeared for centuries—not until the Industrial Revolution had brought about mass production and raised hopes of immensely high profits through mass marketing. Toward the beginning of the 20th century, researchers began to undertake studies of the motivations of many types of consumers and of their responses to various kinds of salesmanship, advertising, and other marketing techniques. From the early 1930s on, there have been "consumer surveys" much in the manner of public-opinion surveys. Almost every conceivable variable affecting consumers' opinions, beliefs, suggestibilities, and behaviour has been investigated for every kind of group, subgroup, and culture in the major capitalist nations. Consumers' wants and habits are beginning to be studied in the same ways in the socialist countries—partly to promote economic efficiency and partly to prevent political unrest. Data on the wants and habits of voters as well as consumers are now being gathered in the same elaborate ways in many parts of the world.

Sample surveys and market research

Large quantities of such information on consumers and voters are stored and statistically processed by computers and are drawn upon for nationwide and international advertising campaigns costing billions of dollars annually. Such advertising—including political advertising—occupies a very high percentage of radio and television time and of newspaper, magazine, and billboard space in countries where it is permitted. By conservative estimates $140,000,000 was spent in the U.S. presidential election of 1952, $155,000,000 in that of 1956, $175,000,000 in 1960, and $200,000,000 in 1964. On radio and television alone the Republican Party was estimated to have spent more than $12,000,000 and the Democratic Party over $7,100,000, for their presidential and vice presidential candidates in 1968. Critics have argued that advertising expenditures on such a scale, whether for deodorants or presidents, tend to waste society's resources and also to preclude effective competition by rival producers or politicians who cannot raise equally large amounts of money. A rising tide of consumer resistance and voter skepticism is leading to various attempts at consumer education, voter education, counterpropaganda, and proposals for regulatory legislation.

As far back as the early 1920s, there developed an awareness among many social critics that the extension of the vote and of enlarged purchasing power to more and more of the ignorant or ill-educated meant larger and larger opportunities for both demagogic and public-spirited propagandists to make headway by using fictions and myths, utopian appeals, and "the noble lie." Interest was aroused not only by the lingering horror of World War I and of the postwar settlements but also by publication of Ivan Pavlov's experiments on conditioned reflexes and of analyses of human motivations by various psychoanalysts. Freud's *Group Psychology and the Analysis of the Ego* (1922) was particularly relevant to the study of leaders, propagandists, and followers, as were Walter Lippmann's *Public Opinion* (1922) and *The Phantom Public* (1925).

In 1927, an American political scientist, Harold D. Lasswell, published a now-famous book, *Propaganda Technique in the World War*, a dispassionate description and analysis of the massive propaganda campaigns conducted by all the major belligerents in World War I. This he followed with studies of Communist propaganda and of many other forms of communication. Within a few years, a great many other social scientists, along with historians, journalists, and psychologists, were producing a wide variety of publications purporting to analyze military, political, and commercial propaganda of many types. During the Nazi period and the period of World War II and the subsequent cold war between the U.S. and the Soviet Union, a great many researchers and writers, both skilled and unskilled, scholarly and unscholarly, were employed by governments, political movements, and business firms to conduct propaganda. Some of those who had scientific training designed very carefully controlled experiments or Intelligence operations, attempting to quantify data on appeals of various types of propaganda to given reactors.

In the course of this theory building and research, the study of propaganda advanced a long way on the road from lore to science. Today several hundred more or less scholarly books and thousands of articles shed substantial light on the psychology, techniques, and effects of propaganda campaigns, major and minor.

Propaganda specialists

In recent decades, nearly every significant government, political party, special-interest group, social movement, and big business firm in the advanced countries has developed its own corps of specialized researchers, propagandists, or "opinion managers" (sometimes referred to as information specialists, lobbyists, legislative representatives, or vice presidents in charge of public relations). Some have become members of parliaments, cabinets, and corporate boards of directors. The most expert among them sometimes are highly skilled or trained, or both, in history, psychiatry, politics, social psychology, survey research, and statistical inference.

Many of the bigger and wealthier propaganda agencies conduct (overtly and covertly) elaborate observations and opinion surveys, among samples of the leaders, the middle strata, and the rank and file of all social groups, big and little, whom they hope to influence. They tabulate many kinds of data concerning those contents of the press, films, television, and organizational media that reach given groups. They chart the responses of reactors, through time, by statistical formulas. They conduct "symbol campaigns" and "image-building" operations with mathematical calculation, using quantities of data that can be processed only by computers. To the ancient art of rhetoric, the "technique of orators," have been added the techniques of the psychopolitical analyst and the media man and the know-how of the administrators of giant advertising agencies, public relations firms, and governmental ministries of information that employ armies of analytic specialists and "symbol-handlers."

It is a commonplace among the highly educated that men in the mass—and even men on high educational and social levels—often react more favourably to utopian myths, wishful thinking, and nonrational residues of earlier experiences than they do to the sober analysis of facts. The average citizen who may be aware of being duped is not likely to have enough education, time, or economic means to defend himself against the massive organizations of opinion managers and hidden persuaders. Indeed, to affect them he would have to act through large organizations himself and to use, to some extent, the very means used by those he seeks to control. The still greater "curse of bigness" that may evolve in the future is viewed with increasing concern by many politically conscious people.

THE ANALYSIS OF CONTEMPORARY PROPAGANDA

The components of propaganda. The contemporary propagandist employing behavioral theory tends to analyze his problem in terms of at least ten questions:

1. What are the goals of the propaganda? (What

changes are to be brought about? In whom? And when?)

2. What are the present and expected conditions in the world social system?

3. What are the present and expected conditions in each of the subsystems of the world social system (such as international regions, nations, lesser territories, interest groups)?

4. Who should distribute the propaganda—the propagandist or his agents?

5. What symbols should be used?

6. What media should be used?

7. Which reactors should the propaganda be aimed at?

8. How can the effects of the propaganda be measured?

9. By what countermeasures can opponents neutralize or suppress the propaganda?

10. How can such countermeasures be measured and dealt with?

In the present state of social science, this ten-part problem can be solved with only moderate confidence with respect to any really major propaganda campaign, even if one has a great deal of money for research. Yet if the propagandist is to proceed as rationally as possible, he needs the best answers that are available.

Goals. Goals are fairly easy to define if the propagandist simply wants to sell a relatively safe, useful, and simple good or service. When the propagandist aims to convert great numbers of people to a religion or a new social order or to induce extremely dangerous collective action like a waɪ or revolution, however, the definition of goals becomes highly complex. It is complicated further by problems about "means–goals" or intermediate goals: probably the campaign will have to go on for a long time and will have to be planned in stages, phases, or waves. The propagandist may find it hard to specify, even to himself, exactly what beliefs, values, or actions he wants to bring about, by what points in time, among different sorts of people. Very large and firmly held complexes of values are involved, such as prestige, peace of mind, income, and even life itself or the military security of entire nations or regions––even, in modern times, the annihilation of all mankind. In such a situation, a mass of intricate and thorny value dilemmas arises: Is military or revolutionary victory worth the price of economic ruin? Can a desired degree of individual liberty be achieved without too much loss of social equality? Is a much quicker achievement of goals worth a much greater amount of human suffering? Are war crimes to be committed in order to win a battle? In short: What is the propagandist willing to risk, for what, across what periods of time?

Particularism versus universalism

Present and expected conditions in the world social system. Under modern conditions, each act of propaganda is apt to have effects in several parts of the world. Some of these may boomerang unexpectedly against the propagandist himself unless he can visualize the global system and its components and anticipate the problems that may arise. The global system, moreover, is inexorably changing. As population, trade, travel, education, and technology evolve, new centres of political, cultural, and economic power emerge. This social evolution, extremely rapid in current times, tends on balance to limit the use of more simplistic and parochial kinds of propaganda and increases the need for more sophisticated, scientifically formulated, and univeisalistic (world-oriented) types. If, for example, there is, as some theorists argue, an evolution everywhere from less rationality and scientism toward more and from the primacy of particularistic loyalties toward the primacy of a universalistic loyalty, is the propagandist to use appeals that resist such trends or accept them? If he resists, what is the cost? If his appeals are far ahead of his time, again what is the cost?

Present and expected conditions in subsystems. In many times and places in the past, the propagandist could profit handsomely by ignoring the welfare of a nation or the world and appealing to extremes of religious, racial, political, or economic fanaticism. This paid off very well, in the short run at least, within many subsystems. Today, however, this kind of propaganda can prove to be useless and even dangerous. The prudent propagandist has therefore to decide what mix of universalistic and particularistic symbolism will best serve his purposes at given times in given places. The choice is never an easy one: parochial or class-conscious or national groups may be aroused to the highest passions; and they are numerous and diverse and often highly incompatible with one another and with the imperatives of the nation or the world.

Propaganda "fronts"

The propagandist and his agents. The use of seemingly reputable, selfless, or neutral agents or so-called front organizations, while the propagandist himself remains behind the scenes, may greatly improve his prospects. If the authorities are after the propagandist, seeking to suppress his activities, he ***must*** stay underground and work through agents. But even in freer circumstances, he may wish someone else to speak for him. The propagandist, for instance, may not speak the reactors' language or idiom fluently. He may not know what they associate with given symbols. Or their cultural, racial, or religious feelings may bias them against him and thus tend to deny him a favourable hearing. In such cases the use of agents is inescapable. Thus, subsidizing a native news commentator or lecturer in a foreign country or furnishing propagandistic music for use by a foreign broadcasting station may be more effective than conducting one's own broadcasts. (There are exceptions, however. Many surveys have shown, for example, that news broadcasts by the British Broadcasting Corporation are considered by various foreign audiences to be more truthful than broadcasts originating in their own countries.) Furthermore, if the propaganda fails or is exposed for what it is, the agent can be publicly scapegoated while the real propagandist continues to operate and develop new stratagems. The prince, said Machiavelli, may openly and conspicuously bestow awards and honours and public offices; but he should have his agents carry out all actions that make a man unpopular, such as punishments, denunciations, dismissals, and assassinations.

A complicated modem campaign on a major scale is likely to be planned most successfully by a collective leadelship—a team of broadly educated and skilled people who have had both practical experience in public affairs and extensive training in history, psychology, and the social sciences. The detachment, skepticism, and secularism of such persons may, however, cause them to be viewed with great suspicion by many reactors. It may be important, therefore, to keep the planners behind the scenes and to select intermediaries, front men, Trojan horses, and "dummy leaders" whom the reactors are more likely to listen to or appreciate.

Image makers

Contemporary social-psychological research, dating from Freud's ***Group Psychology and the Analysis of the Ego,*** makes clear the wisdom of traditional insights concerning the supreme importance of leadership in any group, be it the family, the nation, or the world social system. The rank and file of any group, especially a big one, have been shown to be remarkably passive until aroused by quasi-parental leaders whom they admire and trust. It is hard to imagine the Gallic wars without Caesar, the psychoanalytic movement without Freud, the Nazis without Hitler, or the major Communist revolutions without Lenin and Mao Tse-tung and their politburos. These leaders were real, not dummies invented and packaged by image makers from an advertising agency or public relations firm. In the age of massive opinion researches, however, and with the aid of speech coaches and makeup artists and the magic impact of television, it has become increasingly possible for image makers to create front men who can affect the votes and other behaviour of very large percentages of a national audience. As one knowledgeable participant phrased it in 1970:

> There are now four essential ingredients to a professionally managed political campaign: political polls, data processing, imagery, and money. The polls discover what the voter already believes, and data processing interprets and analyzes the depth of voters' attitudes. After that, an image of the candidate is tailored to meet the voters' demands and desires, and the whole package is then sold by massive expenditures of money in the advertising media, particularly television.
>
> The candidate has become relatively unimportant as long as he can be properly managed. The candidate must be

> bright enough to handle the material furnished to him, but not too intelligent, because there is always the danger that an intelligent candidate may come up with unpopular or controversial ideas of his own, and thereby destroy a carefully contrived campaign strategy. [Excerpt from a public address by Zolton Ferency, chairman and gubernatorial candidate of the Democratic Party of Michigan, June 1970.1

Probably this is an overstatement, but it conveys the flavour of a great deal of contemporary political propaganda. Yet a dummy leader invented by an image maker may not always be invulnerable to counterpropaganda by a real leader, if one should turn up. Even a giant, expensive television campaign may not be able to conceal from all reactors the differences between a dummy and a bona fide leader with high political skills—a Franklin D. Roosevelt, for example, or a Jawaharlal Nehru—whose voice and gestures express a genuine and spontaneous concern for public policy and a determination to "wear no man's collar," and who goes in for great numbers of face-to-face appearances that demonstrate that he has no need for a voice coach and a makeup artist.

Predispositions and inducements

Selection and presentation of symbols. The propagandist must realize that neither rational arguments nor catchy slogans can, by themselves, do much to influence human behaviour. A reactor's behaviour is also affected by at least four other variables. The first is the reactor's predispositions—that is, his stored memories of, and his past associations with, related symbols. These often cause the reactor to ignore the current inflow of symbols, to perceive them very selectively, or to rationalize them away. The second is the set of economic inducements (gifts, bribery, pay raises, threats of job loss, and so forth) which the propagandist or others may apply in conjunction with the symbols. The third is the set of physical inducements (love, violence, protection from violence) used by the propagandist or others. The fourth is the array of social pressures that may either encourage or inhibit the reactor in thinking or doing what the propagandist advocates. Even one who is well led and is predisposed to do what the propagandist wants may be prevented from acting by counterpressures within the surrounding social systems or groups of which he is a part.

In view of these predispositions and pressures, the skilled propagandist is careful to advocate chiefly those acts that he believes the reactor already wants to perform and is in fact able to perform. It is fruitless to call upon most people to perform acts that may involve a total loss of income or terrible physical danger—for example, to act openly upon Communist leanings in a totalitarian fascist country. To call upon reactors to do something extremely dangerous or hard is to risk having the propaganda branded as unrealistic. In such cases, it may be better to point to actions that the reactor can *avoid* taking—that is, to encourage him in acts of passive resistance. The propagandist will thereby both *seem* and *be* realistic in his demands upon the reactor, and the reactor will not be left with the feeling, "I agree with this message, but just what am I supposed to do about it?"

For maximum effect, the symbolic content of propaganda must be active, not passive, in tone. It must explicitly or implicitly recommend fairly specific actions to be performed by the reactor ("buy this," "boycott that," "vote for X," "join Group Y," "withdraw from Group Z"). Furthermore, because the ability of the human organism to receive and process symbols is strictly limited, the skillful propagandist attempts to substitute quality for quantity in his choice of symbols. A brief slogan or a picture or a pithy comment on some symbol that is emotion laden for the reactors may be worth ten thousand other words and cost much less. In efforts to economize symbol inputs, the propagandist attempts to make full use of the findings of all the behavioral sciences. He draws upon the psychoanalysts' studies of the bottled-up impulses in the unconscious mind; he consults the elaborate vocabulary counts produced by professors of education; he follows the headline news to determine what events and symbols probably are salient in reactors' minds at the moment; and he analyzes the information polls and attitude studies conducted by survey researchers.

There is substantial agreement among psychoanalysts that the psychological power of propaganda increases with use of what Lasswell termed the triple-appeal principle. This principle states that a set of symbols is apt to be most persuasive if it appeals simultaneously to three elements of an individual's personality—elements that Freud labelled the ego, id, and superego. To appeal to the ego, the skilled propagandist will present the acts and thoughts that he desires to induce as if they were rational, advisable, wise, prudent, and expedient; in the same breath he says or implies that they are sure to produce pleasure and a sense of strength (an appeal to the id); concurrently he suggests that they are moral, righteous, and—if not altogether legal---decidedly more justifiable and humane than the law itself (an appeal to the superego, or conscience). Within any social system, the optimal blend of these components varies from individual to individual and from subgroup to subgroup: some individuals and subgroups love pleasure intensely and show few traces of guilt; others are quite pained by guilt; few are continuously eager to be rational or to take the trouble to become well informed. Some cautious individuals and subgroups like to believe that they never make a move without preanalyzing it; others enjoy throwing prudence to the winds. There are also changes in these blends through time: personalities change, as do the morals and customs of groups. In large collectivities like social classes, ethnic groups, or nations, the particular blends of these predispositions may vary greatly from stratum to stratum and subculture to subculture. Only the study of history and behavioral research can give the propagandist much guidance about such variations.

Role models and parental symbols

A propagandist is wise if, in addition to reiterating his support of ideas and policies that he knows the reactors already believe in, he includes among his images a variety of symbols associated with parents and parent surrogates. The child lives on in every adult, eternally seeking a loving father and mother. Hence the appeal of such familistic symbolisms as "the fatherland," "the mother country," "the Mother Church," "the Holy Father," "Mother Russia," and the large number of statesmen who are known as the "fathers of their countries." Also valuable are reassuring maternal figures like Queen Victoria of England, the Virgin Mary, and the Japanese Sun Goddess. In addition to parent symbols, it is usually well to associate one's propaganda with symbols of parent substitutes, who in some cases exert a more profound effect on children than do disappointing or nondescript parents: affectionate or amiable uncles (Uncle Sam, Uncle Ho Chi Minh); lively aunts (*la belle France,* Britannia, the Spanish Communist leader La Pasionaria, and Kuan-yin, the Chinese Goddess of Mercy); admired scholars and physicians (Karl Marx, Dr. Sun Yat-sen); politico-military heroes and role models (Abraham Lincoln, Winston Churchill, Mao Tse-tung, "the wise, mighty, and fatherly Stalin"); and, of course, saints (Joan of Arc, Mahatma Gandhi, Martin Luther King, the Buddha). A talented and well-symbolized leader or role model may achieve a parental or even godlike ascendancy (charisma) and magnify the impact of a message many times.

Media of propaganda. There are literally thousands of written, audiovisual, and organizational media that a 20th-century propagandist might use. All human groupings are potential organizational media, from the family and other small organizations through advertising and public relations firms, trade unions, churches and temples, theatres, readers of novels and poetry, special-interest groups, political parties and front organizations to the governmental structures of nations, international coalitions, and universal organizations like the United Nations and its agencies. From all this variety of media, the propagandist must choose those few media (especially leaders, role models, and organizations) to whose messages he thinks the intended reactors are especially attentive and receptive.

In recent years the communications revolution has brought about a massive, worldwide proliferation of school systems and of facilities for news gathering, publishing, broadcasting, holding meetings, and speechmaking. At present, almost everyone's mind is bombarded daily by far more media, symbols, and messages than the human organism can possibly pay attention to. The mind reels under noisy assortments of information bits about rival politicians, rival political programs and doctrines, new technical discoveries, insistently advertised commercial products, and new views on morality, ecological horrors, and military nightmares. This sort of communication overload already has resulted in the alienation of millions of people from much of modern life. Overload and alienation can be expected to reach even higher levels in coming generations as still higher densities of population, intercultural contacts, and communication facilities cause economic, political, doctrinal, and commercial rivalries to become still more intense.

Communication overload

Research has demonstrated repeatedly that most reactors attempt, consciously or unconsciously, to cope with severe communication overload by developing three mechanisms: selective attention, selective perception, and selective recall. That is, they pay attention to only a few media; they fail (often unconsciously) to perceive therein any large proportion of the messages that they find uncongenial; and, having perceived, even after this screening, a certain number of unpleasing messages, they repress these in whole or in part (*i.e.*, cannot readily remember them). The contemporary propagandist therefore tries to find out: (1) what formative experiences and styles of education have predisposed his intended audiences to their current "media preferences"; (2) which of all the publications, television shows, leaders, and role models in the world they do in fact pay attention to; and (3) by which of these they are most influenced. These topics have thus become the subjects of vast amounts of commercial and academic research.

In most cases, reactors are found to pay the most attention to the publications, shows, leaders, and role models with whose views they already agree. People as a rule attend to communications not because they want to learn something new or reconsider their own philosophies of life but because they seek psychological reassurance about their existing beliefs and prejudices. When the propagandist does get their attention by putting his message into the few media they heed, he may discover that, to hold their attention, he must draft a message that does not depart very far from what they already want to believe. Despite the popular stereotypes about geniuses of politics, religion, or advertising whose brilliant propaganda converts the multitudes overnight, the plain fact is that even the most skilled propagandist must usually content himself with a very modest goal: packaging a message in such a way that much of it is familiar and reassuring to the intended reactors and only a little is so novel or true as to threaten them psychologically. Thus, revivalists have an a priori advantage over spokesmen of a modernized ethic, and conservative politicians an advantage over progressives. Propaganda that aims to induce major changes is certain to take great amounts of time, resources, patience, and indirection, except in times of revolutionary crisis when old beliefs have been shattered and new ones have not yet been provided. In ordinary periods (intercrisis periods), propaganda for changes, however worthy, is likely to be, in the words of the German sociologist Max Weber, "a slow boring of hard boards."

Use of reference groups

For reasons just indicated, the most effective media as a rule (for messages other than the simplest of commercial advertising) are not the impersonal mass media like newspapers and television but rather those few associations or organizations (reference groups) with which the individual feels identified or to which he aspires to relate his identity. Quite often the ordinary man not only avoids but actively distrusts the mass media or fails to understand their messages; but in the warmth of his reference groups he feels at home, assumes that he understands what is going on, and feels that he is sure to receive a certain degree of emotional response and personal protection. The foremost reference group, of course, is the family. But many other groups perform analogous functions—for instance, the group of sports fans, the church, the trade union, the alumni group, the clique or gang, the Communist cell. By influencing the key members of such a group, the propagandist may establish a "social relay" channel that can amplify his message. By concentrating thus on the few, he increases his chances of reaching the many—often far more effectively than he could through a plethora of mass meetings, paid broadcasts, handbills, or billboards and at much lower cost. Therefore, one important stratagem involves the combined use of mass media and reference-group channels—writing up materials for such media as news releases or broadcasts in ways designed specifically to reach specified groups (and especially their elites and leaders): who can then relay the messages to other sets of reactors.

The reactors (audiences). The audiences for the propagandist can be classified into: (1) those who are initially predisposed to react as the propagandist wishes, (2) those who are neutral or indifferent, and (3) those who are in opposition or perhaps even hostile.

As already indicated, propaganda is most apt to evoke the desired responses among those already in agreement with the propagandist's message. Neutrals or opponents are not apt to be much affected even by an intensive barrage of propaganda unless it is reinforced by nonpropagandistic inducements (economic or coercive acts) or by favourable social pressures. These facts, of course, are recognized by advocates of civil disobedience; their propagandists would contend that sloganeering and reasoned persuasion must be accompanied by sit-ins and other overt acts of passive resistance; they aim for a new climate of social pressure. These facts are also significantly recognized by Communist regimes; by controlling all means of production, they can offer great economic inducements or threaten a man's livelihood, thus making him a very attentive audience for propaganda. If these copressures are applied too strongly, however, they may become so distasteful to reactors that the associated propaganda will backfire.

Measurement of the effects of *propaganda*. The modern world is overrun with all kinds of competing propaganda and counterpropaganda and a vast variety of other symbolic activities, such as education, publishing, newscasting, and patriotic and religious observances. The problem of distinguishing between the effects of one's own propaganda and the effects of these other activities is often extremely difficult.

Use of experiments

The ideal scientific method of measurement is the controlled experiment. Carefully selected samples of members of the intended audiences can be subjected to the propaganda while equivalent samples are not. Or the same message, clothed in different symbols—different mixes of sober argument and "casual" humour, different proportions of patriotic, ethnic, and religious rationalizations, different mixes of truth and the "noble lie," different proportions of propaganda and coercion—can be tested on comparable samples. Also, different media can be tested to determine, for example, whether results are better when reactors read the message in a newspaper, observe it in a spot commercial on television, or hear it wrapped snugly in a sermon. Obviously the number of possible variables and permutations in symbolism, media use, subgrouping of the audience, and so forth is extremely great in any complicated or long-drawn-out campaign. Therefore, the costs for the research experts and the fieldwork that are needed for thorough experimental pretests are often very high. Such pretests, however, may well save money in the end.

An alternative to controlled experimentation in the field is controlled experimentation in the laboratory. But it may be impossible to induce reactors who are truly representative of the intended audience to come to the laboratory at all. Moreover, in such an artificial environment their reactions may differ widely from the reactions that they would have to the same propaganda if reacting unself-consciously in their customary environment. For these and many other obvious reasons, the validity of laboratory pretests of propaganda must be viewed with the greatest caution.

Whether in the field or the laboratory, the value of all controlled experiments is seriously limited by the problem of "sleeper effects." These are long-delayed reactions that may not become visible until the propaganda has penetrated resistances and insinuated itself deep down into the reactor's mind—by which time the experiment may have been over for a long time. Another problem is that most people acutely dislike being guinea pigs and also dislike the word propaganda. If they find out that they are subjects of a propagandistic experiment, the entire research program, and possibly the entire campaign of propaganda of which it is a part, may backfire.

Use of interviews and observation techniques

Another research device is the panel interview—repeated interviewing, over a considerable period of time, of small sets of individuals considered more or less representative of the intended audiences. The object is to obtain (if possible, without their knowing it) a great deal of information about their life-styles, belief systems, value systems, media habits, opinion changes, heroes, role models, reference groups, and so forth. The propagandist hopes to use this information in planning ways to influence a much larger audience. Panel interviewing, if kept up long enough, may help in discovering sleeper effects and other delayed reactions. The very process of being "panel interviewed," however, produces an artificial environment that may induce defensiveness, suspiciousness, and even attempts to deceive the interviewer.

For many practical purposes, the best means of measuring—or perhaps one had better say estimating—the effects of propaganda is apt to be the method of extensive observation, guided of course by well-reasoned theory and inference. "Participant observers" can be stationed unobtrusively among the reactors. Voting statistics, market statistics, press reports, police reports, editorials, and the speeches or other activities of affected or potentially affected leaders can also give clues. Evidence on the size, composition, and behaviour of the intermediate audiences (such as elites) and the ultimate audiences (such as their followers) can be obtained from these various sources and from sample surveys. The statistics of readership or listenership for printed and telecommunications media may be available. If the media include public meetings, the number of people attending and the noise level and symbolic contents of cheering (and jeering) can be measured. Observers may also report their impressions of the moods of the audience and record comments overheard after the meeting. To some extent, symbols and leaders can be varied, and the different results compared.

Use of "content analysis"

Using methods known in recent years as content analysis, the propagandist can at least make reasonably dependable quantitative measurements of the symbolic contents of his own propaganda and of communications put out by others. He can count the numbers of column inches of printed space or seconds of radio or television time that were given to the propaganda. He can categorize and tabulate the symbols and themes in the propaganda. To estimate the implications of the propaganda for social policy, he can tabulate the relative numbers of expressed or implied demands for actions or attitude changes of various kinds. The 1970 edition of volume 1 of the *Big Soviet Encyclopedia,* for example, had no pictures of Stalin; in the previous edition, volume 1 had four pictures. Did this mean that a new father figure and role model was being created by the Soviet propagandists? Or did it indicate a return to the cult of older father figures such as Marx and Lenin? If so, what were the respective father figures' traits, considered psychoanalytically, and what are the political, economic, and military implications for Soviet policy?

By quantifying his data about contents, the propagandist can bring a high degree of precision into experiments using different propaganda contents aimed at the same results. He can also increase the accuracy of his research on the relative acceptability of information, advice, and opinion attributed to different sources. (Will given reactors be more impressed if they hear 50, 100, or 200 times that a given policy is endorsed--or denounced—by the president of the U.S., the premier of the U.S.S.R., or the pope?)

Very elaborate means of coding and of statistical analysis have been developed by various content analysts. Some count symbols, some count headlines, some count themes (sentences, propositions), some tabulate the frequencies with which various categories of "events data" (newspaper accounts of actual happenings) appear in some or all of the leading newspapers ("prestige papers") or television programs of the world. Some of these events data can be counted as supporting or reinforcing the propaganda, some as opposing or counteracting it. Whatever the methodology, content analysis in its more refined forms is an expensive process, demanding long and rigorous training of well-educated and extremely patient coders and analysts. And there remains the intricate problem of developing relevant measurements of the effects of different contents upon different reactors.

Countermeasures by opponents. Some countermeasures against propaganda include simply suppressing it by eliminating or jailing the propagandist, burning down his premises, intimidating his employees, buying him off, depriving him of his use of the media or the money that he needs for the media or for necessary research, and applying countless other coercive or economic pressures. It is also possible to use counterpropaganda, hoping that the truth (or at least some artfnl bit of counterpropaganda) will prevail.

One special type of counterpropaganda is "source exposure"-informing the audience that the propagandist is ill informed, is a criminal, or belongs to some group that is sure to be regarded by the audience as subversive, thereby undermining his credibility and perhaps his economic support. In the 1930s there was in the U.S. an Institute for Propaganda Analysis that tried to expose such propaganda techniques as "glittering generalities" or "name-calling" that certain propagandists were using. This countermeasure may have failed, however, because it was too intellectual and abstract and because it offered the audience no alternative leaders to follow or ideas to believe.

In many cases opponents of certain propagandists have succeeded in getting laws passed that have censored or suppressed propaganda or required registration and disclosure of the propagandists and of those who have paid them.

Measures against countermeasures. It is clear, then, that opponents may try to offset propaganda by taking direct action or by invoking covert pressures or community sanctions to bring it under control. The propagandist must therefore try to estimate in advance his opponents' intentions and capabilities and invent measures against their countermeasures. If he thinks that they will rely only on counterpropaganda, he can try to outwit them. If he thinks that they will withdraw advertising from his newspaper or radio station, he may try to get alternative supporters. If he expects vigilantes or police persecution, he can go underground and rely, as the Russian Communists did before 1917 and the Chinese before 1949, primarily on agitation through organizational media.

SOCIAL CONTROL OF PROPAGANDA

The democratic ideal and propaganda

Democratic control of propaganda. Different sorts of polities, ranging from the democratic to the authoritarian, have attempted a variety of social controls over propaganda. In an ideal democracy, everyone would be free to make propaganda and free to oppose propaganda habitually through peaceful counterpropaganda. The democratic ideal assumes that, if a variety of propagandists are free to compete continuously and publicly, the ideas best for society will win out in the long run. This outcome would require that a majority of the general populace be reasonably well-educated, intelligent, public-spirited, and patient, and that they not be greatly confused or alienated by an excess of communication. A democratic system also presupposes that large quantities of dependable and relevant information will be inexpensively disseminated by relatively well-financed, public-spirited, and uncensored news gathering and educational agencies. The extent to which any existing national society actually conforms to this model is decidedly an open question. That the world social system does not is self-evident.

In efforts ***to*** guard against "pernicious" propaganda by hidden persuaders, modern democracies sometimes require that such propagandists as lobbyists and publishers register with public authorities and that propaganda and advertising be clearly labelled as such. The success of such measures, however, is only partial. In the U.S., for instance, publishers of journals using the second-class mails are required to issue periodic statements of ownership, circulation, and other information; thereby, at least the nominal owners and publishers become known—but those who subsidize or otherwise control them may not. In many places, paid political advertisements in newspapers or on television are required to include the name of a sponsor—but the declared sponsor may be a "dummy" individual or organization whose actual backers remain undisclosed. Furthermore, agents of foreign governments or organizations engaged in propaganda in the U.S. are required to file forms with the U.S. Department of Justice, naming their principals and listing their own activities and finances—but it is impossible to know whether the data so filed are correct, complete, or significant. In many Western industrial nations, similar registrations and disclosures are required of those who circulate brochures inviting investors to buy stocks and bonds. This principle of disclosure, which appears so useful with respect to foreign agents and securities salesmen, is not often applied, however, to other media of propaganda. (In the U.S. the disclosure of certain types of political campaign advertisements and contributions is required, but the requirement is easily circumvented.) In many countries, claims made in propaganda (including advertising) about the contents or characteristics of foods and drugs and some other products are also subject to registration and to requirements of "plain labelling." In some places, consumer research organizations, privately or publicly supported, examine these claims rigorously and sometimes publish scientifically based counterpropaganda. Finally, there has been an increase in laws and customs requiring that equal space or time or a right of reply be rendered all major contenders in political campaigns or even major spokesmen differing on major issues of the day. In view of the apparently massive effects and the certainly massive expenses of political propaganda on television, there are many movements afoot in democracies to limit expenditures on campaign propaganda and to require networks to give time free of charge for even the minor parties, especially in the weeks immediately preceding elections. There have also been movements to require that political propaganda be halted for a specified number of days before the holding of an election—the idea being that a cooling-off period would allow voters to rest and reflect after the communication overload of the campaign period and would prevent politicians and their backers from using last-minute slander and sensationalism.

Equal time
and equal
space

Authoritarian control of propaganda. In a highly authoritarian polity, the regime tries to monopolize for itself all opportunities to engage in propaganda, and often it will stop at nothing to crush any kind of counterpropaganda. How long and how completely such a policy can be implemented depends, among other things, on the amount of force that the regime can muster, on the thoroughness of its police work, and, perhaps most of all, on the level, type, and distribution of secular higher education. Secular higher education invariably promotes skepticism about claims that sound dogmatic or are made without evidence; and if such education is of a type that emphasizes humane and universalistic values, an ignorant or unreasonable authoritarian regime is not likely to please the educated for very long. If the educated engage in discreet counterpropaganda, they may in the end modify the regime.

World-level control of propaganda. One of the most serious and least understood problems of social control is above the national level, at the level of the world social system. At the world level there is an extremely dangerous lack of means of restraining or counteracting propaganda that fans the flames of international, interracial, and interreligious wars. The global system consists at present of a highly chaotic mixture of democratic, semidemocratic, and authoritarian subsystems. Many of these are controlled by leaders who are ill educated, ultranationalistic, and religiously, racially, or doctrinally fanatical. At present, every national regime asserts that its national sovereignty gives it the right to conduct any propaganda it cares to, however untrue such propaganda may be and however contradictory to the requirements of the world system. The most inflammatory of such propaganda usually takes the form of statements by prominent national leaders, often sensationalized and amplified by their own international broadcasts and sensationalized and amplified still further by media in the receiving countries. The only major remedy would lie, of course, in the slow spread of education for universalist humanism. A first step toward this might be taken through the fostering of an energetic and highly enlightened press corps and educational establishment, doing all it can to provide the world's broadcasters, newspapers, and schools with factual information and illuminating editorials that could increase awareness of the world system as a whole. Informed leaders in world affairs are therefore becoming increasingly interested in the creation of world-level media and multinational bodies of reporters, researchers, editors, teachers, and other intellectuals committed to the unity of mankind.

BIRLIOGRAPHY. Annotated listings of books and articles of all times, countries, and languages, with respect to public opinion and the theory and practice of communication (including propaganda) appear in H.D. LASSWELL, R.D. CASEY, and B.L. SMITH (eds.), *Propaganda and Promotional Activities: An Annotated Bibliography* (1935, reprinted 1969); and in B.L. SMITH, H.D. LASSWELL, and R.D. CASEY, *Propaganda, Communication and Public Opinion: A Comprehensive Reference Guide* (1946). Further listings on general and international propaganda are in B.L. and C.M. SMITH, *International Communication and Political Opinion: A Guide to the Literature* (1956). For more recent periodical literature, see *International Political Science Abstracts* (quarterly); *Psychological Abstracts* (monthly); and *Sociological Abstracts* (8/yr.).

General works of considerable significance include: F.C. BARGHOORN, *The Soviet Cultural Offensive*(1960) and *Soviet Foreign Propaganda* (1964); K.W. DEUTSCH, *The Nerves of Government: Models of Political Communication and Control* (1963); L.A. DEXTER and D.M. WHITE (eds.), *People, Society, and Mass Communications* (1964); L.W. DOOB, *Public Opinion and Propaganda* (1948); J. DRIENCOURT, *La propagande, nouvelle force politique* (1950); J. ELLUL, *Propagande* (1962; Eng. trans., *Propaganda,* 1965); L. FESTINGER, *A Theory of Cognitive Dissonance* (1957); SIGMUND FREUD, *Massenpsychologie und Ich-Analyse* (1922; Eng. trans. by J. STRACHEY, *Group Psychology and the Analysis of the Ego,* 1948); A.L. GEORGE, *Propaganda Analysis: A Study of Inferences Made from Nazi Propaganda in World War II* (1959); R.T. HOLT and R.W. VAN DE VELDE, *Strategic Psychological Operations and American Foreign Policy* (1960); I.L. JANIS *et al., Personality and Persuasibility* (1959); J.T. KLAPPER, *The Effectsof Mass Communication* (1960); H.D. LASSWELL, *Propaganda Technique in the World War* (1927, reprinted 1971); with D. BLUMENSTOCK, *World Revolutionary Propaganda* (1939, reprinted 1970); *et al., Language of Politics* (1949); and with D. LERNER and I. DE SOLA POOL, *The Comparative Study of Symbols* (1952); B.R. BERELSON, P.F. LAZARSFELD and W.N. MCPHEE, *Voting: A Study of Opinion Formatiorz in a Presidential Campaign* (1954); E. KATZ and P.F. LAZARSFELD, *Personal Influence* (1955); D. LERNER (ed.), *Sykewar: Psychological Warfare Against Germany, D-Day to VE-Day* (1949), and (ed.), *Propaganda in War and Crisis* (1951); J. MCGINNISS, *The Selling of the President, 1968* (1969); P. SELZNICK, *The Organizational Weapon: A Study of Bolshevik Strategy and Tactics* (1952); R.K. WHITE, *Nobody Wanted War: Misperceptiorz in Vietnam and Other Wars,* rev. ed. (1970); and TE-CHI YU, *Mass Persuasion in Commrrnist China* (1964).

On propaganda aspects of diplomatic negotiation, see F.C. IKLE, *How Nations Negotiate* (1964); and H.G. NICOLSON, *Diplomacy,* 3rd ed. (1963). For aspects of propaganda problems in less-developed areas, see L.W. PYE (ed.), *Communications and Political Development* (1963); and W.L. SCHRAMM, *Mass Media and National Development* (1964).

On legal aspects and social control of propaganda, see L.J. MARTIN, *International Propaganda: Its Legal and Diplomatic Control* (1958); B.S. MURTY, *Propaganda and World Public Order* (1968); H.D. LASSWELL, *Democracy Through Public Opinion* (1941); J.B. WHITTON and A. LARSON, *Propaganda*

Towards Disarmament in the War of Words (1964); and J.W. BURTON, *Conflict and Communication: The Use of Controlled Communication in International Relations (1969).*

(B.L.S.)

Propertius, Sextus

The most forceful of the Latin elegiac poets, Sextus Propertius emerges from his poetry as a man with a wonderful capacity for living, eager to embrace every aspect of life and experience. Although best known for its sensuousness, his work also reveals a deep understanding of life's tragic underside: that there is sometimes only a step between happiness and misery. He is often self-absorbed and self-pitying, and his poetry can suffer from an overdose of pathos, but these defects are largely neutralized by his welcome sense of humour and, especially, by his poetic power and range of imagination. Technically a very competent poet, his handling of the elegiac couplet is unmistakable for its vigorous, masculine rhythm.

Sextus Propertius was born in or near Assisi, in Umbria, between 55 and 43 BC (most probably *c.* 50 BC). Very few details of his life are known. His father died when Propertius was still a boy, but he was given a good education by his mother. Part of the family estate was confiscated (*c.* 40 BC) to satisfy the resettlement needs of the veteran troops of Octavian, later the emperor Augustus, after the civil wars. Propertius' income was thus severely diminished, though he was never really poor. With his mother, he left Umbria for Rome, and there (c. 34 BC) he assumed the dress of manhood. Some of his friends were poets (including Ovid and Bassus), and he had no interest in politics, the law, or army life. His first love affair was with an older woman, Lycinna, but this was only a passing fancy when set beside his subsequent serious attachment to the famous "Cynthia" of his poems.

The first of Propertius' four books of elegies (the second of which is divided by some editors into two) was published in 29 BC, the year in which he first met "Cynthia," its heroine. It was known as the *Cynthia* and also as the *Monobiblos* because it was for a long time afterward sold separately from his other three books. Complete editions of all four books were also available. *Cynthia* seems to have had an immediate success, for the influential literary patron Maecenas invited Propertius to his house, where he doubtless met the other prominent literary figures who formed Maecenas' circle. These included the poets Virgil (whom Propertius admired) and Horace (whom he never mentions). The influence of both, especially that of Horace in Book III, is manifest in his work.

Meeting with Virgil and Horace

Cynthia's real name, according to the 2nd-century writer Apuleius, was Hostia. It is often said that she was a courtesan, but elegy 16 in Book I seems to suggest that she belonged to a distinguished family. It is likely that she was married, though Propertius only mentions her other lovers, never her husband. From the poems she emerges as beautiful, passionate, and uninhibited. She was intensely jealous of Propertius' own infidelities and is painted as a woman terrible in her fury, irresistible in her gentler moods. Propertius makes it clear that, even when seeking pleasures apart from his mistress, he still loved her deeply, returning to her full of remorse, and happy when she reasserted her dominion over him.

After many violent scenes, it appears that Propertius finally broke off his tempestuous affair with her in 24 BC, though inferring dates from the poems' internal evidence cannot be undertaken with real confidence, as this kind of personal poetry often interweaves fact with fancy. He was to look back on his liaison with her as a period of disgrace and humiliation. This may be more than a mere literary pose, although after Cynthia's death (she does not seem to have lived for long after their break) he regretted the brusqueness of their separation and was ashamed that he had not even attended her funeral. In a most beautiful and moving elegy (IV:7), he conjures up her ghost and with it recreates the whole glamour and shabbiness of the affair. While he makes no attempt to brush over the disagreeable side of her nature, he also makes it clear that he loves her beyond the grave.

Propertius' poetic powers matured with experience. The poetry of Book II is far more ambitious in scope than that of Book I and shows a richer orchestration. His reputation grew, and the emperor Augustus himself seems to have taken notice of him, for, in Books III and IV, the poet laments the premature death of Marcellus, Augustus' nephew and heir apparent (III:18), and he composed a magnificent funeral elegy (IV:11) in praise of Cornelia, Scribonia's daughter by her first marriage—the "Queen of Elegies" as it is sometimes called.

Content of the poetry

As his poetic powers developed, so also did Propertius' character and interests. In his earliest elegies, love is not only his main theme but is almost his religion and philosophy. It is still the principal theme of Book II, but he now seems a little embarrassed by the popular success of Book I and is anxious not to be thought of simply as a gifted scoundrel who is constantly in love and can write of nothing else. In Book II he considers writing an epic, is preoccupied with the thought of death, and attacks (in the manner of later satirists, such as Juvenal) the coarse materialism of his time. He still loves to go to parties and feels perfectly at ease in the big city with its crowded streets, its temples, theatres, porticoes, and its disreputable quarters. In a way he is a conservative snob, in general sympathy with Roman imperialism and Augustan rule; but he is open to the beauties of nature and is genuinely interested in works of art; though he disapproves of ostentatious luxury, he also appreciates contemporary fashions.

Some of his contemporaries accused him of leading a life of idleness and complained that he contributed nothing to society. But Propertius felt it his duty to support the right of the artist to lead his own life, and he demanded that poetry, and art in general, should not be regarded simply as a civilized way of passing the time. In elegy 3 of Book III he gives deep meaning to the process of artistic creation and emphasizes the importance of the creative artist.

Influence of Alexandrian poets on Propertius

In Books III and IV Propertius demonstrates his command over various literary forms, including the diatribe and the hymn. Many of his poems show the influence of such Alexandrian poets as Callimachus and Philetas. Propertius acknowledges this debt, and his claim to be the "Roman Callimachus," treating Italian themes in the baroque Alexandrian manner, is perhaps best shown in a series of elegies in Book IV that deal with aspects of Roman mythology and history, and were to inspire Ovid to write his *Fasti,* a calendar of the Roman religious year. These poems are a compromise between the elegy and the epic. Book IV also contains some grotesque, realistic pieces, two unusual funeral elegies, and a poetic letter.

Two of the lasting merits of Propertius seem to have impressed the ancients themselves. The first they called *blanditia,* a vague but expressive word by which they meant softness of outline, warmth of colouring, a fine and almost voluptuous feeling for beauty of every kind and a pleading and melancholy tenderness: this is most obvious in his descriptive passages and in his portrayal of emotion. His second and even more remarkable quality is poetic *facundia,* or command of striking and appropriate language. Not only is his vocabulary extensive but his employment of it is extraordinarily bold and unconventional: poetic and colloquial latinity alternate abruptly, and in his quest for the striking expression he frequently seems to strain the language to the breaking point.

Propertius' handling of the elegiac couplet, and particularly of the pentameter, deserves especial recognition. It is vigorous, varied, and picturesque. In the matter of the rhythms, caesuras, and elisions that it allows, the metrical treatment is more severe than that of Catullus, but noticeably freer than that of Ovid, to whose stricter usage, however, Propertius increasingly tended (particularly in his preference for a disyllabic word at the end of the pentameter). An elaborate symmetry is observable in the construction of many of his elegies, and this has tempted critics to divide a number of them into strophes.

As Propertius had borrowed from his predecessors, so his successors, Ovid above all, borrowed from him; and *graffiti* on the walls of Pompeii attest his popularity in the 1st century AD. In the Middle Ages he was virtually forgotten; and since the Renaissance he has been studied by professional scholars more than he has been enjoyed by the general public. To the modern reader, acquainted with the psychological discoveries of the 20th century, the self-revelations of his passionate, fitful, brooding spirit are of peculiar interest.

Almost nothing is known about Propertius' life after his love affair with Cynthia was over. It is possible that he married her successor in his affections (perhaps in order to qualify for the financial benefits offered to married men by the *leges Juliae* of 18 BC) and had a child, for an inscription in Assisi and two passages in the letters of the younger Pliny (AD 61/62–*c.* 113) indicate that Propertius had a descendant called Gaius Passennus Paulus Propertius, who was also a poet. During his later years he lived in an elegant residential area in Rome on the Esquiline Hill. The date of his death is not certain, though he was still alive in 16 BC, for two events of that year are mentioned in his fourth book, which was perhaps edited posthumously.

Sextus Propertius is generally thought of as a difficult poet, partly because of the uncertainty of many manuscript readings, partly because of the often obscure mythological allusions in which his work abounds. But such allusions are not there for ornament; they are a kind of emotional shorthand by which Propertius expresses his deep insight into human nature. The mellifluous Greek names, moreover, are skillfully woven into the rich verbal pattern that, along with his sincerity, is one of his outstanding characteristics.

MAJOR WORKS

Four Books of Elegies: Book I, known in antiquity as the *Cynthia Monobiblos* (*c.* 33–c. 28 BC); Book II (*c.* 28–c. 25 BC); Book III (c. 24–c. 22 BC); and Book IV (*c.* 21–c. 16 BC). Special mention may be made of Elegy 1, Book I, which introduces Cynthia; of Elegy 28 (both parts), Book II, a prayer for her recovery from illness; of Elegy 10, Book III, on Cynthia's birthday; of Elegy 24, Book III, on their final parting after a quarrel; and of Elegy 7, Book III, on Cynthia's death.

BIBLIOGRAPHY

Important editions and textual criticism: JOSEPHUS SCALIGER (1577) contains many textual conjectures and transpositions of lines that were accepted by most subsequent editors before CARL LACHMANN (1816), who restored essentially the text of the best manuscripts. E.A. BARBER (1953) is the most commonly used critical edition today. D.R. SHACKLETON BAILEY, ***Propertiana*** (1956), gives valuable critical and exegetical discussions of difficult passages.

Commentaries: JANUS BROUKHUSIUS (1727), still recommended for its elegance and learning; W.A. CAMPS, 4 vol. (1961–67), original 'and penetrating.

Translation: A.E. WATTS, ***The Poems of Sextus Propertius,*** rev. ed. (1966), in rhymed couplets—often refreshing but makes Propertius sound too harmless.

Literary criticism: G. LUCK, ***The Latin Love Elegy,*** 2nd ed. (1969), has two chapters on Propertius, the man and the artist, and a bibliographical survey of recent research.

(G.Lu.)

Property, Law of

Property may be defined as an exclusive right to control an economic good; it is the name for a concept that refers to the rights and obligations, privileges and restrictions that govern the relations of men with respect to things of value. People everywhere and at all times desire the possession of things that are necessary for survival or valuable by cultural definition and that, as the result of the demand placed upon them, become scarce. Customs as well as legislation enforced by organized society control the competition for, and guarantee the enjoyment of, these desired things. What is guaranteed to be one's own is, in a broad sense, property.

The word property is frequently used indiscriminately to denote not only ***objects*** of rights that have a pecuniary content but also ***rights*** that persons have with respect to things. Thus, lands and chattels are said to be property, and rights, such as ownership, life estates, and easements, are likewise said to be property. Accurate legal terminology, however, usually reserves the use of the word property for the designation of rights that persons have with respect to ***things.***

The concept of things

Not every thing is controlled by property rights, nor are these rights themselves always governed by the law of property. Legislation, doctrine, or jurisprudence in various legal systems defines the things that can become objects of property rights. In most legal systems. including common law jurisdictions and systems of the French family, the word things applies to both physical objects and intangibles. In legal systems following the model of the German Civil Code, such as the Japanese system, the word things applies only to corporeal objects that are susceptible to appropriation. In these systems, intangibles, such as rights and obligations, are not things nor are they, technically, objects of property rights. Accurate definition of the word things is indispensable because only things in the legal sense can be objects of property rights.

With respect to things, persons may have a variety of rights, some of which confer a direct and immediate authority over a thing whereas others merely confer the possibility of enjoyment through the intervention of another person. An owner and a lessee, according to appearances, for example, seem to have the use and enjoyment of a house in much the same way. But technically the owner has a direct and immediate authority over the house; the lessee has a right against the owner of a house to let him enjoy the house. All rights with respect to things, if they are susceptible to monetary evaluation, are property in the sense that they are guaranteed by the legal order and form part of a person's patrimony. But only rights that confer a direct and immediate authority over a thing are governed by the law of property.

For the purposes of this article, the law of property may thus be defined as a branch of private law that deals with rights conferring a direct and immediate authority over things. This definition of the domain of property law distinguishes it from other branches of private law, namely, from the law of persons, the law of contracts or conventional obligations, the law of torts, the law of family, and the law of successions. These branches deal with relations that may, and frequently do, give rise to property rights. Because of their origin and purpose, however, these property rights are often subject to special rules outside the law of property.

The article contains the following sections:

I. Historical development of property rights

Property rights have developed along with the social organization of mankind. As a legal institution, private property has been known in ancient and even primitive legal systems; as a social fact, private property has been observed in the most primitive societies, tribes, clans, or other groups. Of course, definitions of what constitutes property, and attitudes toward property rights, vary with different cultures and different historical periods.

PROPERTY RIGHTS IN DIFFERENT TYPES OF SOCIETIES

Primitive and nomadic societies. From a partial study of primitive groups and cultures, the German philosopher Friedrich Engels (1820–95) advanced his political rather than scientific thesis that the most primitive form of social and economic organization was that of communism, and this thesis is still expounded today. Modern research has shown, however, that such simple labels as communistic or individualistic are not adequate for the complex scheme of privileges and duties that characterize the relations between individuals and the primitive community with respect to things. It is true, of course, that the institutions of primitive societies often show a much closer relationship between the individual and his group than they do in Western cultures, and they show a greater readiness to yield to the claims of the group. Among hunting or fishing tribes, the community may have the cverlordship of hunting grounds, of fishing vessels, and even of dwellings or of any domesticated animals, although individual claims and privileges are not excluded. At the same time, private property is clearly recognized in weapons, articles of clothing, and ornaments, Individual claims and privileges to other things, such as hunting grounds, are often accompanied by specified duties and responsibilities toward the things and the community. Nowhere has there been discovered an irrational or undifferentiated absorption of the individual into the group.

Community control

In nomadic societies, property rights are distributed among the tribe, clans or families, and the individual members. Individual property rights are largely confined to chattels, especially weapons, utensils, and ornaments. Domesticated animals, ceremonial items, and medicine bundles manifesting the possession of supernatural powers may be the objects of either tribal, family, or individual property rights, depending on the economic conditions and social structure of each community. When land is plentiful and when the whole system of husbandry is based on roaming about large tracts of land, there is no reason to carve out individual plots for permanent occupation and use. Even with the appearance of agriculture, territorial rights remain for some time vested in the community or in clans and families.

Agricultural societies. In developed agricultural societies, in addition to property rights in chattels, there are well-defined property rights in land, which, by virtue of cultivation, become a most important economic good. Rights in land, at any historical stage, may be vested in the community, in clans or families, and in individual members with a varying degree of intensity. It would be dangerous to generalize from a number of observed types of social and economic organization and to project conclusions as to the most prevalent form of property rights in land, namely, as to the communal, family, or individual ownership of the soil. Vestiges of communal tenure have been observed in various parts of the world, ranging from exclusive control to mere supervision of the use of lands by families or individuals. Clan, family, and individual rights to the exclusive use of parcels of land, subject to a variety of obligations and responsibilities, have also been observed. The intensity of individual claims to the exclusive use and enjoyment of lands, and of attending duties and responsibilities, may vary with each society, but one may assert with a degree of certainty that the phenomenon of individual property rights in lands appeared for the first time in the framework of agricultural societies.

Urban societies. In urban and commercial societies, as exemplified by ancient Greek and Roman communities, family and individual property rights in lands and chattels tend to become the rule. Vestiges of communal tenure within the tribe are scarce in the case of Greek and Roman communities, because the culture of these peoples developed in connection with towns surrounded by small plots of intensive cultivation. Individual ownership of lands and chattels achieved the status of an exclusive right in ancient Greek communities and that of dominium, an absolute right of ownership, in Rome.

In the long history of property law in Western civilization, simple and unencumbered ownership was known almost exclusively at the beginning and end of the Roman period and. again, after the French Revolution. In the Middle Ages, lands in western Europe were subjected to feudal tenures, namely, to a regime of division of ownership into perpetual interests of landlords and tenants. Landowners retained the ownership of lands, but they did not have possession. Originally land reverted to the landowner on the death of the tenant, to be redistributed at the landowner's will, but by the 9th and 10th centuries tenants began to hold lands by virtue of a perpetual and heritable right of enjoyment even though they did not have ownership. At the end of the feudal period (12th and 13th centuries), the nature of the tenant's interest came to be identified as true ownership, although it remained subject to a tax or relief in favour of the former landlord at the time of inheritance. Although feudalism in the strict sense ceased to exist, some perpetual tenures and land reliefs continued to exist in France until the beginning of the Revolution, and in Prussia until almost the same time. Since that time, the only right of ownership recognized in the legal systems belonging to the French family has been full ownership corresponding to the *dominium* of the early Roman law. Analogous developments took place in most parts of continental Europe where feudalism had prevailed. In the United States, tenures have been largely abolished by statutes declaring all land to be allodial (freely held, without obligation to another); and in American jurisdictions in which tenures may be said to have survived, the only incident of tenure is escheat (property that reverts to the state if there is no one competent to inherit it). In England, reform legislation designed to suppress the few remaining vestiges of feudal law was enacted in 1925.

Feudal development

Modern reforms

Neither codification of the law of property in civil law jurisdictions nor reform legislation in common law jurisdictions has hindered further evolution. Modern legislation and judicial practice in the Western world and countries with comparable or derivative systems exhibit a tendency toward a limitation of the intensity of ownership in the interest of all and, at the same time, toward extension of ownership to new forms of wealth. These developments in contemporary systems tend to indicate that ownership is not an absolute right but involves, rather, social as well as individual responsibilities. The right of ownership does not confer today, as it did in the feudal period, political, social, and economic privileges. Yet the possession of wealth still exerts much de facto power in society. In spite of limitations, ownership allows considerable freedom for the satisfaction of purely private interests and endows the owner with a type of limited sovereignty.

In Socialist countries, the institutions of property underwent drastic changes. Traditional concepts were suppressed, and new kinds of property rights emerged in conformity with Marxist philosophy. Property in Socialist countries is fundamentally divided into goods of production and goods of consumption and into personal property, cooperative property, and Socialist property. Superficial comparison between these divisions and property institutions in the Western sense, without reference to underlying basic economic, political, and social philosophy, would be both misleading and inaccurate.

ACQUISITION OF PROPERTY RIGHTS

Property rights may be acquired in a variety of ways: by the occupancy of things that belong to no one, by transfer from a previous owner or even by a nonowner, by operation of law, by the effect of judgments, and by acts of public authorities. For systematic purposes, distinction may be made between *original* and *derivative* acquisition of property rights. An original acquisition involves the creation of new property right; it is independent of any pre-existing rights over the same thing. A derivative acquisition involves a transfer of a pre-existing right from one person to another.

The distinction between the two modes of acquisition of property rights is important in the light of the maxim, prevailing in most legal systems, that no one can transfer a greater right than one has. This means that, ordinarily, the transferor must be owner and must transfer the property as it may be burdened with rights of third persons. In a number of contemporary legal systems, the scope of the maxim has been narrowed by exceptions. Thus, under the laws of France and Germany, property rights in movables that are neither lost nor stolen may be transferred by a nonowner to a good-faith purchaser for value. In American jurisdictions, analogous results have been reached by a so-called bona fide (good-faith) purchaser doctrine. Lands, as a rule, must be transferred by the true owner; yet, by way of exception, a good-faith purchaser may be protected in several legal systems if he has relied on entries in land registers or other public records. These problems do not arise in cases of original acquisition of property rights.

Original acquisition. There is a variety of original modes of acquisition of property rights. For the purposes of this article, attention may be focussed on occupancy, finding, accession, acquisitive prescription, expropriation, and the establishment of property rights by acts of public authorities.

Occupancy. Occupancy—namely, the taking of possession of things that belong to no one—is perhaps universally recognized. In Roman law things without owner became the property of the first possessor. In contemporary legal systems, this mode of acquisition of property rights is largely limited to movable things, such as wild animals, birds, fish, and abandoned chattels. In times past, lands could also be acquired by mere occupancy or cultivation in various parts of the world; today, however, as a rule, the acquisition of property rights in lands is subject to license or grant by the state, which is supposed to hold title to all unclaimed lands. Akin to occupancy is the finding of lost things and the trove of a treasure. Lost things have an owner, but laws in various jurisdictions ordinarily attribute ownership to the finder after the lapse of a certain period of time or upon the completion of certain formalities, such as advertisements or reports to the authorities. Likewise, laws provide for the apportionment of a treasure trove between the finder and the owner of the property in which the treasure was hidden.

Owner's right to fruits of earth

Accession. Accession is another broadly recognized mode of acquisition of property rights. It is based on the principle that the ownership of a thing, either movable or immovable, carries with it the right to whatever the thing produces and to certain other things that are united with it, whether naturally or artificially. Thus, the fruits of the earth, whether spontaneous or cultivated, and the increase of animals belong to the owner by right of accession. The ownership of the land carries with it the ownership of all that is directly above and under it, unless the contrary is established by provision of law or contract; therefore, buildings and other constructions erected by trespassers on the land of another become the property of the landowner. Detailed provisions in various legal systems deal with accession to movables, which, apart from the increase of animals, ordinarily takes place in cases of joining materials belonging to different owners; mixing grains or fluids; and, in cases of production of new things, bestowing labour on the materials of another person.

Acquisitive prescription. Acquisitive prescription, a civil law method of acquisition, is predicated on the possession of a thing over a designated period of time with the intention to own it. Acquisition of property rights in immovables ordinarily requires a longer period of possession than the acquisition of property rights in movables. The required period of time may also vary with the nature of possession: a good-faith possessor ordinarily acquires property rights in a shorter period of time than a bad-faith possessor. In common law jurisdictions, the institution of "adverse possession" performs the same function as acquisitive prescription. Technically, adverse possession extinguishes the right of the previous owner and bars his remedy against the possessor; it does not confer title upon the adverse possessor. In all legal systems today, however, those in adverse possession are equally well protected and are able to transfer their rights to their heirs or administrators and such.

Power of eminent domain

Expropriation. Expropriation of property for purposes of public utility, with or without compensation, is known in all contemporary legal systems. Even when made without compensation, expropriation is distinguished from confiscation, the taking of property by the authorities arbitrarily or as a penalty for the violation of law. Ordinarily, constitutional provisions and other legislative texts insist upon notice and the payment of an adequate, fair, or just compensation to the owner for the expropriation to be valid. In effect, expropriation is made in favour of the state, its political subdivisions, or private utilities enjoying the so-called power of eminent domain because of the services they render to the general public. The possibility of expropriation in all legal systems indicates that property rights are not absolute.

Privileges conferred by public authorities. Finally, an original acquisition of property rights occurs when public authorities confer upon certain persons entirely new economic privileges or recognize privileges that had existed in fact but not in law. Examples are grants of property rights to lands of the public domain, including ownership and rights for the exploitation of mineral resources; grants for the exploitation of natural resources, such as hydroelectric energy or radio waves; and patents, registered designs, trademarks, trade names, and copyright, which form so-called industrial or intellectual property. (Material on these subjects can be found in the following articles: COPYRIGHT LAW; TRADEMARK LAW; PATENT LAW.)

Derivative acquisition. All legal systems establish methods for the acquisition of property rights by transfer from a previous owner. The transfer may be voluntary, as in the case of a last will and testament or an agreement between the previous owner and the transferee; it may be involuntary, as in the case of a judicial sale; or it may take place by operation of law, as in the case of intestate succession.

Sale. One of the most prevalent modes of acquisition of property rights is by the contract of sale. A sale involves the transfer of a thing for a sum of money or the promise of a sum of money; if the transfer is for something other than money, the transaction is technically designated as exchange. In legal systems of the French family and in common law jurisdictions, the contract of sale transfers the ownership of the things sold. Insofar as third persons are concerned, however, transfer of ownership may depend on delivery of movables or the recording of the title of immovables. In Roman law and in systems following the German Civil Code, however, the contract of sale merely involves a promise to transfer the peaceable possession of the thing sold; the actual transfer of property rights may depend on delivery or on compliance with certain formalities. In most legal systems, contracts of sale are free of any formalities and may be based on verbal agreement, but the rules of evidence may exclude the proof of certain verbal contracts if the object exceeds a specified value. Moreover, in most contemporary systems, the sale of immovable property must be recorded, either for its validity against third persons or for the actual transfer of rights.

Acquisition through wills and testaments

Donation. Another prevalent mode of acquisition of property rights is by donation. Donations may be *inter vivos,* intended to take effect while the donor and the donee are living, or *mortis* causa—that is, in contemplation of death. Donations of all sorts are ordinarily subject to strict requirements of form intended to ensure that their execution and the intent of the donor are genuine. *Moıtis causa* donations are ordinarily contained in last wills and testaments, which must be executed everywhere in strict compliance with the requisite formalities. Formal requirements are usually dispensed with in cases in which the possession of tangible movables is transferred to the donee. As in the case of sales, the transfer of property rights may be effective as between the donor and the donee upon completion of a donation or it may require for its effect the delivery of movables and the recording of necessary documents.

Judicial Sale. Anothei prevalent mode of acquisition of property rights is by judicial sale. Judicial sales take place in a variety of circumstances, such as in cases when the property of an insolvent debtor is seized and sold for the satisfaction of his creditors, when property is sold by the state or by its political subdivisions for the payment of tax claims, when property is judicially partitioned among co-owners, or when inherited property or that of a minor is sold by an administrator or tutor. Provisions of law ordinarily establish the requisite formalities and rules of substance for the validity of a judicial sale. As a derivative mode of acquisition, the judicial sale transfers to the acquirer only the property rights that the previous owner had.

Intestate succession. Finally, derivative acquisition of property rights takes place by operation of law in cases of intestate succession; that is, when a person dies without leaving a will. Rules of law in developed as well as primitive legal systems determine which rights are heritable and specify the order of succession, which varies with the culture and the structure of society. In Socialist legal systems, the devolution of property by inheritance tends to be restricted. In most other modern legal systems, death taxes have, to some extent, restricted the devolution of large fortunes, but the right of inheritance remains one of the fundamental precepts of law. In civil law systems, the legal heirs of a deceased person are supposed to continue the "personality" of the deceased and to succeed to all of his rights and obligations that are considered to be heritable. Moreover, certain close relatives—whether descendants, ascendants, or a surviving spouse—are entitled to a forced share of the estate, even against the will of the deceased. In common law systems, the property of an intestate is placed under the authority of an administrator who pays all debts and charges and who puts the legal heirs into possession upon completion of the administration. Although the concept of a forced share is foreign to common law, the children and surviving spouse share the estate·of one who dies intestate. Modern legislation in a number of common law jurisdictions has also granted the spouse a right to elect what the law would allow her in the case of intestacy rather than accept the provisions of an unfavourable will.

Forced share in civil law countries

II. Types of property rights

OWNERSHIP

The rights of individuals, groups of individuals, or other entities, for the exclusive enjoyment of economic goods may be designated as rights of ownership. The word ownership has a precise technical meaning in most legal systems and, especially, in civil law countries. It is seldom used in the professional literature of English law and most common law jurisdictions, though in the United States ownership is frequently used synonymously with property. In this article the word ownership will be used in a broad sense to designate rights for the exclusive enjoyment of economic goods.

Rights of ownership may be classified according to a variety of criteria. From the viewpoint of the *subjects* of these rights—that is, the persons or entities that hold them—rights of ownership may be distinguished according to whether they are held by individuals, by tribes and clans, by family groups, by collectives and cooperatives, by unincorporated associations, by corporations, or by the state and public corporations. Each right of ownership in this respect may have peculiar characteristics of its own.

Varieties of owners. *Ownership by individuals.* Ownership by individuals has been recognized by all societies and at all times, although the scope, extent, and incidents of individual claims have varied with the culture and structure of each society. Moreover, variations have been common with respect to the objects that may be owned by individuals. For example, individual ownership of lands or of other means of production may be excluded, but individual ownership of chattels may be allowed.

In developed legal systems, ownership may be vested in a single individual, in which case one speaks of individual ownership purely and simply; or it may be vested in a number of individuals, in which case one speaks of individual co-ownership. Co-ownership has been recognized from early times, and, in a sense, the ownership of lands and chattels by tribes, clans, or other groups, may be regarded as a form of co-ownership. But it was in ancient Rome that the institution of co-ownership by individuals assumed certain typical characteristics that survive in contemporary lega! systems.

Co-ownership

In civil law systems, the share of each individual co-owner is a distinct right of property; it gives rise to a claim for the proportional use or enjoyment of the thing held in common; it is freely disposable unless there is an agreement to the contrary; and the share of each owner devolves to his legal heirs or legatees at death. The rights and obligations of the co-owners are ordinarily determined by directly applicable provisions in civil codes. If one co-owner wishes to sell his part of a thing held in common, as distinguished from selling individual shares, consent of all the co-owners is ordinarily required, unless, of course, the law or contractual provisions establish a different rule. Co-ownership is essentially voluntary; thus, ordinarily co-owners have the right to demand partition of the thing held in common, whether in kind or by judicial sale.

In common law jurisdictions, the institution of co-ownership may assume either the form of an *ownership in common*, corresponding to the civil law notion of co-ownership, or the form of *joint ownership*, which is an original common law institution. The difference between the two forms of co-ownership depends upon the manner in which an individual interest devolves upon the death of a co-owner. If an owner in common dies, his share passes to his successors, whether by will or on an intestacy; but if a joint owner dies, his interest accrues to the remaining owners, so that when all owners but one are dead the survivor becomes the sole owner.

Ownership by tribes and clans. In tribal societies, ownership by individuals tends to be limited to certain chattels only. Apart from cultural, geographical, and historical variations, the exclusive control of lands and of .most chattels is vested in the tribal community. Tribes on the move and tribes engaged in agriculture as an occasional pursuit have been known to assign lands to clans or kindreds for temporary occupation. In a more settled state of society, portions of the tribal territory may be assigned for permanent occupation, but occasional redistributions of land indicate that the overlordship remains vested in the tribe. Once a tribe is fully settled, there is frequently a gradual transition from tribal communalism to clan or kindred communalism; that is, transfer of the ownership of lands from the tribe to the clans or kindreds.

Vestiges of tribal and of clan ownership persisted in most European countries in the form of the open-field system, until enclosures put an end to it in relatively modern times. This system was based on the idea that the ownership of land was vested in the collective body, which was charged with the duty to see that every individual or household would have the means and opportunity for profitable labour. Survival of tribal custom may still be observed in isolated parts of the continent in the form of village ownership of pastures and forests.

Effect of the enclosures

Ownership by family groups. Ownership by family groups is predicated on the existence of an "extended" family—a family of many blood relatives—that tends to curtail individual interests in order to preserve the unity of the household. Extended family arrangements have been prevalent in India, where they continue to exist today; extended-family arrangements were known to exist among the southern Slavs in the Balkan Peninsula up to the first part of this century, and comparable institutions have been observed in the early history of the Germans, the eastern Slavs, and certain Romance nations.

Since Indian family ownership contains features that distinguish it from property institutions of the Western world, mention of these features is appropriate here. Ordinarily the property of an extended family in India is owned by the body of the members. The head of the family has merely the management of the propeity rather than title to it. When the head of the family dies, the members may not be said to be heirs, because they do not take property that did not already belong to them. The only change is that the family is reduced in size, a result that would also follow the death or departure of a junior member. According to some local customs, the head of the family may be regarded as nominal owner of the property, but even in this case the ownership of the head of the family is subject to stringent collective controls that exclude the idea of individual ownership in the Western sense.

Therights of individual members

The rights of individual members may be described as undivided shares in the whole of the family property. In principle, shares are assigned to members in the light of their relative seniority and are given to branches of the family rather than to individuals. Members of junior generations may not get their shares until the death of their parents, who, as long as they live, represent the whole branch.

Since ordinarily the head of the family is manager rather than owner of the family property, his acts are supervised and at times overruled by the body of members. As a rule, alienation or encumbrances such as sales or mortgages require the consent of all members.

The purpose of the institution of family ownership is to secure the means of subsistence for all of the members of the extended family. As a consequence, the members have a right of maintenance that is quite alien to Western views: weaker members of the family are given the means of support without regard to their contribution of labour; spinsters are provided with a dowry; and children are given the means of education. To achieve these ends, the family sets aside a certain amount of the property or its income and does not allow such funds to be used for other purposes. The beneficiaries, however, have no direct property interest in the funds; their position, in effect, is similar to that of beneficiaries of charitable institutions.

Family ownership does not preclude individual ownership. In certain circumstances, family ownership may be converted into individual ownership by dividing the assets among the individual members. Moreover, in the framework of any extended family, there may be a distinction among objects belonging to individual members (as a result of their personal labours) and objects belonging to the family community as a whole.

Ownership by collectives and cooperatives. Property owned by collectives and cooperatives is found in primitive as well as in developed legal systems and in capitalistic as well as in socialistic countries. Since the beginning of the 19th century, this type of property has proliferated as a result of pronounced movements toward collectivism and cooperation.

In capitalist countries, collectives and cooperatives are associations of individuals, such as farmers, labourers, consumers, home owners, or small entrepreneurs, formed for the pursuit of some productive enterprise, the benefits of which are to be shared in accordance with the capital or labour contributed by each member. The rights and obligations of members may be specified in contractual provisions or in legislation. Although these associations are ordinarily formed voluntarily, there are instances in certain countries where they may be formed under legal compulsion. In the United States during the 20th century, the most common type of cooperatives have been formed for the production and distribution of electricity in rural areas and for the building and enjoyment of apartment units in metropolitan areas.

In Socialist countries, collectives and cooperatives of a variety of sorts flourish as realizations of Marxist theory under the protection or under the compulsion of law. In Socialist countries, the property of cooperatives exists alongside personal property and property belonging to the state. The prime example of cooperative property in the U.S.S.R. is that of the kolkhozes, or collective farms, which have the perpetual right of enjoyment of state lands. The kolkhozes must be organized and administered in strict compliance with legislation; they are bound to cultivate or exploit the lands assigned to them in certain designated ways and to give certain payment to the state; and in this respect they differ from corresponding institutions in capitalistic countries.

Rights of kolkhozes

Ownership by unincorporated associations. According to the Roman law tradition, rights of ownership, as indeed any rights and obligations, may exist only in those entities possessing legal personality. Such entities are living human beings (natural persons), associations of human beings, and foundations established for the realization of a general interest (juristic persons). The state and its political subdivisions, profit and nonprofit corporations, and partnerships are examples of associations capable of possessing legal personality and of owning property. Hospitals and charitable institutions are examples of foundations likewise capable of possessing legal personality and of owning property.

Unincorporated associations formed for the pursuit of vocational, artistic, scientific, or religious purposes may or may not possess legal personality depending on applicable laws and on compliance with required formalities. If they do not possess legal personality, they cannot hold property in their own name; their property belongs directly to the members of the association and is usually held in undivided shares. If they possess legal personality, they have property of their own that is distinct and distinguishable from the individual property of the members.

Ownership by corporations. Corporations are juristic persons—that is, artificial entities formed by human beings to which the law attributes legal personality for the pursuit of social and economic purposes. They have no physical existence, but, as persons, they participate in legal life and are capable of holding property in their own name. There are various sorts of corporations: private and public, secular and religious, and profit and nonprofit corporations.

The proliferation of corporations since the 19th century has been said to reflect, throughout the world, a movement from the individual ownership of capital to the collective ownership of capital. This movement has been observed in capitalist countries, in which wealth tends to be concentrated in the hands of private corporations, as well as in Socialist countries, in which wealth is concentrated in the hands of the state and other public corporations.

Distinction between property of corporation and shareholder

The corporate device allows for a clear distinction between the property of the corporation and the property of a shareholder. Technically, no one owns the corporation, which has an independent existence of its own. A share merely represents a fractional interest in the property of the corporation and confers a corresponding interest in its administration, and even when all shares are concentrated in the hands of a single person, there is still a clear distinction between the property of the corporation and the property of the sole owner of its shares. When incorporation is completed, the individual property of a shareholder is completely insulated from the risks of the corporate enterprise. Creditors of the corporation may not take the property of a shareholder for the satisfaction of their claims, nor may creditors of the shareholder take the corporate assets. The only risk that a shareholder assumes is that represented by the amount of his investment in the corporation, and it is only his own shares that constitute property that may be seized by his own creditors.

Ownership by state and state corporations. In all legal systems, the state and its political subdivisions are political corporations that have the capacity to own property. This property may be of two sorts: public property (such as navigable waterways, national parks, and monuments), which is ordinarily inalienable and is held in trust for the benefit of all citizens, and private property (such as state-owned enterprises and buildings housing various agencies), which does not differ in essence from property held by private persons.

In modern times, states have asserted ownership over a variety of things that previously were considered to be without owner, such as running waters and wild animals. This form of state ownership, asserted in an effort at conservation of natural resources, confers mainly administrative advantages and prevents the private appropriation or ownership of certain assets of society except under regulations that protect the social interest.

In socialist countries, lands, industrial capital, and vari-

Public corporations

ous service fields are ordinarily owned by the state and its political subdivisions to the exclusion of private interests. In capitalist countries, the state and its political subdivisions may enjoy certain economic monopolies and may directly own a variety of enterprises whether in competition or in cooperation with private interests. Moreover, the state and its political subdivisions may undertake economic activities through a variety of public corporations. Thus, for example, postal services, telecommunications, and the exploitation of hydraulic energy may be in the hands of public corporations. In recent times, there has been much experimentation toward the development of public corporations with the cooperation of private capital; the state may preserve the control and administration of a public corporation in a service field, but the profits of the enterprise may be shared by private shareholders.

Legal distinctions between movable (personal) and immovable (real) properly. In civil law systems, in mixed jurisdictions, and in Scandinavian countries, things are divided into movables and immovables, a division that was also known in ancient legal systems. In common law jurisdictions, property (rather than things) is divided into personal property and rea! property (realty), but these terms may be taken as roughly equivalent to the civil law notions of movables and immovables. In socialist legal systems, the traditional divisions of things into movables and immovables tends to be suppressed. In the U.S.S.R., this division has been expressly discarded and replaced by new classifications reflecting prevailing economic and political theory. The significance of the division of things into movables and immovables lies primarily in the fact that special rules of property law apply to the various categories of things classified as immovables. Almost universally, the acquisition, possession, transfer, or encumbrance of immovables is subject to rules of property law that are materially different from those governing movables. Moreover, immovables may be subject to special rules in the fields of obligations and contracts, family law, successions, civil procedure, taxation, criminal law, and conflict of laws.

Special position of land

The rules of property law applicable to immovables tend, on the one hand, to delimit narrowly individual rights in this type of wealth and, on the other hand, to enhance security of title. In modern times, there has been more emphasis on security of title, but, even today, rights in immovables may be subject to special limitations concerning acquisition, alienation, seizure, testamentary disposition, and partition. The modem trend toward security of title has led to an increased emphasis on public records. In central European countries and in part's of the British Commonwealth the very existence of interests in immovables depends, ordinarily, on the recording of formal acts in land registers. Moreover, all entries in the land registers are accorded full faith and credit (that is, are accepted in other states or countries) so that innocent third parties may rely upon them. In the United States, in legal systems of the French family, in Scandinavian countries, and in other parts of the world in which a land register system has not as yet been established, the transfer or encumbrance of immovable property is ordinarily effective toward third persons from the time pertinent documents are placed in public records. The degree of faith and credit accorded to entries in these records varies from country to country.

Since ancient times and up to the era of the Industrial Revolution, land (an immovable) was regarded as the most important type of wealth; hence, particular rules were developed to safeguard interests connected with the possession, use, and enjoyment of land. In modern times, however, economic emphasis has shifted to values other than land, and the law tends to accord to these values the type of protection traditionally reserved for land.

Ancient divisions of property. In ancient Roman law, emphasis on the relative significance of the elements of wealth led to the division of things into res mancipi and res nec mancipi. Land subject to Roman ownership, cattle, beasts of draft and burden, and rural servitudes attached to land subject to Roman ownership were classified, in the light of their importance for the realization of economic, political, and social interests, as res mancipi; all other elements of wealth were res nec mancipi. This division of things was kept alive by classical jurists in spite of profound changes in the economy and society. The Emperor Justinian formally abolished it in the 6th century and replaced it with the distinction between movables and immovables, a distinction that had been known but was relatively insignificant under earlier law.

Roman law scholars elaborating on Justinian's *Corpus juris civilis* during the era of reception of Roman law in continental Europe (14th to 16th centuries) paid little attention to the old division of things and quite naturally focussed attention on the distinction between movables and immovables. This distinction was elaborated further in the following centuries and found its way into the developing legal systems of various continental countries.

Modern variations. In legal systems of the French family, the distinction between movables and immovables rests, in principle, on physical notions of mobility and on inherent characteristics of things. An immovable is defined as a thing having a fixed place in space, and a movable as one that can either move itself or be removed in space. For reasons of policy, however, the law may treat as movables things that according to lay notions could be regarded as part of an immovable, such as standing crops. On the other hand, the law may qualify as immovable things broadly considered to be movables, such as farm implements and animals. Rights that attach to an immovable object are also immovables.

In Germany, and in legal systems following the model of the German Civil Code, immovables are tracts of land and their essential component parts; things that are neither tracts of land nor essential component parts of tracts of land are movables. Since things in these legal systems are defined as tangible objects, rights are neither movables nor immovables; nevertheless, for certain practical purposes, rights closely connected with immovables may be subject to the rules governing immovable property.

In England, the distinction of property into personal and real was considered as late as 1925 to be the keystone of property law. The distinction arose in the formative era of the common law and corresponds roughly to the distinction between feudal and nonfeudal property. Because land was the foundation of feudalism, the distinction between real property and personal property corresponds roughly to that between land and chattels. The distinction between the two kinds of property was originally procedural; lawyers distinguished between real actions, tending to secure the recovery of property in kind, and personal actions, tending to secure the payment of damages for unlawful interference with one's possessions. Because feudal relations were largely determined by the manner in which particular tracts of land were held, it was necessary for an ousted landholder to recover his own particular tract. Hence, real actions were largely available for the protection of interests in land, whereas personal actions were available for the protection of interests in chattels; in time, the terms real and personal property came to denote the objects of property rights protected by these actions.

Real property

Real property (realty) may thus be defined as interests in land other than leasehold interests (held for a definite duration); personal property (personalty) includes interests in movables and leasehold interests in land. Each type of property was, in effect, governed by a distinct set of rules. With the exception of certain technical rules of minor significance, however, the differences between real and personal property were swept away in England as a result of statutes enacted after 1925. The differences between the notions of movables and immovables in civil law and real and personal property in English common law have thus been minimized. In spite of differences in methodology and conceptual technique, the law of real property in England today performs a function essentially similar to that performed by the law of immovable property in civil law jurisdictions.

Similarly, legislation in the United States and in other common law jurisdictions has simplified the law of real property and has transformed it into a law applicable to land. The distinction between real and personal property is still drawn in the United States in the light of the historical past, but it has lost most of its original significance. In contemporary American practice, distinction is made between land and movables, and modern treatises of property law include consideration of both elements of wealth.

In the U.S.S.R., the distinction between movables and immovables was expressly abolished by the Russian Civil Code of 1922, which was largely reproduced in the various civil codes of the constituent republics. Since all land had been nationalized and, therefore, removed from commerce, it was thought that the distinction between movables and immovables had lost its reason of existence. Today, property in the U.S.S.R. and other Socialist countries is divided into goods of production and goods of consumption and into personal property, cooperative property, and socialist property. As such, these distinctions reflect Marxist philosophy. Land, which can no longer be privately owned, is subject to special rules. Buildings and other constructions may be personal property, cooperative property, or socialist property, depending on their function. In each case, they are subject to appropriate legislation rather than a common set of rules, although it is true that residential homes, which may be privately owned, are subject to a number of special rules that are inapplicable to other types of personal property. The transfer of a residential home, for example, is subject to the requirements of notarial act and registration in public records, whereas the alienation of other things is effected by informal agreement and delivery.

Choice of law. The laws of a state or country apply, in principle, to all persons, acts, and things within its geographical limits. This principle of the territoriality of laws, however, is not absolute. In cases involving foreign elements, courts must decide whether they have jurisdiction to pass upon the controversy before them and whether the controversy is governed by the court's own law or by some foreign law. These questions form part of the discipline known as conflict of laws (*q.v.*) or private international law.

Law of the situs

The division of things into movables and immovables has legal consequences in the field of conflicts of laws. According to a universally recognized rule of choice of law, the determination of whether a particular thing is movable or immovable and the determination of most disputes concerning rights in immovable property are made by application of the law of the place in which the property is situated—the law of the situs. In common law jurisdictions, almost all claims with respect to iinmovable property, whether they arise from contractual transactions or from inheritance, are said to be governed by the law of the situs. Because common law courts refuse to exercise jurisdiction when immovable property is situated in a foreign state or country, this law is ordinarily the court's own law. In civil law countries the law of the situs ordinarily governs claims arising from transactions among living persons only; succession rights in immovable property are governed, as a rule, by the law of the nationality of the deceased. In contrast with common law courts, courts in civil law countries exercise jurisdiction quite freely when immovable property involved in litigation is situated on foreign soil, applying the law of the situs.

Choice of law in civil law countries

There is a variety of law rules in existence for movable property. The traditional common law rule is that movables, having no fixed location, are governed by the law of the domicile of their owner. In modern times, however, the courts tend to pay lip service to the traditional rule while they actually fashion new rules; movables are now governed by such criteria as their original or current placement or the place where the parties involved carried out their transactions, depending on particular facts and circumstances. In civil law countries, as a rule, the creation, modification, or termination of rights in individual corporeal movables is determined by application of the law of the situs. Succession rights in movable property, however, are ordinarily determined by application of the law of the nationality of the deceased. In these countries, the division of things into movables and immovables tends to lose much of its significance in conflicts law, since, as a rule, all types of property are governed either by the law of the situs or by the law of the nationality of the deceased. In all jurisdictions, rights in industrial property are governed either by the law of the place in which these rights were created or by directly applicable international conventions.

Legal **distinctions** between tangible **and** intangible property. The Romans classified objects of property rights as either corporeal or incorporeal. Physical objects that could be felt or touched were given as illustrations of corporeals. Incorporeals were abstract conceptions, objects having no physical existence but having monetary value; the illustrations given were rights of various kinds, among them inheritance, obligations, and all proprietary interests in things with the exception of the right of ownership, which was considered as corporeal because its object was corporeal.

In legal systems of the French family, which follow the Roman tradition, things are divided into corporeals and incorporeals. A thing is corporeal if it is perceivable by any of the senses; incorporeals are rights that are conceived by the intellect. Under the German Civil Code, however, and those modelled after it, ownership and other proprietary interests may exist for corporeal objects only. Incorporeal objects, such as rights having monetary value, though part of a person's patrimony, are not governed by the law of property.

Types of intangible property

In common law jurisdictions, a distinction is made between tangible and intangible property that corresponds roughly to the Roman division of things into corporeals and incorporeals but does not coincide with it. Tangible property includes lands and certain chattels known as choses in possession; that is, rights in definite tangible things over which possession may be taken. Intangible property is largely one of two types: (1) choses in action, namely, rights of property that can only be claimed or enforced by legal action and not by taking physical possession, such as bank accounts; debts generally; stocks and shares; and industrial property, including patents, registered designs, trademarks and trade names, and copyright and (2) *incorporeal hereditamenis,* namely, heritable rights, or rights passed by way of descent to heirs, such as those relating to lands, buildings, minerals, trees, and all other things that are part or affixed to land and also easements, profits, and rents.

Servitudes and easements. From among the various species of incorporeal things or intangible property, attention may be focussed at this point on servitudes in civil law jurisdictions and on easements and profits in common law jurisdictions. A servitude is a proprietary interest, which amounts to a real right on the property of another person. It may be a charge laid on one estate in favour of another estate, such as a right of passage, or a right of diverting water or drainage through the other estate. Or it may be a charge laid on an estate in favour of a person, such as a right of usufruct, use, or habitation. An easement may be defined roughly as a property right in a person or a group of persons to use the land of another for special purposes, such as rights of way, rights of diverting the course of a stream for irrigation purposes, and rights for the support of buildings—that is, any right not inconsistent with the general property right of the owner of the land.

Profits. A profit, technically known as *profit-d-prendre,* is the right to use another's land by removing a portion of the soil or its products; it includes the right to take fish or game; to pasture one's animals on the land of another; and to take wood, sand, coal, and other mineral substances. At least in American jurisdictions, easements as well as profits may be established in favour of an estate or in favour of a designated person.

Rights of way. Rights of passage or rights of way are some of the most frequently encountered servitudes and easements. These are rights to pass over the land of another. They are not natural rights and must be created by application of one of the methods that are available for the acquisition of property rights. Rights of passage or of way may be either public, for all persons to enjoy, or

Water rights

private, for the benefit of a designated person or a class of persons. Rights with respect to the use of waters are likewise some of the most frequently encountered servitudes or easements. According to civil law conceptions, which have been broadly adopted in Anglo-American jurisdictions, landowners have certain well-defined rights with respect to surface waters; underground waters; and natural watercourses, such as rivers, streams, and lakes traversing or bordering their lands. Modifications of these well-defined rights may take the form of servitudes or easements. Thus, the owner of an estate situated above may acquire, from the owner of an estate situated below, the right to pollute a natural stream traversing both estates or the right to obstruct the flow of the stream or the flow of surface waters; the owner of the upper estate may acquire the right to use an unreasonably great amount of water for his own purposes or to discharge drain and refuse waters upon the lower estate; and any landowner may acquire the right to divert the flow of natural water courses and to maintain on another estate an aqueduct for his needs.

Mineral rights

Finally, mineral rights, especially rights for the exploitation of oil and gas, may take the form of servitudes in civil law jurisdictions and of profits in common law jurisdictions. In most parts of the world, however, mineral exploration depends on prior license, concession, franchise, or grant by the state that has asserted its ownership of valuable natural resources. Only seldom are oil and gas rights profits in American jurisdictions; ordinarily, persons undertaking such operations acquire from the owner of the land the ownership of the minerals in place.

TEMPORAL DIVISION OF OWNERSHIP

In all legal systems, the right of ownership is susceptible to division among co-owners holding undivided interests over the same thing. In common law juridictions, however, ownership is capable of being divided in a great variety of ways from the viewpoint of time as well as from the viewpoint of space. The law governing temporal division was worked out in the first instance for land and later applied to funds comprising both land and certain chattel interests of a permanent nature, such as stocks and shares.

Common law: estates. The historical basis of common law property is that only the Crown can own land. A landowner, strictly speaking, does not own land but a ***time*** in the land, an interest called an ***estate.*** Lawyers manipulated this concept of time in the land and devised an intricate system of estates, which may be either successive in time or concurrent, in the sense that two or more persons may own property jointly or they may have separate rights of ownership in the same property at the same time. Ownership is not attached to the land itself but to an abstract entity, the estate, that is interposed between the tenant and the land. The estate is purely conceptual, yet it is treated by the law as if it were a real thing.

Three simple classifications of estates may be mentioned at this point. The first is into estates in ***possession*** and estates in ***expectancy.*** If a person's estate gives him a right to the immediate possession of the land, it is said to be an estate in possession; if, on the other hand, he must wait for his right to possession to take effect, he has only an estate in expectancy. The second classification concerns only estates in expectancy, which are divided into ***remainders*** and ***reversions.*** A remainder is an estate that becomes effective when the estate of a previous owner expires. A reversion is similar to a remainder in that the right to enjoy the land is postponed to a future date; when there is a reversion, however, the land eventually returns to the grantor, whereas with a remainder the land goes to some other person. The third classification is into ***freehold*** and ***leasehold*** estates, the former having an indefinite and the latter a definite duration.

Fee simple, fee tail, and life estate

Freehold estates. The simplest examples of freehold estates are the estate in fee simple, the estate tail, and the estate for life. The estate in fee simple is, in the ordinary course of events, the nearest thing to full ownership of land found in common law jurisdictions. It confers full rights of possession, enjoyment, and disposition during life and by will. If the owner of the estate in fee simple dies intestate, the land passes to the relatives entitled under statute in that event. Only in the event of the owner dying intestate and without relatives will the estate come to an end and will the land pass to the Crown in England or to the state in other common law jurisdictions. The estate tail or fee tail is an estate less than a fee simple in the sense that it is limited to the person and the heirs of his body. The tenant in tail has full rights of possession and enjoyment, and the estate does not come to an end at his death; it passes to his heirs but only a limited class of heirs, his descendants. The line of descent may be further limited by allowing the estate to descend only through males or through females. If the tenant dies without leaving descendants, the estate may revert to the grantor, or it may go to someone enjoying a remainder. A life estate is one whose duration may be measured by the lifetime of the tenant or of another person. It may also be cut off by the occurrence of a specified contingency, as, for example, the remarriage of the life tenant. The life tenant has the right to obtain the profits of the property, the right to possession, and the right to dispose of his interest. When a life estate is created, there is always a corresponding estate in expectancy that will come into existence when the life estate is terminated. The owner of this estate may have rights of reversion or of remainder.

Leaseholds. Leaseholds or estates less than freehold are those that have a definite duration or a duration that may be made definite a the will of one of the parties concerned. Not all leaseholds arise from leases in the strict sense of the word; some arise from other arrangements and are more properly called ***tenancies.*** The most important and the most frequent are the so-called periodical tenancies, for example, from week to week, from month to month, or from year to year; they may be terminated by notice given by the landlord or by the tenant only on the anniversary of the original grant. When the proper day for notice passes without notice being given, the tenancy is renewed for an additional definite term. Other possible tenancies are those at will and at sufferance. A tenancy at will may terminate at any moment by notice given either by the landlord or by the tenant; a tenancy by sufferance is one held by a person whose lawful term has ended and who continues to possess without right.

Rights of reversion

Whenever a person holds a terminable estate, such as a life estate, questions arise as to his duties toward persons holding rights of reversion or remainder. Ordinarily, a life tenant is entitled only to the income of the property, including the physical use of tangible things; he may not touch the capital that must be transmitted intact at the termination of his interest. Of course, it is frequently difficult to draw a clear line of distinction between capital and income, and the problem is even more complicated when the enjoyment of land is divided among several persons successively. To resolve this difficulty in a practical manner, and in order to afford a measure of protection to the interests of successive holders of land, courts developed remedies for waste. The ever-present temptation of a temporary holder of land to exhaust the soil or to let buildings fall into disrepair may be checked by an action for damages or, even more important, by an injunction restraining the commission of waste.

Statutes enacted in various jurisdictions since the 19th century have worked out substantial modifications of the common law governing estates over land. In most American jurisdictions, the estate in fee tail has been practically abolished, and landholdings have been declared to be allodial (owned and heritable without obligation). In England, the Law of Property Act of 1925 reduced the number of legal estates that can exist over land to two, namely, (1) a fee simple absolute in possession, and *(2)* possession for an unlimited period of time; the number of legal interests or charges was also reduced to five years. All other estates, interests. and charges must exist in trusts (see below ***Ownership for the benefit of others: trusts as property),*** the trustees having the legal estate in fee simple.

Civil law. Compared with the common law doctrine of estates, the institution of ownership in civil law systems is the essence of simplicity. No such abstraction as an estate is interposed between the owner and his property, nor is ownership built on ownership as in the case of a trust. Thus, there is no temporal division of ownership; a person either owns a thing or he does not; it is as simple as that. The content of ownership may be limited, as when burdens on the land such as servitudes are created, but these burdens are regarded as restrictions on the use of the land rather than as rights of separate or concurrent ownership.

Dominium: absolute ownership. The idea of absolute or at least exclusive ownership has been inherited from Roman law. One of the most striking institutions of that legal system was dominium, upon which all the civil law systems have modelled their treatment of ownership. Dominium was, in the final form that it received in the Justinian legislation, as near to being absolute as any institution of private law can be. The owner had an almost absolute right to use, enjoy, and dispose of a thing as he saw fit, subject to exceedingly few restrictions imposed by the state. The kinds of encumbrances (servitudes, mortgages) with which a thing could be burdened were kept to the lowest possible number, and when created they were carefully distinguished from ownership; upon the disappearance of an encumbrance, ownership resumed its original plenitude, leaving no room for the concept of a remainder or a reversionary interest. In contemporary civil law systems, ownership is ordinarily defined as an exclusive right that exhausts the utility of a thing. This right may be either perfect or imperfect. Perfect or full ownership implies that the owner has the prerogatives of use, enjoyment, and disposition of a thing. Imperfect or naked ownership exists when the power of disposing of the property rests with the owner hut when the use or enjoyment is vested in whole or in part in another person. The right of ownership may be dismembered in a number of ways specified by law. Among the permissible dismemberments of ownership in any civil law jurisdiction are included servitudes and real security rights, such as pawn, pledge, and mortgage. All dismemberments of ownership are proprietary interests, which, by their nature, confer direct and immediate authority over a thing belonging to another person.

Perfect and imperfect ownership

Usufruct. Prominent among the personal servitudes in any civil law jurisdiction is the right of perfect usufruct, which corresponds roughly with the concept of a life estate in common law jurisdictions. A perfect usufruct is a real right of enjoyment of limited duration that is exercised on a thing belonging to another person. The usufructuary has a right to draw from the thing all the advantages and utility that it may produce, as if he were owner, but he must restore the thing to the owner or his successors at the end of the usufruct. The creation of a usufruct does not deprive the owner of his ownership; it merely deprives him of the prerogatives of exclusive use and enjoyment.

The difference between usufruct and life estate relates to theory and structure rather than to function. In spite of historical, theoretical, and structural differences, civil law and common law tend to give in most instances strikingly similar solutions to problems of everyday life. A comparison of the function of the life estate with usufruct furnishes an illustration: the two systems, though starting with different basic premises have arrived at substantially the same position.

Obligations to owner or successive heir

Although technically there are neither remainders nor reversions in civil law systems, there are persons whose positions correspond roughly with those who enjoy remainder or reversion. When the proprietary interest of a person in a piece of property is certain to terminate, as in the case of a usufruct, the property eventually is restored to the owner or to another person. Further, cases of terminable ownership are not rare in civil law jurisdictions: any ownership acquired by juridical act may be stipulated to terminate upon the lapse of a term or upon the occurrence of a condition. The duties of the temporary owner toward his successor are ordinarily governed by contractual stipulations rather than by rules of property law, but there is a notable exception concerning the duties of an heir instituted by a will toward the person (successive heir) designated by the will to receive the property after the death of the instituted heir. His obligations to preserve and transmit the inheritance to the persons entitled to take it after him are governed by the rules of succession.

The usufructuary is under obligation to ensure that the property will be restored intact to the owner at the cessation of his interest. He must not destroy or alter the character of the buildings or of the land, and, if the land is agricultural, he may not even change the type of farming practiced. Further, the usufructuary is under obligation to keep the property in a good state of repair, though in this respect there are some variations in detail from country to country. The instituted heir is entitled to use and enjoy the property as owner, but, like the usufructuary, he is under duty to manage the property with a certain degree of care.

OWNERSHIP FOR THE BENEFIT OF OTHERS: TRUSTS AS PROPERTY

In all legal systems, the ownership of a thing normally carries with it the power to enjoy the thing and to manage it; but, since not all persons are capable or willing to manage their property, the law allows management to be detached from the enjoyment of the property. This may be accomplished everywhere by vesting property in an artificial person, such as a corporation or a foundation. There is no necessity, however, to interpose an artificial person between human beings and a fund; property may be vested in human beings for the benefit of other human beings or indeed for themselves. This is accomplished in common law jurisdictions by means of a trust.

A trust is a legal relationship whereby a person called *trustee* undertakes the obligation to deal with property over which he has control (called the *trust property)* for the benefit of persons called *beneficiaries,* of whom he may himself be one. Generally speaking, all kinds of property, real or personal and tangible or intangible, may be held in trust; but the property most frequently so held includes land, stocks, and shares. A trust generally involves separation of the management from the beneficial enjoyment of property. although in civil law jurisdictions this can occur only when a person is incompetent to manage his own affairs because of absence, age, or unsound mind.

The separation of management and beneficial enjoyment. In all cases in which management and beneficial enjoyment are separated, the question arises as to the nature of the respective interests of the manager and of the person having the beneficial enjoyment. It might be possible to say that the manager owns the property but that he is under duty to manage it for the benefit of another who has nothing more than a personal right to take action against the manager. It might also be possible to say that the beneficiary owns the property, but the manager has full powers of administration without any proprietary interest in the property. This second possibility exists in civil law jurisdictions, where the property administered by the tutor of a minor or by the curator of a person of unsound mind belongs to the incompetent. Neither possibility exists in common law jurisdictions. In these jurisdictions, in effect, both the trustees and the beneficiaries own the property in different ways; or, more accurately, neither owns the property in the sense of the Roman dominium, but each owns an interest in it, called, respectively, the legal estate and the equitable interest.

For historical reasons as well as functional purposes, it became necessary to treat the right of the beneficiary as a separate equitable estate comparable to the legal estate; the desire to make a temporal division that had led to the development of the doctrine of estates was also found in the area of trust property. Thus there is not one object of ownership, a physical thing, but two separate abstract things, the legal estate owned by the trustee for the purpose of managing the property and the equitable interest owned by the beneficiary for the purpose of enjoying the same. The legal estate is a way of explaining that the trustees may act commercially as owners of the property, enjoying wide powers of alienating it in the market; the equitable estate means that the beneficiaries have the beneficial ownership, which implies that they can enjoy the use and possession of it and draw an income from it.

Obligations of trustees. Property may be given to trustees for the benefit of persons or for the accomplishment of certain purposes, usually charitable. In either case, the trustees may not take any benefit themselves in their character as trustees, unless they are entitled to charge for their services under the instrument creating the trust. They must obey the terms of the trust and are under personal duty toward the beneficiaries for the proper management of the trust property. If the trustees do not carry out their duties toward the beneficiaries, the beneficiaries may bring proceedings in a court of equity to have the trust enforced.

Rights of beneficiaries

The fragmentation of ownership into a legal estate and an equitable interest is an original common law institution that serves a variety of functions. There is no equivalent institution in civil law jurisdictions, unless one may regard as equivalent a substitution or fideicommissum, whereby an instituted heir or legatee is bound to restore the inheritance to another person. Yet it ought to be noted that the instituted heir, though temporary owner of property, has both management and bexeficial enjoyment as long as his right lasts. In recent years, because of economic considerations, the institution of trusts has been introduced by legislation in a limited number of civil law jurisdictions having close ties with the United States or with countries of the British Commonwealth (*e.g.*, Mexico and Liechtenstein). As a single device to accomplish a variety of purposes, the trust has no equal, but nearly everything that the trust can accomplish may actually be achieved within the framework of a civil law system, the policies of the law permitting. When something achievable with the trust is not achievable in civil law jurisdictions, it is usually because the result is forbidden for reasons of social policy.

Vesting of titles. The law does not tolerate a high degree of uncertainty as to who is entitled to the enjoyment of material goods, and it develops devices tending to prevent goods from being kept out of the market for long periods of time. These devices relate to the *vesting* of titles. In civil law jurisdictions and in Socialist countries, the rules governing the vesting of property rights are very simple. Property at any given moment must be owned either by a natural person or by an artificial person; contingent ownerships, likely to arise upon the lapse of a term or upon the occurrence of a condition, are ordinarily based on transactions among living persons. As a rule, property must be transmitted by contract, by will, or by operation of law to a person that is living or that at least is conceived in the womb.

In common law jurisdictions, the word vesting has different meanings, and the rules concerning the vesting of property rights are more complicated than in civil law jurisdictions. All interests in property are divided into vested interests and contingent interests. For an interest to be vested, the person entitled to it need not have a right to the immediate possession of land or to draw an income immediately from a fund; all that is required is that the nature and amount of the interest and the identity of the person entitled to it should be known. If the identity of the person or the amount of his interest is not yet known and will be known only upon the happening of a contingency, the interest is qualified as contingent.

RESTRICTIONS ON THE OWNERSHIP OF PROPERTY

In all legal systems, the scope, incidents, and content of ownership are subject to restrictions made in the general interest of society in the enjoyment of its material goods. It is in the light of these restrictions that ownership is ordinarily qualified as an exclusive rather than an absolute right. Some of these restrictions are imposed by rules of public law, others by rules of private law, and still others by the effect of lawfully executed private covenants.

Limitations on what may be owned. In the first place, there are limitations everywhere as to the kinds of things that may become objects of property rights. In almost all modern civilized legal systems (the exceptions being such aberrations as Nazi Germany), human beings may not be owned by other human beings. According to modern conceptions, a living human body and any member or part thereof are regarded as incidents of a comprehensive "right of personality" rather than objects of property rights, other than that one may have an ownership right in his own body. Contemporary legislation, jurisprudence, and doctrine indicate that transactions concerning parts or members of human bodies, such as donations of blood or legacies of eyes, are valid, unless, of course, they conflict with social mores. With regard to dead bodies, the possibility of private ownership is broadly admitted, although such bodies are ordinarily subject to compulsory disposal by burial or cremation. Apart from human bodies and members or parts thereof, there are things that may not become objects of property rights either because of a physical impossibility of appropriation or because the law so provides. Running waters, the atmospheric air, and the open sea, for example, may not be appropriated in their entirety by an individual or artificial person. According to traditional civilian ideas, these are common things and insusceptible of any ownership.

Rights of states to property. Second, the state may, and frequently does, assert primary rights to property, thns excluding from the sphere of private ownership a wide variety of elements of wealth. In civil law as well as in common law jurisdictions, the state usually asserts its ownership of the seashore, the continental shelf, and inland navigable waters and bottoms. These are ordinarily designated as public things or public property held by the state in trust for the benefit of all. Further, the state may assert its ownership of a variety of mineral substances (including, in some jurisdictions, oil and gas), unclaimed lands, and such natural resources as hydroelectric energy and radio waves; and it may monopolize a growing number of service fields, such as telecommunications. These are ordinarily designated as the private property of the state or of public corporations. In recent years, statutes have been enacted in various capitalist countries asserting state ownership over running waters and wildlife—things that have been traditionally regarded as belonging to no one in particular. This form of state ownership is a new conception, giving expression to the demand for conservation of natural resources, which are assets of society and capable of appropriation only under regulations that protect the general interest.

State ownership of certain minerals

The state's assertion of primary rights to property is much more intense in Socialist countries, where the bulk of the wealth constitutes Socialist property. Land, productive capital, and most service fields belong, as a rule, to the state, its political subdivision, or to public corporations. The kinds of objects that may become personal property of individuals are thus severely limited in comparison with capitalist countries, and there are even limitations as to the quantity of things that a citizen may own. In the U.S.S.R., for example, a citizen may own only one dwelling, be it a house or an apartment. In no circumstances may personal property be used to derive nonlabour income, such as rents, dividends, or interests.

Limitations on uses of property. Finally, in all legal systems there are limitations pertaining to the use of lands by private landowners. Until the era of the Industrial Revolution, lands constituted the bulk of wealth and the foundation of the social order; hence, there was a need for limitations that would safeguard the general interest. In spite of changed economic conditions, most of these limitations are still relevant and continue to prevail. They may derive from rules of public law, such as zoning regulations and provisions concerning the expropriation of lands for purposes of public utility; from rules of private law, such as provisions prohibiting excessive emissions of smoke, noise, vibrations, and odours; and from protective covenants entered into among landowners under the protection of the law. The matter of excessive emissions is dealt with in civil law jurisdictions under the heading of neighbourhood rights, because the restrictions are imposed in favour of adjoining owners. In common law jurisdictions, the same matter forms the object of the law of nuisance, a topic of tort law. Protective covenants (covenants that are attached to a property deed and go with

Protective covenants

the property when sold, usually containing restrictions on maintenance or sale to specific parties) are rare in civil law jurisdictions, because the field is adequately covered by directly applicable legislation; moreover, the scarcity of protective covenants may be attributed to the fact that restrictions on the use of lands by juridical act necessarily must take the form of a servitude, namely, of a charge laid on an estate in favour of another estate. In common law jurisdictions, however, protective covenants are quite common in subdivision developments. They are ordinarily established by a subdivider or by a group of landowners for the purpose of preserving and enhancing real property values by the maintenance of certain building standards and uniformity in the use of lands. Protective covenants also were once used in the United States to perpetuate racial discrimination in housing, but the United States Supreme Court ordered an end to this practice on constitutional grounds.

SOCIAL CONSEQUENCES OF PROPERTY RIGHTS

The issue of the social consequences of property rights, inextricably connected with the questions of the nature, origin, and justification of private property, has given rise to interminable discussions throughout the ages. A deep analysis of this issue properly belongs to the domains of the philosophy of law, sociology, economics, and political theory. At this point, it suffices to discuss a few social effects of property rights.

Political effects. The political organization of society often determines the kinds, scope, and content of property rights that may be enjoyed by individuals; in turn, it is inevitable that the distribution of wealth should exert influence on the structure and the evolution of the political organization. The fact that property rights exercise political power has not escaped the attention of philosophers since the time of Plato and Aristotle. Plato attacked private property because he thought that, without its abolition, his philosopher kings, the future guardians of his ideal state, would become enemies and tyrants instead of allies of the people. Aristotle criticized his contemporaries' proposals to equalize individual fortunes in order to prevent revolution, but he fully concurred in the view that all civic dissensions arise from inequality of wealth. Consequently, he suggested that all gifts of property should be outlawed and that severe restrictions should be imposed upon the right of inheritance, the means by which an oligarchy preserves itself. The idea that private property was at the root of political and economic evils was carried into modem times by several theorists and philosophers, including, of course, Karl Mam.

Economic effects. The effect of property rights on the economic system is as profound as that on the political organization of society, and the two are frequently inseparable and indistinguishable. Certainly, whether a political and economic system qualifies as capitalistic or socialistic depends on the kinds, scope, and extent of property rights available to individuals. And both capitalistic and Socialist countries control economic activities in terms of their philosophies regarding property rights. Antitrust legislation enacted in the United States, for example, is intended to encourage competition and to prevent the concentration of wealth in the hands of a few powerful individuals or corporations. In the Soviet Union, however, legislation consistent with Marxist philosophy narrowly delimits individual property rights and secures the concentration of wealth in the hands of the state; as a result, the influence of private property on social and economic organization is minimized.

Effects on social stability. The effect of property rights on social stability and, conversely, the effect of internal security on economic development have been discussed extensively by economic theorists, political scientists, and social philosophers. Classical economists, for instance, believed that social stability was possible only if individuals might control, for purposes beneficial to themselves, what they have discovered and appropriated to their use, what they have created by their own labour, and what they have acquired under the existing social and economic order. Private property was deemed to be the cement of society. Socialists, of course, have often had arguments precisely to the contrary.

Effects on initiative and production. Economists and political thinkers have for many years debated the effect of property rights on individual initiative and productivity. Classical economists, for instance, believed that the right of property was the most powerful of all encouragements to the increase of wealth. This view has been consistently challenged by theoretical socialists who have postulated the existence of altruistic motives in the production of goods. Pragmatic considerations, however, seem to have led some Socialist countries, including the Soviet Union, to recognize that, at least to some degree, profit taking does provide incentives. Individual motives for the acquisition of personal property thus help to make the wheels of industry go round in Socialist as well as in capitalist countries.

BIBLIOGRAPHY. For doctrinal works on property in general, and on the foundation, incidents, effects, and function of property rights in particular, see R.T. ELY, *Property and Contract in Their Relations to the Distribution of Wealth,* 2 vol. (1914, reprinted 1971); G. HUSSERL, *Der Rechtsgegenstand* (1933); C.R. NOYES, *The Institution of Property* (1936); L.F. VINDING KRUSE, *The Right of Property* (Eng. trans. 1939); R. GONNARD, *La Propriété dans la doctrine et dans l'histoire* (1943); K. RENNER, *The Institutions of Private Law and Their Social Functions* (Eng. trans. 1949, reprinted 1976); F. GRACE, *The Concept of Property in Modern Christian Thought* (1953); J.R. COMMONS, *Legal Foundations of Capitalism* (1924, reprinted 1974); R.P. CALLIESS, *Eigentum als Institution* (1962); G. DIETZE, *In Defense of Property* (1963); and A.A. BERLE and G.C. MEANS, *The Modern Corporation and Private Property,* rev. ed. (1968).

On the historical development of property rights, the classic works are: L. FELIX, *Entwicklungsgeschichte des Eigentums,* 4 vol. (1883–1903); E.L.V. DE LAVELEYE, *De la propriété et de ses formes primitives,* 4th rev. ed. (1891); C.J.M. LETOURNEAU, *Property: Its Origin and Development* (Eng. trans. 1901); P. VINOGRADOFF, *Outlines of Historical Jurisprudence,* 2 vol. (1920–22, reprinted 1971); and N.D. FUSTEL DE COULANGES, *The Origin of Property in Land,* 2nd ed. (Eng. trans. 1892, reprinted 1927). Modern works include: F. CHALLAYE, *Histoire de la propriété,* 5th ed. (1958); R.B. SCHLATTER, *Private Property: The History of an Idea* (1951, reprinted 1973); and w. NIPPOLD, *Die Anfänge des Eigentums bei den Naturvölkern und die Entstehung des Privateigentums* (1954).

On the contemporary law of property in the United States, see A.J. CASNER (ed.), *American Law of Property, 8* vol. (1952–) ; R.R.B. POWELL, *Law of Real Property,* 7 vol. (1949–61); J.E. CRIBBET, *Principles of the Law of Property,* 2nd ed. (1975); R.A. BROWN, *The Law of Personal Property,* 3rd ed. (1975); in England, see G.C. CHESHIRE, *The Modern Law of Real Property,* 12th ed. (1976); D.C. JACKSON, *Principles of Property Law* (1967); J.C. VAINES, *Personal Property,* 5th ed. (1973); and F.H. LAWSON, *Introduction to the Law of Property* (1958); in selected civil law jurisdictions, including France, Germany, Greece, and Louisiana, see A.N. YIANNOPOULOS, *Civil Law of Property* (1966); C.M.B.A. AUBRY and C.F. RAU, *Property* (Eng. trans. 1966); and M.F. PLANIOL and G. RIPERT, *Les biens* (1952); in the U.S.S.R., see R.O. KHALFINA, *Personal Property in the U.S.S.R.,* (Eng. trans. 1966); and N.D. KOLESOV, *Social Property in the Soviet Union* (Eng. trans. 1961); in Arab countries, see F.J. ZIADEH, *Property Law in the Arab World* (1979). For a comparative study of the institution of ownership in common law and civil law jurisdictions, see K. RUDOLPH, *Die Bindungen des Eigentums* (1960).

On the law of trusts, see A.W. SCOTT, *The Law of Trusts,* 3rd ed., 6 vol. (1967, supp. 1980); G.W. KEETON and L.A. SHERIDAN, *The Law of Trusts,* 10th ed. (1974); and c. DeWOLF, *The Trust and Corresponding Institutions in the Civil Law* (1965). On the civilian notion of exclusive ownership, and on the law governing usufruct, see A.N. YIANNOPOULOS, *Personal Servitudes,* 2nd ed. (1978). On the law governing restraints on the use of property in common law jurisdictions, see c. EDWARD CLARK, *Real Covenants and Other Interests Which Run with Land,* 2nd ed. (1947); and G.H. NEWSOM, *The Discharge and Modification of Restrictive Covenants* (1957).

(A.N.Y.)

Property Tax

A property tax is a tax levied upon land and buildings. In some countries, including the United States, the tax is also levied upon business and farm equipment and inventories. Sometimes the tax extends to automobiles, jewelry, furniture, and even to such intangibles as bonds, mortgages, and shares of stock that represent claims on, or ownership of, tangible wealth.

National differences in taxable property

In most countries, property taxes are used by local or state rather than national governments. Property-tax receipts in 1970 supplied about 55 percent of the revenue raised by local governments in the United States. Throughout much of Europe and Latin America and parts of Africa and Asia, one finds taxes that may be broadly classified as property taxes in their functioning and that supply significant proportions of total tax revenue.

In several countries the property tax applies in fact primarily to urban real property. The intensity of use varies widely—revenues total about 4 percent of national income in Canada, Ireland, the United Kingdom, and the United States; about 2 percent in Australia, Denmark, and New Zealand; somewhat more than 1 percent in Belgium, Japan, South Africa, and Taiwan; and lesser but still significant amounts in a dozen or more other countries.

In some countries, property-tax revenues have lagged far behind the growth of national income because the tax has been based on measures that have not responded to changes in the general level of prices. The original land surveys were designed to serve for long periods, and the taxes were based on surface area or presumed income at rates that might have served moderately well in a world of stable economies. War, inflation, and other forces, however, have made them obsolete; and popular resistance and lack of administrative capacity have generally prevented their modernization.

Levies not ordinarily included in property taxes are those on transfer of property (by sale, gift, or death), on net wealth, and on capital; special charges for some public service or improvement (such as special assessments in the United States); certain types of agricultural imposts; and portions of income taxes that apply to presumed or actual yield of farm or urban land.

THE DEVELOPMENT OF PROPERTY TAXATION

The tax base

One of the most difficult problems in taxing property is to find a reasonable basis of assessment. The problem has grown more difficult as the complexities of economic life have increased. The taxes of the ancient world, of parts of medieval Europe, and of the American colonies were originally land taxes based on area rather than on value. Eventually gross output came to serve as the base. At a later stage, attempts were made to find a measure of what would now be called the individual "ability to pay"; thus, other forms of wealth, such as farmhouses, animals, and implements, were included. At various times, governments have tried to make the tax base one of general property value rather than of specific amounts of different types of particular properties. Yet to reach movable property effectively for taxation has always been difficult; and taxing intangible forms of wealth has proved even harder.

The New England colonies developed taxes that sought to reach all of the "visible estate," real and personal. The "general property tax," applying to all property, was on the statute books of some states by 1800. During the colonial period, the southern and middle colonies made relatively little use of property taxation, but by the middle of the 19th century it had become the principal source of revenue in all the states. The base of the general property tax was defined to include intangible wealth. Yet the value of mortgages and other intangibles consisted largely of claims to rights in real estate and tangible personal property, which were also taxed. Since the double burden seemed excessive, and since concealment was easy, enforcement of the tax on intangibles became difficult and ultimately almost impossible. Disintegration of the property tax as a general tax began early and continued into the 20th century as more and more property escaped, legally and nonlegally. Today the tax in most U.S. states applies to real property; to machinery and inventory of businesses; to relatively small amounts of furniture and jewelry and somewhat more to the value of autos; and in a small degree to bank balances, securities, and other intangible personal property. A national census in 1966 found that real estate accounted for 85 percent of the property-tax base.

The property tax is ultimately a tax upon persons; property serves only as the basis of assessment. The amount payable is based not on a person's or a company's total net wealth but on gross value without regard to debts. The tax will ordinarily be paid from income (either from the property or from other sources).

The property tax in the United States is the chief source of revenue for local government. State governments once used the tax as an important source of revenue, but only a dozen now get as much as 2 percent of their revenue from this source. Forty state governments, however, assess some or all of the operating property of railroads and other utilities. For many years, few if any states took serious interest in the way in which local governments administered the tax; but active efforts to improve it expanded after World War II, and by 1969 significant leadership was found in half or more of the states. Some authorities favour a state takeover of the tax, partly because they believe that states would administer it better and partly in order to remove inequalities in taxing capacity among local governments (especially for financing schools).

The scope of the tax in different countries varies greatly, depending upon legal factors, administrative realities, tradition, availability of other sources of revenue, the organization of government—especially the relative role of local government, for which this levy is of key significance—and the public services provided. The attempt to extend the tax to other than real property (land and buildings) is almost unique to the United States. There is a strong argument in principle for broad coverage of tangible property, since otherwise the tax discriminates (is not neutral) among types of consumption and investment. Administrative difficulties limit what is possible in practice. Classification of property by different types has served as a basis for varying the effective burdens, sometimes by providing for the exclusion of a fraction of the value of some kinds of property (machinery, forests, mines, securities, furniture, etc.), sometimes by adjusting the rates of tax.

ADMINISTRATION

Problems of tax administration

Responsibility for the various phases of administration rests almost entirely upon government officials. Administration involves the discovery or identification of the property to be taxed, its valuation, the application of the appropriate tax rate, and collection. Where the amount of tax is measured by income, as in Great Britain and some of the British Commonwealth countries, income rather than capital value must be determined. The methods of self-assessment or withholding by a third party (an employer or seller) that are used in other forms of taxation are seldom applied in this area. Important aspects, especially valuation, are a matter of judgment rather than of fact. The determination of value for tax purposes is not an incidental result, or an automatic by-product, of a transaction entered into for other purposes, such as a wage payment or a retail sale.

Difficult administrative problems arise in determining (1) what actually exists in a physical sense (the location, topography, and area of a piece of land; the size, materials, and condition of buildings; the number and types of machines or items of inventory) and (2) the value of the property. To do this well requires skilled professional personnel, access to information of various types, and appropriate facilities. The quality of most property-tax administration is far below satisfactory levels. Valuation procedures in Australia, New Zealand, and Great Britain appear superior to those often accepted in the United States, although some states and communities in the United States have shown considerable improvement.

Better administration involves a number of things. One is better mapping and the improvement of other means of getting accurate property descriptions. Another is more sources of data about values and more sophisticated approaches to valuation. For some types of properties, such as single-family residences, sales of generally similar properties provide a good basis for valuation. Some properties, such as office and apartment buildings, can be valued on the basis of the income they yield. For unique and highly specialized properties, including factory and other buildings that are integral parts of a business operation, the value for tax purposes must rest on estimates of reproduction cost less depreciation. Business inventories may be valued on the basis of company records and so may machinery and equipment.

Assessing the tax

Good assessment requires the skills of a permanent professional staff, selected and promoted on a merit basis, working full time at pay comparable to that in private industry and free from political pressures. The United States assessor has typically been a part-time official, usually elected, poorly paid, and without the special training now recognized as essential. Rarely has he had the basic information and other facilities needed. Incompetence has sometimes been compounded by favouritism and corruption. The area under the responsibility of one assessment staff needs to be large enough to permit specialization and the development of expertise. Rarely are staffs large enough to make reasonably complete coverage oftener than every four or five years. Yet the pace of change and the amount of new constiuction are so great as to make many assessments significantly obsolete before a new cycle can correct them. Keeping maps and records up-to-date calls for more continuing work than most governments will support, though modern data-processing techniques offer hope of reducing the burden.

Practice in most countries has been to assess at only a small fraction of the full, current market value even when the governing law specifies that assessment shall be at 100 percent. Low valuations mean that tax rates must be higher. This tends to make inequities in the tax burden greater because a large rate difference magnifies any inequity growing out of poor valuation. As a practical matter, it is more difficult to compare the assessments of different properties when all are far from market levels; and it is more difficult to criticize an assessment as being unfairly high when it is, say, one-fourth of market price while others are one-fifth or less.

Tax review

Because the tax base, and hence the amount of tax payable, depends upon an official's estimate of value rather than on a free-market test (as with a sales tax) or on the taxpayer's report (as with an income tax), the taxpayer will not have participated in the determination of the assessment. The law usually provides facilities for appeal before the tax becomes final, but these are often of little worth. Such review often is farcical. The taxpayer may be ignorant of procedure, or he may not consider the possible saving worth the trouble. Occasionally an owner comes armed with legal talent, piles of evidence that the board cannot check, and perhaps political influence. The elements of unchecked discretion, in both the original assessment and in its later review, provide opportunities for abuse. A owner who hires the right attorney may get a reduction easily, especially if the assessor has deliberately overestimated the value.

TAX RATES

Generally in the United States, but rarely in other countries, the tax rate consists of two or more superimposed elements accruing to separate jurisdictions—to city or town, school district, county, perhaps the state, and to one or more special districts (sewer, drainage, etc.). Each unit of government determines the tax rate it will apply, ordinarily expressed as a percentage of the assessed value. Since assessments are usually much below market values, nominal tax rates give a misleading impression of the burden. For example, 20 U.S. cities in 1966 had nominal rates averaging 9.33 percent, but assessments averaged 15 to 20 percent of market values; consequently, the true rates were 1.6 to 2.1 percent. In eight cities with average tax rates of 5.23 percent, the property was generally assessed at 50 to 60 percent of market prices; effective rates, then, ranged from 2.6 to 3.1 percent of full value.

When government functions were narrow and the property tax the sole source of local revenue, tax rates were determined simply by dividing the figure for estimated expenditure by that for assessed valuation. If spending was to be $400,000 and assessments were $40,000,000, a rate of 1 percent would suffice. Political attitudes have changed. 'Today officials are more likely to estimate the amount that will be available if the existing tax rate is maintained and then try to judge whether taxpayers will consider additional spending to be worth a higher rate of tax. U.S. state governments formerly used the property tax as a flexible element, relying for the most part on other taxes. According to whether these were inadequate or in surplus, the state would raise or lower its property-tax rate. Many states still have constitutional power to do so. When a strong demand for some particular service appears but officials prefer not to raise their "general fund" rates, the legislature may vote to mandate a "special" rate.

Rate limitations are common, imposed sometimes by the state constitution, often by statute. For each class of government—counties, cities, school districts—a maximum ceiling rate will be set. Sometimes the limit may be changed upon referendum or by special legislative action. It is difficult to judge whether such limitations have been effective in restraining the growth of spending. One result, however, has been the establishment of special districts that have independent taxing power and thus lie outside the limitations.

THEORY OF PROPERTY TAXATION

Shifting and incidence

Most taxes can be shifted in some degree from the taxpayer to other persons. As a rule, the higher the tax on land, the lower the price. But if land is in great demand, the buyer may absorb some or all of the tax. The actual amount a buyer will pay for a plot of land depends upon the net income he can expect from it in relation to the yields available from other investments. If the net income from a plot of land is expected to be $1,200 a year indefinitely, and if the prevailing yield on long-term assets is 6 percent, then the land will be worth $20,000. If the tax then rises by $200, the net yield drops to $1,000 and the worth of the land falls to $16,667. The tax increase is said to have been capitalized. To the buyer of the land, the tax in effect at the time of purchase will not be a burden thereafter because he has already discounted it in the purchase price. Since land prices generally have gone up over time, the property tax has not so much lowered land prices as retarded their rise. In the United States, in the ten years up to 1966, land prices are estimated to have risen on the average by 6.9 percent a year.

The extent to which taxes on buildings and other improvements can be shifted involves quite different factors. The construction of buildings depends upon the willingness of investors to make capital available for them, and taxes affect that willingness. A property tax will be treated as a cost of doing business. It must generally be recovered in higher prices from consumers (or in lower prices paid to suppliers or workers). Firms that do not succeed in passing the tax on to customers will suffer a lower rate of return on invested capital. Companies in competition with others located where rates are lower may be unable to shift the tax fully to consumers. New investment will go where net earnings after tax are greatest. The supply of capital, and thus of goods and services, grows when investment yields are promising; it lags when profits are low. As output and prices adjust to changes in tax rates, the taxes will tend to be shifted to consumers. The length of time it takes for a change in a property tax on buildings to be reflected in prices paid by consumers varies from a few months to a number of years. For regulated public utilities, the shifting of a change in tax will usually require some time because new rates will have to be authorized by an official agency. The regulatory process typically requires many months. In some cases, notably United States railroads, after-tax returns have been markedly lower than those in the economy generally; the taxes have been heavier than on business generally, and they have fallen largely on owners (and on suppliers of debt capital) rather than on customers.

The homeowner cannot shift the tax on his dwelling. The price he paid for the land, of course, will have been adjusted to the tax that was in effect when he purchased the property; this tax is no burden on the owner after his purchase (though few homeowners realize that if the tax had been lower, the price paid for the land would have been higher). The tax on the house closely resembles a tax on other items of consumption, although it tends in the United States to be higher than the burdens on most other consumer goods: Deduction of the tax in computing income for the individual income tax helps reduce the net burden for the homeowner.

Burden of the property tax

The amounts of property tax borne by persons at different levels of income cannot be determined accurately. There is almost no way to take account adequately of the element represented by capitalized land tax in the price of land. The portion of property tax falling on businesses (often more than **40** percent) is presumably shifted to consumers according to their purchases, including those of telephone, electric, and other utility services.

In general, the property tax seems to be either roughly proportional to income or slightly regressive (in that the increases in amounts of tax paid do not rise as rapidly, in percentage terms, as income). Regressive taxes are widely believed to be inequitable. While property taxes burden persons with low incomes more than may seem wise, their ultimate benefits have a "pro-low-income" bias. This means that their total redistributive effect from higher to lower income groups is substantial when account is taken of the degree to which property taxes pay for schools and other services for low-income groups. There is also widespread "horizontal inequity" in property taxes because of unequal assessments upon owners. The tax falls more heavily on some kinds of business (*e.g.*, railroads and other utilities) and some types of consumption (*e.g.*, housing) than on others. In the United States, property taxes on farming as a business tend generally to be low relative to the value of property but high in relation to income. Because property taxation is of such long standing that its many elements have worked themselves into the economy, some portions being capitalized and others variously adjusted to, the inequities have to some extent been reduced.

Exemptions in the United States

The property tax has been increasingly weakened by exemptions. In the United States exemptions apparently remove about **30** percent of the land area in the average locality. Most of this, however, consists of streets, schools, parks, and other property of local government; therefore, to apply the tax would merely transfer funds from one government account to another. In some localities, state- or federal-government real estate is important, although these bodies sometimes make payments in lieu of local taxes. Property owned and used for religious, educational, charitable, and some other purposes is generally exempt.

Some exemptions are made in order to attract new businesses or to encourage low-income housing. Some states grant exemptions for part of the value of a "homestead," perhaps with a limitation based on income of the owner-occupant; several give some exemption to persons over age *65* or to veterans.

Economic effects. Property taxation finances local government, not fully but enough to make the independence of local government meaningful. This permits decentralization of government, which may be considered a good because it enables a community to exercise a degree of choice.

The property tax may have substantial nonrevenue effects. Where it is heavy enough to bring large revenues, it leads to changes in behaviour, not just because taxpayers have less to spend and save but also because individuals and businesses conduct their affairs differently because of it. Although property-tax rates expressed as percentages are usually small, in the United States they apply to capital values and are effectiveiy much higher: if a property that yields 9 percent gross is taxed at **3** percent, the tax is equal to **33** percent of the pretax income—and *50* percent of the 6 percent remaining after tax. A tax of 20 cents for each 80 cents paid for the costs of housing—not as high as actually prevails in many urban areas—is *25* percent when expressed on the same basis as a retail sales tax.

A community with high tax rates on buildings will be at a disadvantage in the national (and international) competition for capital unless it can offer compensatory advantages. The supply of capital for the economy as a whole comes from saving. Whether the property tax affects it materially is not clear. Methods of production requiring relatively large amounts of capital will be discouraged if they are subject to tax.

The effect on new building

The tax on buildings and property other than land distorts resource allocation where older property exists. New, high-quality buildings are taxed more heavily per unit of space than are old ones, including slums. There is no justification for this in the costs that the two types of property and their occupants impose on local government in terms of police, fire protection, etc. Thus the user's payment for the services of local government goes down, relatively, as the building he occupies gets worse, even though public expenses attributable to the property are unchanged or may even increase. Likewise, residents who shift from poorer to better quality housing or business property must pay more toward the costs of government even though they will not ordinarily receive more government services. Cities that urgently need to replace obsolete buildings paradoxically base much of their financing upon a tax that encourages owners to hold on to deteriorated structures and penalizes owners of new ones.

Every increase in the property-tax rates on structures (not land) reduces the desirability of putting capital funds into new buildings, creates an incentive against upgrading quality by new construction, and discourages maintenance. It also leads to the construction of rooms, apartments, and buildings somewhat smaller than would be the case in the absence of tax.

Differences in effective tax rates among localities may have the effect of creating islands of relatively low tax rates. Some communities may have tax bases above average in relation to governmental obligations and can get by with lower tax rates. They attract capital. Some communities, perhaps by the use of zoning, exclude types of property associated with high governmental expense such as high-density housing, which brings many children and requires more schools. Tax rates elsewhere must then be higher. The existence of such enclaves will add to the fiscal imbalance of neighbouring localities and accentuate the difficulties of older areas.

Lower tax rates on the fringes of an urban area encourage suburbanization. Property nearer the centre will be subject to high tax rates, aggravating the troubles of central-city business properties. High taxes on structures also favour horizontal over vertical growth of metropolitan areas.

Where, as in Great Britain, the property tax rests on income, land held idle or far below its best use will yield little revenue. In such cases, the tax incentive for efficient use is notably lacking. The rates at which timber is cut and minerals extracted can be influenced materially by property taxation.

Taxing the "unearned increment"

Site-value taxation. The use of a land tax as the chief source of revenue has often been proposed. It was favoured by the Physiocrats in 18th-century France. Probably the best known exponent was a 19th-century American, Henry George. His Progress and Poverty (1879) drew upon economic analysis in the tradition of Ricardo and Mill to argue persuasively for a single tax on land and the abolition of other taxes (then predominantly levied on other property). More recently, proposals for heavier taxation of land—site-value taxation—have found increasing support. One argument is that much of what is paid for the use of land reflects socially created demand and is not a payment to bring land into existence. The community can capture in land taxes some of the values it has createci—including ihose resulting from streets, schools, and other facilities. This, it is maintained, would be a more equitable way of financing local government. Another argument is that the revenue from a tax on land would permit *a* reduction of taxes on buildings, which tend to deter new construction. A third argument is that higher land taxes would make for more efficient use of land.

There is a great deal to be said in favour of increasing taxes on land and thus lowering land prices. Economically, of course, a "high" price for some land is essential in order to encourage the best employment of it. The user of land ought to pay the amount of its worth in its best use; but the owner, facing no cost of production, need not receive all that is paid. Government can reasonably take much of the total paid by the user.

A heavier land tax would change the conditions of ownership. The total collected from users would not change, but private owners of land would retain less, the public treasury getting more. The price system would still allocate land use. Taxes on improvements could then be reduced greatly. The tax relief for deteriorated buildings would be slight, but for those of high quality the reduction could be large in relation to net return on investment. More buildings, new and better ones, would be supplied. Modernization and maintenance of existing buildings would become more profitable.

Over the longer run, landowners would get less of the increments in land values and the public would get more. Socially created values would be channelled into governmental rather than private uses.

The use of elements other than value as the tax base would offer the possibility of more rational and efficient taxation. Taxes could be related more closely to the cost of governmental services. If land area, the distance from some base point, and perhaps such factors as the floor space or volume of a building were used as criteria—and perhaps related to the specific services of local government—considerable revenue could be collected with much less administrative expense.

The opponents of site-value taxation point out that the unearned increment in land value has been capitalized and question the fairness of imposing a heavy tax on present land values for which owners have paid in good faith. They doubt the ability of assessors to make fair enough appraisals to support much heavier rates on land. They also doubt whether land alone, excluding buildings, would be an adequate tax base.

Moves in the direction of site-value taxation have been made in Australia and New Zealand, South Africa, and parts of western Canada. The case for site-value taxation being a strong one, other areas will probably adopt it. (C.L.Ha.)

PROPERTY TAXATION IN EUROPEAN COUNTRIES

The most common forms of property taxation in Europe are general net-wealth (net-worth) taxes and taxes on specific types of property, such as real estate. Some European countries, however, leave property itself untaxed—although the extent of an individual's property holdings may be used as an index of taxable capacity for personal and corporation income-tax purposes.

The property-tax figures chiefly in the revenues of local governments. For central and federal governments it is not an important source of revenue. In 1969, for example, taxes on net worth and on real estate supplied only 4.5 percent of the revenues of the federal government of West Germany and 4 percent of the revenues of the national government of France. On the other hand, these taxes supplied from 20 to 30 percent of local-tax revenues in Belgium, Denmark, France, and Germany.

Methods of assessment

The three principal approaches to the assessment of property are rental value, capital value, and market value. Swiss cantons assess a property tax on agricultural land based on the income from the land, which is treated as rent. Denmark has to a certain extent adopted the rental-value concept for assessing agricultural land subject to the estate tax and seagoing vessels subject to the net-wealth tax. A modified rental-value concept is used in Spain to assess urban real estate by cadastral income (contribucidn territorial *urbana*), which is assumed to reflect expected rents.

A more common approach to the assessment of real property in European countries is that of capital value. The traditional idea is that capital value can be estimated on the basis of rental values, treating them as earnings on capital. The capital-value concept (Einheitswert) prevails in real-property and net-wealth taxation in West Germany and Luxembourg, in the net-wealth taxation of Austria, and in the taxation of real estate by the net-worth tax of Sweden (*förmögenhetsskatt*).

Most European countries endeavour to assess property according to its market value. Applications of the principle differ. In Swiss cantons, for example, market value is used in net-worth and real-estate taxation of land and buildings, while The Netherlands makes use of market value in net-wealth taxation only for agricultural and other open land. Still another way has been adopted in Austria, where a legally defined "fair-market value" (*gemeiner* Wert) is applied in an effort to come close to the actual selling value of real property.

In Europe as elsewhere, the chief problem in the administration of property taxes is to keep assessed valuations up-to-date. This is especially true for real estate, in which the principal assessments were made decades ago (in The Netherlands at the end of the 19th century, in Germany in 1935, in Spain in 1929, in France in 1925 and 1939). Although new and supplementary assessments have been made since, the relation between the assessed values has remained unchanged and does not reflect current realities.

Efforts to reform the system of property taxation have repercussions upon the rest of the fiscal system and touch the interests of all economic and social groups. Future changes will be concerned with three sets of problems: (1) the need to adapt tax bases to changing property values; (2) the requirement that different kinds of property be treated equally, so as not to discriminate among them; and (3) the need to harmonize taxes among the countries of the Common Market. In the early 1970s various proposals were being considered in European legislative bodies, but there was a lack of unanimity among groups whose interests were directly involved. (H.Fec.)

PROPERTY TAXATION IN ASIAN COUNTRIES

Property taxation has long been a feature of the tax systems of some Asian countries. Its significance as a revenue source has been growing in recent decades.

The tax is imposed mostly at the local government level, although in a few countries, such as Indonesia, Laos, and Singapore, it is levied by the central government. In the Philippines the property tax is imposed by the national government but is collected by, and the entire proceeds distributed among, local units.

Generally, property taxes in Asian countries are paid in cash. In the Republic of China on Taiwan, however, the niral land tax on rice land is payable in kind.

Importance of property taxation in Asia

The ratio of property taxes collected to total local revenues is indicative of the importance of this tax. India, which first imposed the property tax in 1793, drew an average of 57 percent of local revenues from it in the years 1966–68; it is the principal source of revenue of about 2,000 autonomous local bodies. Collections in Japan averaged 15.6 percent of the total revenue of local governments from 1968 to 1970. In the Philippines the property tax brought in about 15 percent of local-government revenues from 1967 to 1969. In Pakistan the ratio averaged 14.3 percent in 1966–69; in South Korea, 9.2 percent in 1967–69; in Thailand in 1968, only 5 percent. In countries where the tax is levied nationally, the ratio of collections to total national revenue is insignificant.

While the property tax serves as a major source of income for local government, other revenue sources are more productive. In Thailand, for example, local governments draw twice as much revenue from the business tax. In Japan they derive more revenue from the municipal inhabitant tax. In Pakistan the yield from the property tax was exceeded by that from the sales tax and income tax in 1966–69. In South Korea the acquisition tax, entertainment and restaurant tax, and automobile tax were more productive in 1967–69.

In Asian countries the property tax applies to land, buildings, and other improvements. In the Philippines, machinery is also included. In general, personal property, including intangibles, is not subject to the tax. There are exceptions to this, as in Indonesia, where net wealth exceeding 2,000,000 rupiahs is taxed, and in Japan, where certain types of personal assets are taxable.

Bases of assessment

The bases used by Asian countries in the assessment of the property tax include market value (in Japan and the Philippines), annual rental value (in India, Singapore, South Vietnam, Malaysia, South Korea, and Laos), and capital value (in Thailand and some states in India).

As a rule, the bases of assessment in Japan are "normal" market value for land, replacement cost less depreciation in the case of buildings, and purchase price less normal depreciation for depreciable assets. In the Philippines the assessment law stipulates fair market value as the basis of assessment, but in practice the basis is far below the market value because assessments have not been carried out periodically.

Most Asian countries use annual rental value as the basis of assessment. Under this principle, the tax is based on the average gross-rental income the property is expected to generate under prevailing market conditions.

In Indonesia the basis of the property tax is the net wealth owned by the taxpayer at the beginning of the calendar year. Net wealth is defined as the value of all property less debt.

Other Asian countries collect a fixed amount based on a particular unit of land measurement. Laos collects a specific amount per square metre of land. In West Malaysia the annual tax on land is a certain amount per 1,000 square feet.

Problems of administration. A number of difficulties have been encountered in the administration of property taxes. In general, these relate to the valuation of property, collection of taxes, and the application of exemptions.

Valuation and collection

In Japan it was noted in a 1959 survey that the methods of land valuation used by different municipalities resulted in an unequal tax burden among landowners. Thus, the assessed value of land averaged 16 percent of the market price for urban land and 25 percent for rural land. To correct this, a new method of valuation was introduced in 1964, but it led to 600–700 percent increase in the tax on urban land. New adjustments had to be made, scaling down the urban land tax to a more equitable level.

Real property in the Philippines is assessed at fair market value, but only about 45 percent of the value is taxed. In many localities, the existing assessment schedules are based on values of a number of years ago. Moreover, many taxable properties have not been included in the assessment rolls because of the absence of tax maps and other basic assessment tools. The need for these tools has long been felt, but funds for their purchase and for the employment of more assessment personnel have been lacking. Another problem in the Philippines is under-collection, resulting from inadequate collection techniques, the low pay levels of collecting officials, and a failure to apply sanctions for noncompliance.

A study of the property tax in South Korea in 1966 disclosed that the assessment base was very low in relation to market value. The actual property-tax rate had continuously decreased until it was lower than its legal rate. Buildings were found to be more heavily assessed than land. The report recommended fundamental reforms in assessment procedures that would lead to the adoption of a system based on market value.

In India the main problem of property taxation is basically that of assessment. Judicial decisions have required the standard rent fixed under the Rent Control Act to be considered the reasonable rent for the purpose of valuation of the property. As a result, the annual rental value of almost all urban property has remained frozen. There is also a lack of proper machinery for correcting and updating the assessment lists in most local bodies, which have also been backward in collecting the tax. There is also a lack of competent personnel. Various study committees set up in India from 1951 to 1966 have recommended: (1) imposition of a minimum statutory rate to be uniformly applied; (2) abolition of the method of assessment based on the Rent Control Act; (3) establishment of an impartial and independent machinery for assessment; and (4) the vesting of independent agencies with powers to recover arrears.

Prominent among the problems besetting the administration of the property tax in many Asian countries is the lacklustre performance of assessors and collectors. Many of the personnel are untrained, and many assessment and collecting units are undermanned. Sometimes considerations other than merit play a decisive role in the appointment of assessors and collectors.

Exemptions in Asian countries

Another problem area of property taxation is that of exemptions. Asian countries allow exemption for a wide range of properties and assets. These can be generally classified as follows: (1) those belonging to the government—state, central, or local—and used for administrative purposes; *(2)* those used for religious, charitable, scientific, and educational purposes; (3) those with values below a certain limit; (4) cemeteries or burial grounds; and (5) property used by international organizations, embassies, and consulates.

The first category is common to almost all Asian countries. Belonging to the second category are pagodas in Laos and places used for sacrificial rites in South Korea. In Japan exemption from the tax is accorded to assets owned by organizations operated for public benefit; *e.g.*, the Japan Red Cross Society and the National Health Insurance Association. Lands held by homesteaders in the Philippines, and reclaimed lands in South Korea, are also exempt.

Preferential tax treatment is also used to encourage building construction and land development. In Singapore, property-tax rates are lower on buildings and vacant land in the less developed areas. In Japan, partial exemptions are given to new fireproof residential buildings having apartments of less than a certain size.

The future of property taxation. While in Europe and the United States the property tax has long been a stable and productive source of revenue, in Asia its full revenue potential is still to be tapped. There are serious deficiencies in the legislation and administration of property taxes. One barrier to reform is a dearth of statistical information that would enable a systematic assessment of the system's effectiveness. Any thoroughgoing reform will also require resolute steps to rid the tax-collection machinery of the inept and the corrupt. Specific changes are needed in the following areas: rate structure, classification of property, exemptions, administration, and technical expertise. (A.Q.Y.)

BIBLIOGRAPHY. A general introduction to the subject of property taxation in the United States and Europe is DICK NETZER, *Economics of the Property Tax* (1966). Descriptions of property taxation in various countries may be found in the "World Tax Series" published in recent years by the Harvard Law School. A summary of current economic theory is contained in C. LOWELL HARRISS, *Property Taxation: Economic Aspects* (1968). Works on specific countries include: GEORGE C.S. BENSON *et al., The American Property Tax: Its History, Administration, and Economic Impact* (1965); FREDERICK L. BIRD, *The General Property Tax: Findings of the 1957 Census of Governments* (1960), on the problems of property-tax administration in the United States; GEORGE F. BREAK and RALPH TURVEY, *Studies in Greek Taxation* (1964); HENRY J. GUMPEL and CARL BOETTCHER, *Taxation in the Federal Republic of Germany*, 2nd ed. (1968); JAMES HEILBRUN, *Real Estate Taxes and Urban Housing* (1966), an analysis of the economic aspects of alternative possible methods of taxing urban housing in the United States; RICHARD W. LINDHOLM (ed.), *Property Taxation, U.S.A.* (1968); ANNEMARIE MENNEL, *Die Steuersysteme in den EWG-Staaten, EFTA-Staaten und den USA* (1971), dealing with the countries of the European Economic Community and the European Free Trade Association, and with the United States; J.F.N. MURRAY, *Valuation and the National Economy* (1967), covering the history of assessing practices in various countries since biblical times; MARTIN NORR and PIERRE KERLAN, *Taxation in France* (1966); and H.P. WALD, *Taxation of Agricultural Land in Underdeveloped Economies* (1959). ARTHUR D. LYNN, JR. (ed.), *The Property Tax and Its Administration* (1969), contains the proceedings of a symposium sponsored by the Committee on Taxation, Resources, and Economic Development (TRED), held at the University of Wisconsin in 1967.

(C.L.Ha./H.Fec./A.Q.Y.)

Prophecy

Prophecy, a religious phenomenon generally associated with Judaism and Christianity, is found throughout the religions of the world, both ancient and modern. In its narrower sense, the term prophet (Greek *prophētēs,* "forthteller") refers to an inspired person who believes that he has been sent by his god with a message to tell. He is, in this sense, the mouthpiece of his god. In a broader sense, the word can refer to anybody who utters the will of a deity, often ascertained through visions, dreams, or the casting of lots; the will of the deity also might be spoken in a liturgical setting. The prophet, thus, is often associated with the priest, the shaman (a religious figure in primitive societies who functions as a healer, diviner, and possessor of psychic powers), the diviner (foreteller), and the mystic.

In a much broader sense, the term prophet has been used in connection with social and religio-political reformers and leaders.

GENERAL CHARACTERISTICS OF PROPHECY

Nature and significance. A primary characteristic of prophetic self-consciousness is an awareness of a call, which is regarded as the prophet's legitimization. This call is viewed as ultimately coming from a deity and by means of a dream, a vision, an audition, or through the mediation of another prophet. The Old Testament prophet Jeremiah's call was in the form of a vision, in which Yahweh (the God of Israel) told him that he had already been chosen to be a prophet before he was born (Jer. 1:5). When the call of the deity is mediated through a prophet who is the master of a prophetic group or an individual follower, such a call can be seen as a mandate. Furthermore, such mediation means that the spirit of the prophet master has been transferred simultaneously to the disciple. In the case of cult prophets, such as the prophets of the gods Baal and Yahweh in ancient Canaan, the call may be regarded as a mandate of the cult.

Prophets were often organized into guilds in which they received their training. The guilds were led by a prophet master, and their members could be distinguished from other members of their society by their garb (such as a special mantle) or by physical marks or grooming (such as baldness, a mark on the forehead, or scars of self-laceration).

The nature of prophecy is twofold: either inspired (by visions or revelatory auditions), or acquired (by learning certain techniques). In many cases both aspects are present. The goal of learning certain prophetic techniques is to reach an ecstatic state in which revelations can be received. That state might be reached through the use of music, dancing, drums, violent bodily movement, and self-laceration. The ecstatic prophet is regarded as being filled with the divine spirit, and in this state the deity speaks through him. Ecstatic oracles, therefore, are generally delivered by the prophet in the first-person singular pronoun and are spoken in a short, rhythmic style.

Prophetic ecstasy

That prophets employing ecstatic techniques have been called madmen is accounted for by descriptions of their loss of control over themselves when they are "possessed" by the deity. Prophets in ecstatic trances often have experienced sensations of corporeal transmigration (such as the 6th-century-BC Old Testament prophet Ezekiel and the 6th–7th-century-AD founder of Islām, Muhammad). Such prophets are believed to have a predisposition for such unusual sensations.

The functions of the prophet and priest occasionally overlap, for priests sometimes fulfill a prophetic function by uttering an oracle of a deity. Such an oracle often serves as part of a liturgy, as when ministers or priests in modern Christian churches read scriptural texts that begin with the proclamation: "Thus says the Lord." The priest, in this instance, fulfills the prophetic function of the cult. Not only do the role, of the prophet and priest overlap but so do the roles of the prophet and shaman. A shaman seldom remembers the message he has delivered when possessed, whereas the prophet always remembers what has happened to him and what he "heard."

The diviner, sometimes compared with the prophet, performs the priestly art of foretelling. His art is to augur the future on the basis of hidden knowledge discerned almost anywhere, as in the constellations (astrology), the flight of birds (auspices), in the entrails of sacrificial animals (haruspicy), in hands (chiromancy), in casting lots (cleromancy), in the flames of burning sacrifices (pyromancy), and other such areas of special knowledge (see also DIVINATION; SHAMANISM).

Mystics and prophets are similar in nature in that they both claim a special intimacy with the deity. The mystic, however, strives for a union with the deity, who usurps control of his ego, whereas the prophet never loses control of his ego. On occasion mystics have delivered messages from the deity, thus acting in the role of a prophet, and have been known to use ecstatic trances to reach the divine or sacred world; *e.g.*, many Roman Catholic saints and Muslim Siifis (see also MYSTICISM; SAINT).

Uniqueness of Israelite prophecy

In the Western world, Israelite prophecy is regarded as unique, for not only did it oppose institutionalized religion but it is understood as having propagated an ethical religion emphasizing individual freedom, a religion not dependent on mechanical ritual and legalism.

The term prophecy also has been used in a strictly predictive sense, not necessarily dealing with religious themes. In this sense, *The Communist Manifesto,* by Karl Marx and Friedrich Engels, was viewed as a "prophecy" of things to come; a new approach that goes against the traditional in literature, art, politics, and other areas may —in this wider sense—be termed "prophetic."

Types of prophecy. *The divinatory prophet.* Types of prophecy can be classified on the basis of inspiration, behaviour, and office. Divinatory prophets include seers, oracle givers, soothsayers, and mantics (diviners), all of whom predict the future or tell the divine will in oracular statements by means of instruments, dreams, telepathy, clairvoyance, or visions received in the frenzied state of ecstasy. Predictions and foretellings, however, may also be the result of inspiration, or of common sense by the intelligent observation of situations and events, albeit interpreted from a religious point of view.

The cult prophet. Of broad importance to the religious community is the cult prophet, or priest-prophet. Under the mandate of the cult, the priest-prophet (who may be an ordinary priest) is part of the priestly staff of a sanctuary, and his duty is to pronounce the divine oracular word at the appropriate point in a liturgy. As such, he is an "institutional" prophet. The difference between a cult prophet and a prophet in the classical sense is that the latter has always experienced a divine call, whereas the cult prophet, pronouncing the word of the deity under cultic mandate, repeats his messages at a special moment in the ritual. Because of the timeless character of cultic activity, however, every time he prophesies, his message is regarded as new.

The role of prophecy in the founding of new religions

The missionary prophet. Missionary (or apostolic) prophets are those who maintain that the religious truth revealed to them is unique to themselves alone. Such prophets acquire a following of disciples who accept that their teachings reveal the true religion. The result of this kind of prophetic action may lead to a new religion; *e.g.*, Zoroaster, Jesus, and Muhammad. The founders of many modern religious sects also should be included in this type.

The reformative prophet. Another type of prophet is of the reformative or revolutionary kind (looking to the past and the future), closely related to the restorative or purificatory type (looking to the past as the ideal). The best examples are the Old Testament classical prophets; *e.g.*, Amos and Jeremiah. Many of these so-called literary prophets were working to reform the religion of Yahweh, attempting to free it from its Canaanite heritage and accretions. In the Arab world Muhammad is included in this category. The social sympathy found among such prophets is rooted in their religious conscience, What may have been preached as religious reform, therefore, often took on the form of social reform. This kind of prophecy is also found in India and Africa, where prophets in modern times have arisen to restore or purify the old tribal religious forms, as well as the customs and laws that had their sources in the older precolonial religious life. Many of these movements became revolutionary not only by force of logic but also by force of social and political pressure (see also MESSIAH AND MESSIANIC MOVEMENTS).

Though there may be several categories of prophecy according to scholars, no sharp line of demarcation differentiates among these different types. Any given prophet may be both predictive and missionary, ecstatic as well as reformative.

PROPHECY IN THE ANCIENT NEAR EAST AND ISRAEL

The ancient Near East. In ancient Egypt, charismatic prophecy apparently was not commonplace, if it occurred at all, though institutional prophecy was of the greatest importance because life was regarded as depending upon what the gods said. Some ancient texts contain what has sometimes been regarded as prophetic utterances, but these more often are considered to be the product of wise men who were well acquainted with Egyptian traditions and history. Among Egyptian sages, historical events were thought to follow a pattern, which could be observed and the laws of which could be discerned. Thus, times of hardship were always thought to be followed by times of prosperity, and predictions were made accordingly.

Egyptian texts

In Egyptian mantic (divinatory) texts there are prophetic sayings, but the particular concerns of these texts are more political than religious. Some are fictitious, and many are considered to have been prophesied after the event has already taken place. The papyrus text "The Protests of the Eloquent Peasant" is considered by some authorities as a prophecy, since the peasant is forced to deliver speeches, saying: "Not shall the one be silent whom thou hast forced to speak." This compulsion to speak in the name of the divine is called by some scholars the "prophetical condition."

In a Hittite text King Mursilis II (reigned 1339–06 BC) mentions the presence of prophets, but there is no information about the type of prophecy. More informative are texts from Mari (Tall al-Ḥarírí, 18th century BC) in northwest Mesopotamia, where some striking parallels to Hebrew prophecy have been discovered. The Mari prophets—believed to be inspired—spoke the word of the god Dagon just as Israelite prophets spoke the word of Yahweh.

In Mari, the two key words for prophet are *muḫḫum* (an ecstatic, a frenzied one) and *āpilum* (the one who responds). Both may be connected with the cult, but there are incidents indicating that the *muḫḫum* was not bound to the cultic setting but received his message in a direct revelation from his god. The *āpilum* usually acted within a group of fellow prophets. Many of their sayings are political in nature, but there are also oracles that deal with the king's duty to protect the poor and needy, indicating that an ethical dimension was present among the Mari prophets. The messages could also contain admonitions, threats, reproofs, accusations, and predictions of either disaster or good fortune.

The Mari texts are important in the history of prophecy because they reveal that inspired prophecy in the ancient Near East dates back 1,000 years before Amos and Hosea (8th century BC) in Israel. From Mesopotamia there is evidence of the *maḫḫu*, the frenzied one, known in Sumerian texts as the *lú-gub-ba*. Mention also is made of some prophets who spoke to Assyrian kings, and their message is sometimes introduced with the clause: "Do not fear." Omina (omens) texts containing promises or predictions are also known. In one of the *maqlu* ("oath") texts, in which an *āšipu* priest is being sent forth by his god, the deity first asks "Whom shall I send?"

The *baru* (a divinatory or astrological priest) declared the divine will through signs and omens, and thus by some is considered to have been a prophet. Though he might possibly have had visions, he was not in actuality an ecstatic. The art of divination became very elaborate in the course of time and required a long period of training.

Zoroaster, the 7th–6th-century-BC Iranian founder of the religion that bears his name, is one of the least well-known founders of a religion because of the character of the existing textual materials and because some scholars have advocated that Zoroaster is a mythical figure. He may have been, however, an ecstatic priest-singer, or *zaotar*, who used special techniques (especially intoxication) to achieve a trance. Zoroaster found the priests and cult of his day offensive, and opposed them. He preached the coming of the kingdom of the god Ahura Mazdā (Ormazd), who is claimed to have revealed to Zoroaster the sacred writings, the Avesta. In the *Yasna* (a section of the Avesta), Zoroaster refers to himself as a Saoshyans, a saviour. Messianic prophecies of the end of the world are found in Zoroastrian literature, but these are more a literary product than actual prophetic utterance (see also ZOROASTRIANISM AND PARSIISM).

Prophets were a common phenomenon in Syria-Palestine. In an Egyptian text (11th century BC), Wen-Amon (a temple official at Karnak) was sent by the pharaoh to Gebal (Byblos) to procure timber. While there, a young noble of that city was seized by his god and in frenzy gave a message to the king of Gebal that the request of Wen-Amon should be honoured. In another instance, an Aramaic inscription from Syria records that the god Ba'al-shemain told King Zakir (8th century BC) through seers and diviners that he would save the king from his enemies. These chapters reveal the close connection between sacrificial rites and divine inspiration. In the Old Testament book of Numbers, chapters 22-24, the Mesopotamian prophet Balaam (who may have been a *maḫḫu*) from Pethor, whom the Moabite king Balak had asked to curse the invading Israelites, is mentioned. In chapter 27, verse 9, of Jeremiah, another Old Testament book, it is said that prophets, diviners, and soothsayers were in the neighbouring countries of Judah: in Edom, Moab, Ammon, Tyre, and Sidon. Since so little is known about these prophets, the question of the uniqueness of Hebrew prophecy is difficult to assess (see also NEAR EASTERN RELIGIONS, ANCIENT).

Origins and development of Hebrew prophecy. The Hebrew word for prophet is *navi'*, usually considered to bc a loan word from Akkadian *nabū, nabāʾum*, "to proclaim, mention, call, summon." Also occurring in Hebrew are *ḥoze* and *ro'e*, both meaning "seer," and *neviʾa* ("prophetess").

Though the origins of Israelite prophecy have been much discussed, the textual evidence gives no information upon which to build a reconstruction. When the Israelites settled in Canaan, they became acquainted with Canaanite forms of prophecy. The structure of the prophetic and priestly function was very much the same in Israel and Canaan. Traditionally, the Israelite seer is considered to have originated in Israel's nomadic roots, and the *nnvi'* is considered to have originated in Canaan, though such judgments are virtually impossible to substantiate. In early Israelite history, the seer usually appears alone, but the *navi'* appears in the context of a prophetic circle. According to I Samuel, there was no difference between the two categories in that early time; the terms *navi'* and *roʾe* seem to be synonymous. In Amos, *ḥoze* and *navi'* are used for one and the same person. In Israel, prophets were connected with the sanctuaries. Among the Temple prophets officiating in liturgies were the Levitical guilds and singers: the "sons" of Asaph, Heman, Jeduthun, who are said to "prophesy with lyres, with harps, and with cymbals" (I Chronicles). Other prophetic guilds are also mentioned. Members of these guilds generally prophesied for money or gifts and were associated with such sanctuaries as Gibeah, Samaria, Bethel, Gilgal, Jericho, Jerusalem, and Ramah. Jeremiah mentions that the chief priest of Jerusalem was the supervisor of both priests and prophets, and that these prophets had rooms in the Temple buildings. In pre-exilic Israel (before 587/586 BC), prophetic guilds were a social group as important as the priests. Isaiah includes the *navi'* and the *qosem* ("diviner," "soothsayer") among the leaders of Israelite society. Divination in the pre-exilic period was not considered to be foreign to Israelite religion.

Prophets connected with sanctuaries

In reconstructing the history of Israelite prophecy, the prophets Samuel, Gad, Nathan, and Elijah (11th to 9th centuries BC) have been viewed as representing a transitional stage from the so-called vulgar prophetism to the literary prophetism, which some scholars believed represented a more ethical and therefore a "higher" form of prophecy. The literary prophets also have been viewed as being antagonistic toward the cultus. Modern scholars recognized, however, that such an analysis is an oversimplification of an intricate problem. It is impossible to prove that the *nevi'im* did not emphasize ethics simply because few of their utterances are recorded. What is more, none of the so-called "transitional" prophets was a reformer or was said to have inspired reforms. Samuel was not only a prophet but also a priest, seer, and ruler ("judge") who lived at a sanctuary that was the location of a prophetic guild and furthermore was the leader of that *navi'* guild. In the cases of Nathan and Gad there are no indications that they represented some new development in prophecy. Nathan's association with the priest Zadok, however, has led some scholars to suspect that Nathan was a Jebusite (an inhabitant of the Canaanite city of Jebus).

Elijah was a "prophet father" (or prophet master) and a prophet priest. Much of his prophetic career was directed against the Tyrian Baal cult, which had become popular in the northern kingdom (Israel) during the reign (mid-9th century BC) of King Ahab and his Tyrian queen, Jezebel. Elijah's struggle against this cult indicated a religio-political awareness, on his part, of the danger to Yahweh worship in Israel; namely, that Baal of Tyre might replace Yahweh as the main god of Israel.

Classical prophecy in Israel and Judah

The emergence of classical prophecy in Israel (the northern kingdom) and Judah (the southern kingdom) begins with Amos and Hosea (8th century BC). What is new in classical prophecy is its hostile attitude toward Canaanite influences in religion and culture, combined with an old nationalistic conception of Yahweh and his people. The reaction of these classical prophets against Canaanite influences in the worship of Yahweh is a means by which scholars distinguish Israel's classical prophets from other prophetic movements of their time. Essentially, the classical prophets wanted a renovation of the Yahweh cult, freeing it from all taint of worship of Baal and Asherah (Baal's female counterpart). Though not all aspects of the Baal-Asherah cult were completely eradicated, ideas and rituals from that cult were rethought, evaluated, and purified according to those prophets' concept of true Yahwism. Included in such ideas was the view that Yahweh was a jealous God who, according to the theology of the psalms, was greater than any other god. Yahweh had chosen Israel to be his own people and, therefore, did not wish to share his people with any other god. When the prophets condemned cultic phenomena, such condemnation reflected a rejection of certain kinds of cult and sacrifice—namely, those sacrifices and festivals not exclusively directed to Yahweh but rather to other gods. The prophets likewise rejected liturgies incorrectly performed. The classical prophets did not reject all cults, per se; rather, they wanted a cultus ritually correct, dedicated solely to Yahweh, and productive of ethical conduct. Another important concept, accepted by the classical prophets, was that of Yahweh's choice of Zion (Jerusalem) as his cult site. Thus, every cult site of the northern kingdom of Israel and all the sanctuaries and *bamot* ("high places") were roundly condemned, whether in Israel or Judah.

Amos, whose oracles against the northern kingdom of Israel have been misunderstood as reflecting a negative attitude toward cultus per se, simply did not consider the royal cult of the northern kingdom at Bethel to be a legitimate Yahweh cult. Rather, like the prophet Hosea after him. Amos considered the Bethel cult to be Canaanite.

Prophets of the ancient Near East generally interjected their opinions and advice into the political arena of their countries, but in this regard the classical Hebrew prophets were perhaps more advanced than other prophetic movements. They interpreted the will of God within the context of their particular interpretation of Israel's history, and on the basis of this interpretation often arrived at a word of judgment. Important to that interpretation of history was the view that Israel was an apostate people—having rejected a faith once confessed—from the very earliest times, and the view that Yahweh's acts on behalf of his chosen people had been answered by their worship of other gods. In this situation, the prophets preached doom and judgment, and even the complete destruction of Israel. The source of prophetic insight into these matters is the cultic background of liturgical judgment and salvation, wherein Yahweh judged and destroyed his enemies, and in so doing created the "ideal" future. What is totally unexpected is that the prophets would go so far as to include Israel itself as among Yahweh's enemies, thus using these ideas against their own people. Usually, however. the prophets allowed some basis for hope in that a remnant would be left. The future of this remnant (Israel) lay in the reign of an ideal king (as described in Isaiah), indicating that the prophets were not antiroyalists. Though they could and did oppose individual kings, the prophets could not make a separation between Yahweh and the reign of his chosen king or dynasty. Their messianic ideology, referring to the messiah, or anointed one, is based on old royal ideology, and the ideal king is not an eschatological figure (one who appears at the end of history). In this respect, the prophets were nationalistic; they believed that the ideal kingdom would he in the promised land, and its centre would be Jerusalem.

Concepts of the remnant and of the messiah

With the Exile of the Judaeans to Babylon of 586 BC, prophecy entered a new era. The prophecies of what is called Deutero-Isaiah (Isaiah 40–45), for instance, were aimed at preserving Yahwism in Babylonia. His vision of the future went beyond the pre-exilic concept of a remnant and extended the concept into a paradisiacal future wherein Yahweh's new creation would be a new Israel. This tone of optimism is continued in the prophetic activity (late 6th century BC) of Haggai and Zechariah, prophets who announced that Yahweh would restore the kingdom and the messianic vision would come to pass. Prerequisite to this messianic age was the rebuilding of the Temple (which was viewed as heaven on earth). When, however, the Temple had been rebuilt and long years had passed with neither the kingdom being restored nor the messianic age initiated, Israelite prophecy declined.

There is a tendency in prophetic preaching to spiritualize those aspects of religion that remain unfulfilled; herein lie the roots of eschatology, which is concerned with the last times, and apocalyptic literature, which describes the intervention of God in history to the accompaniment of dramatic, cataclysmic events. Since the predictions of the classical prophets were not fulfilled in a messianic age within history, these visions were translated into a historical apocalypse, such as Daniel. Why prophecy died out in Israel is difficult to determine, but Zechariah offers as good an answer as any in saying that the prophets "in those days" told lies. Prophets did appear, but after Malachi none gained the status of the classical prophets. Another reason may be found in Ezra's reform of the cult in the 5th century BC, in which Yahwism was so firmly established that there was no longer any need for the old polemics against Canaanite religion.

Reinterpretation of the prophetic heritage

Prophecy and apocalyptic literature. With the advent of postexilic Judaism (Ezra and after), including its emphasis on law and cult, there was not much room left for prophecy. The prophetic heritage was channelled through the teaching of their words. What remained of prophetic activity was expressed in various literary works that claimed esoteric knowledge of the divine purpose. The apocalyptic writers saw themselves as taking over and carrying on the prophetic task, but they went beyond the prophets in their use of old mythological motifs. The events they described had usually occurred long ago, but their recounting of these events was for the purpose of hinting and even predicting the events of the future. There was a far greater emphasis upon predictive speculation about the future than on the prophetic analysis and insight into history. The apocalyptic authors wrote pseudonymously, using the names of ancient worthies (such as Adam, Enoch, Abraham, Daniel, and Ezra). The literature is predominantly prose, but that of the classical prophets was predominantly poetry. Apocalyptic language is lavish in its use of fantastic imagery, frequently using riddles and numerical speculations. In apocalyptic literature angelology came into full blossom, with accounts of fallen angels (fallen stars) caught up in the forces opposed to God, frequently pictured in the old mythological motif of the struggle between darkness and light. Wild beasts symbolized peoples and nations, and there were esoteric calculations and speculations about the different eras through which history was passing as the world approached the eschaton (the consummation of history).

Dominant in apocalyptic literature is the theme of God's sovereignty and ultimate rule over all the universe. The message of the apocalyptic writers is one of both warning of the doom to come at the end of history, and hope in the new age beyond history under the rule of God, when the righteous will be vindicated.

Prophecy and prophetic religion in postbiblical Judaism. Though prophecy did not cease functioning in early Judaism, rabbinical Judaism—that influenced by rabbis, scholars, and commentators of the Bible—sought to limit it by advocating the pre-exilic era as the classical time of prophecy. Prophecy was not suppressed, but it came to be encircled by the law (Torah) in that all prophecy had to be in harmony with Torah, which was the definitive revelation of God's will. Thus, rabbinical Judaism gave prophecy its place of importance, but only as a phenomenon of the past. Such a theological stricture could not restrain the charismatic, eschatologically oriented patriots who arose during the time of Roman hegemony (mid-1st century BC–4th century AD). One rabbi, Akiba ben Joseph, joined with a messianic pretender, Simeon bar Kokhba (ber Kochba), in a revolt (132–135) and functioned as a prophet within that movement.

Eschatologically oriented prophets

Some prophets are known from the period of Hellenistic Judaism. I Maccabees, chapter 14, relates that Simon Maccabeus, who finally secured political independence for Judaea in 142 BC, was chosen as "leader and high priest forever, until a trustworthy prophet should arise." The same notion of a prophet soon to appear is expressed in chapter 1 of I Maccabees. The Hasmonean (Maccabean) prince John Hyrcanus (reigned 135–104 BC) was regarded as fulfilling these expectations and was called a prophet by the 1st-century AD Jewish historian Josephus (Jewish *War).* Josephus also mentions some Zealots (Jewish revolutionaries) as prophets and also one Jesus, son of Ananias, who in AD 62 predicted the destruction of the Temple and the defeat of the Jews. Josephus also mentions the seer Simon, a prophet leader (Antiquities), and Menahem, who prophesied in the 1st century BC. Among the followers of Judas Maccabeus, the leader of the 2nd-century-BC revolt, there apparently were persons who divined knowledge of the future. These and other notations indicate that seers and prophets played an important role in the intertestamental and postbiblical periods.

Jewish theology in Alexandria (Egypt) took up early rabbinical ideas and postulated that the will of God was to be discerned in the Torah and affirmed that the interpretation of law succeeded both the prophetic office and the role of sages. The law was thus considered to be superior to prophetic teaching. The Jewish philosopher Philo of Alexandria (*c.* 30 BC–after AD 40) affirmed that the Jews are a people of prophets. He also asserted that when a prophet has reached the fourth and final stage of ecstasy he is ready to become an instrument of divine power. Though Philo was influenced by Hellenistic concepts of prophecy, his basic foundation was still the Old Testament. Later rabbis believed that prophecy, though it was a gift from the world beyond, still required some knowledge. In rabbinic discussions of the nature of truth, it was generally held that reason alone was necessary but insufficient; prophecy could supply what was missing.

Superiority of law in relation to prophecy

The medieval Jewish philosopher Maimonides understood prophecy as an emanation from God to the intellect of man. Thus, prophecy could not be acquired by human effort. The divine gift of prophecy was bestowed upon those with both mental and moral perfection, combined with the presence of superior imagination. Opponents of this view advocated that Maimonides' concept of prophecy was not Jewish because Jewish prophecy always showed itself to be miraculous (see also JUDAISM; JUDAISM, HISTORY OF).

PROPHECY IN CHRISTIANITY

Divination and prophecy in the Hellenistic world. The problem of false prophets that occurred in Old Testament times also occurred in the early Christian communities. Prophets and diviners were widespread throughout the Hellenistic world. The Greek *prophētēs* was not only a forthteller but also an interpreter of divine messages. In addition, there were mantics (from the Greek mantis)—*i.e.*, visionary seers—whose visions were interpreted by prophets, soothsayers, diviners of all kinds, and especially astrologers. The impetus for much of this activity came from Babylonia. The influx of new religions from the East brought a profusion of astrologers and prophets. Many schools of astrology were founded throughout the Hellenistic world, and old schools of philosophy became very much occupied with astrology.

New Testament and early Christianity. Prophecy in the New Testament is seen as both a continuation of Old Testament prophecy as well as its fulfillment. For New Testament authors, the correct interpretation of Old Testament prophecy is that it speaks in *toto* of Christ. To prove their point, they often cite passages of Old Testament prophecy that are then elucidated as the words of God about Christ. New Testament writers follow Jesus himself in this matter, and Jesus is taken to be the prophet that was promised in Deuteronomy (see John 1:45, cf. 5:39, 6:14; Acts 3:22 *ff.*). Jesus regarded himself as a prophet, and so did some of his contemporaries. One special aspect of the prophetic image, however, is missing in Jesus: he was not an ecstatic, although supernatural revelations are found in connection with him; *e.g.*, the transfiguration of Jesus as witnessed by some of his Apostles on Mt. Tabor. In these New Testament descriptions of the transfiguration, Jesus is proclaimed to be the Son of God in words borrowed directly from Old Testament enthronement ritual. As a prophet, Jesus predicted his own death, his return as the Son of man at the end of the world, and the destruction of the Jerusalem Temple. At many points, Jesus is compared with and interpreted by the classical prophets in New Testament writings: his death—seen as the martyrdom of a prophet, his sufferings, and even his identity.

Though the New Testament describes Jesus as a prophet, he is at the same time believed to be more than a prophet: he is the expected Messiah (Greek *christos*, "anointed one"), predicted by prophets of old, who should reign as the Son of David and the Son of God. The royal ideology of the Old Testament was most important to early Christianity, for herein lay the seeds of its doctrines of Christ (see also MESSIAH AND MESSIANIC MOVEMENTS).

Several prophets are mentioned in the New Testament. One, Zechariah, is said to have perished "between the altar and the sanctuary" (Luke). Reference to his death is included by the Gospel writers because he was the last prophet before Jesus to have been killed by the Jews. Zechariah, the father of John the Baptist, uttered the Benedictus ("Blessed," the initial Latin word of the prophetic song) under the inspiration of the spirit. His wife, Elizabeth, also was described as being inspired by the spirit.

Others are Simeon, the prophetess Anna, and John the Baptist. These prophets are conceived by the New Testament writers as the termination of Old Testament prophecy, a concept also expressed by Jesus with reference to John the Baptist.

New Testament prophetic figures

The New Testament mentions several prophetic figures in the early church. Among them are Agabus of Jerusalem; Judas Barsabbas and Silas, who also were elders of the Jerusalem Church; the four prophesying daughters of

Philip the evangelist; and John, the author of Revelation. The term prophet is used with reference to an office in the early church along with evangelists and teachers, and the recipient of the letter bearing his name, Timothy, is called both a minister and a prophet. The prophet's role in the early church was to reveal divine mysteries and God's plan of salvation. Paul instructed his followers in the correct use of prophecy, and evaluated it as more beneficial to the life of congregations than ecstatic glossolalia (speaking in tongues). He considered prophecy to be the greatest spiritual gift from God, and in his view a prophet therefore ranks ahead of evangelists and teachers. With all this prophetic activity, the problem of false prophecy was crucial, and warnings against it abound in the New Testament. The most dangerous of the false prophets is predicted in the book of Revelation to John as yet to come. Many of these prophets, viewed as magicians and exorcists, are condemned for inducing chaos and for leading people astray. Therefore all prophetic activity had to be examined.

In the period immediately after the Apostles, prophets continued to play an important leadership role in the church, sometimes being called high priests. They were the only ones permitted to speak freely in the liturgy because of their inspiration by the Holy Spirit. Gradually, however, the liturgy became more and more fixed, and less freedom and innovation was permitted; this change, combined with the threat of false prophecy, eliminated these charismatic personalities. Among the heretical sects that advocated a return to prophetic activity, Montanism (2nd century), led by the prophet Montanus, advocated that the spirit of truth had come through Montanus. The freedom of doctrinal innovation that Montanus advocated could well have led to doctrinal anarchy, and the result of the struggle against this heresy was the suppression of charismatic prophecy, wherein ecstatic inspiration came to be viewed by the church as demonic.

Montanism

Another prophet who created a problem in the early church was Mani—the 3rd-century founder of a dualistic religion that was to bear his name (Manichaeism)—who considered himself to be the final messenger of God, after whom there was to be no other.

Prophetic and millenarian movements in later Christianity. In Western medieval church doctrines and rituals, active prophecy had no place. Prophetic activity was carried on, however, through holy orders. Mystically oriented holy men would sometimes appear as prophets with a special message, and even ecstatics found their places within the monasteries. In Eastern Christianity, monastic life stressed training in mystical experience.

Throughout Christian history there have been millenarian movements, usually led by prophetic-type personalities and based on the New Testament belief in Christ's return. Their basic doctrine is chiliasm (from Greek *chilioi,* "thousand"), which affirms that Christ will come to earth in a visible form and set up a theocratic kingdom over all the world and thus usher in the millennium, or the 1,000-year reign of Christ and his elect.

The early and medieval church hierarchy generally opposed chiliasm because such movements often became associated with nationalistic aspirations. Though the key leaders of the Protestant Reformation opposed chiliasm. and therefore minimized its effects upon the emergent denominations (*e.g.,* Lutheran, Calvinist, and Anglican), chiliasm did influence Anabaptist circles (radical reformation groups), and through them chiliastic ideas influenced Protestant Reformed theology and have appeared in reform movements, such as Pietism in Lutheran churches, and various revivalistic movements.

PROPHECY IN ISLAM

The centrality of prophecy in Islām. Pre-Islāmic prophecy in Arabia was no different in character from other Semitic prophecy. Pre-Islāmic terms for prophet are *ʿarrāf* and *kāhin* ("seer," cognate to Hebrew ***cohen,*** "priest"). The *kāhin* could often be a priest, and as a diviner he was an ecstatic. The *kāhin* was considered to be possessed by a *jinnī* ("spirit"), by means of whose power miracles could be performed. Also, poets were considered to be possessed by a *jinnī* through whose inspiration they composed their verses. The importance of the seers and diviners was noted in all aspects of life. Any problem might be submitted to such men, and their oracular answers were given with divine authority. It is not surprising, therefore, to find that a *kāhin* often became a sheikh, a temporal leader, and there were instances in which the position of *kāhin* was hereditary.

The role of the *kāhin*

It was against this background that the founder of Islām, Muhammad, appeared. During his early career in Mecca (in Arabia) he was considered by his tribesmen, the Quraysh, to be only another jinni-possessed *kāhin.* His utterances during this time were delivered in the same rhymed style as that used by other Arab prophets and were mostly the products of ecstatic trances. At about 40 years of age Muhammad experienced the promptings of the one god, Allāh, and retreated into the solitude of the mountains. These retreats served psychologically as preparations for his later revelations. The central religious problem of Muhammad was the fact that Jews had their sacred scriptures in Hebrew, and Christians had theirs in Greek, but there was no written divine knowledge in Arabic. Muhammad's preoccupation with this concern, along with a sense of the coming Day of Judgment, became the seeds of his new religion. Contemplation had matured Muhammad, and biographers point out that, as one may conclude from the Qur'Bn, Muhammad received the divine call in a vision. His ecstatic revelations were in the form of auditions, usually involving the angel Gabriel reading the divine message from a book. The illiterate Muhammad had his wife Khadijah, who was 15 years his senior, record them, and they are preserved in the Qur'Bn. Because this is believed to be a verbatim copy of the Heavenly Book, literally the words of Allāh himself, it cannot be questioned.

Muhammad considered himself to be more than a mere prophet (*nābi*); he thought of himself as the messenger (*rasūl*) of Allāh, the final messenger in a long chain that had begun with Noah and run through Jesus. As Allāh's *rasūl,* Muhammad saw his first mission to be that of warning the Arab peoples of the impending doomsday. No doubt Muhammad was influenced by the Judeo-Christian tradition in his concept of the Day of Judgment, as well as in his concept of himself as a prophet. Muhammad, who had felt at one time that Arabs were religiously inferior to Jews and Christians, became the medium of revelations that created Islām, and raised the Arabs in Muhammad's own evaluation to a status equal with that of the other two religions.

After AD 622, when Muhammad left Mecca and found refuge in Medina, ecstatic revelations began to play a secondary role in his prophecy—due to his political concerns—and not only does the rhymed prose of his message give way to more conventional prose but the content is more obviously the product of reasoned reflection on all aspects of life.

The Qur'ānic doctrines of prophecy. An official Islāmic view, and also that of Muhammad himself, was that Muhammad was the final Prophet. The Qur'Bn mentions those men who are considered to have imparted divine knowledge: Adam, Noah, Abraham, Isaac, Jacob, Moses, David, and Jesus. None of these revealed Allāh's message in full, since they were sent only to one nation. Muhammad, on the other hand, was sent to all nations and also to the *jinn.* The messages of the prophets before Muhammad were believed to have been either forgotten or distorted, but Islām claims that the Qur'Bn both corrects and confirms the sayings of the earlier prophets; Muhammad is the "seal of the prophets"; *i.e.,* the end of prophecy. All prophecy before Muhammad is incomplete and points to the coming of the final revelation.

The final Prophet

The prophetic activity of Muhammad serves as the foundation of Islām and Muslim society. The incomparable revelations of Muhammad are believed to have brought true monotheism into the world, to which nothing can be added or taken away. Thus, there is no more need of prophets or revelations (see also QUR'AN).

Later theological and philosophical doctrines. After the death of Muhammad, the expansion of Islām brought it into contact with the world at large, and a Muslim culture (involving science, philosophy, and literature) emerged, partially as a result of the Muslim acquisition of Byzantine culture. Christians and Jews became advisers and officials in Muslim courts. Christian philosophers introduced Muslim students to the works of the 4th-century-BC Greek philosopher Aristotle and to Neoplatonism (a philosophical system concerning the complex levels of reality), to theories about the nature of man, to theology, to the nature of existence, and to cosmology. Philosophical discussions about God, however, leave little or no room for prophets, and the savant displaced the prophet as the one proclaiming the will of God. As religious leaders, the savants were the keepers of sunnah (the life and habits of the prophet) and hadith (traditions about Muhammad's utterances and actions), which are supplements to the Qur'ān. Study of *ḥadīth* and *sunnah* contributed to the beginning of scholarly and scholastic activities in Islām, from which study emerged the Muslim system of duties and obligations (*fiqh*). Muslim theology began in the formulation of the doctrine of the general consensus (*ijmāʿ*), which was used to determine what was genuine sunnah. None ventured to question that Allāh was the only God, that Muhammad was his prophetic messenger, or that the Qur'ān was Allāh's word; to have done so would have been tantamount to admitting that one was not a Muslim.

Displacement of prophets

Scholastic philosophy was first introduced openly into Muslim theology by al-Ash'ari (10th century) who was the first to give Islām a systematic exposition. Another theologian, Ibn Sina (Avicenna), considered prophecy still to be a fundamental aspect of Islām, but for him, a prophet was not the spirit-possessed spokesman of God but rather an intelligent, intuitive man whose insight results in a place of leadership in society. Another philosopher, Ibn Rushd (Averroes), denied the belief that man's knowledge could ever be the same as God's knowledge; he also denied doctrines of predestination and corporeal resurrection, both of which were aspects of Muhammad's message.

Prophetic figures after Muḥammad. The fact that Muhammad was considered to be the final prophet did not end prophecy in Islām. After Muhammad's death, several seers proclaimed themselves his successors. Muhammad had designated no one to succeed himself, and left no sons. Abū Bakr, the father of Muhammad's wife ʿA'ishah, was chosen caliph (Arabic khalifah, "substitute, deputy"), but this did not discourage others from claiming that they were called of Allāh and thus trying to lead their own tribes as Muhammad had led his. Such movements were crushed by force, which contributed to the rapid expansion of Islām.

The *mahdī*

Some prophets claimed that they were long-awaited saviour-deliverers (mahdi, "restorer of the faith") and even gained some following beyond their own local tribes. Muhammad Aḥmad ibn as-Sayyid 'Abd Allāh of the Sudan preached a holy war against Egypt (1881) and fought and defeated the British governor-general C.G. Gordon at Khartoum in 1885. In India (Punjab), Mirza Ghulan Aḥmad claimed that he had received the spirit of Jesus and that he was a prophet-messiah. He recorded his revelations from Allāh in a book. Considering himself to be the Christ to his generation, he set out to reform Islām by liberalizing strict orthodoxy, yet avoiding the extremes of the pro-western movements of his time. He gained a large following among middle-class Muslims, but was soon disowned by orthodox Islām. His sect (Ahmadiyah), though small in numbers, has through its missionary activities spread over much of the world. Its socio-political stance is similar to that of the Black Muslims of the United States (see also ISLAM).

PROPHECY IN OTHER RELIGIONS

Pronhetic movements and figures in the Eastern religions. Buddhist literature contains predictions of a certain Buddha Maitreya, who will come as a kind of saviour-messiah to inaugurate a paradisaical age on earth. Gautama the Buddha himself, the 6th-century-BC founder of Buddhism, mentioned this prediction. Among the Hindus, the *Purāṇa* literature ("old history") contains prophetic passages, but these are to be understood as predictions after the event has occurred. Hindu religion has had many prophetic reformers, and the tribes of India, in their struggle for freedom, have produced prophets who combined the ideas of religious freedom with the hope of political and social freedom. The Oraons, a tribe in Chota Nāgpur, saw several prophets (bhagats) appear around the turn of the 20th century. Their intent was to free their people from foreign culture and political rule, returning to the older Hindu culture and religion. Such efforts often led to armed rebellion and ended in disaster.

In ancient China, divination was commonplace. One Confucian book involving divination, the "Classic of Changes," may have been connected with pre-Han Confucianism (before the 3rd century BC). Classical Confucian religion, however, emphasized the importance of rational process over inspiration and divination. Autocratic governments eliminated any such revolutionary, prophet-led movements as occurred in India, and any prophecy against the establishment was regarded as heretical. Inspired prophecy found little place in the official state religion. This situation did not rule out prophecy in folk religion, in which prophets appeared and promised their followers the good life in this world and in the next. In modern times, some of these movements became religio-political movements, as when Hung Hsiu-Ch'üan, an ecstatic epileptic noble of the middle 19th century, started a movement called the Taiping ("Great Peace"), a sect claiming that it was establishing the correct political order anew. Hung's movement—perhaps under the impact of Protestant missions—was quite austere, and it opposed magic, idols, and belief in spirits. He considered the New Testament to be authoritative for his new sect, and its rapid growth—aided by connections with other revolutionary movements—soon resulted in a genuine danger to the Manchu ruler of China. The Taiping Rebellion was crushed by Gordon in 1864.

Folk religion prophets

Diviners and shamans (male and female) are well represented in old Japanese Shintb. Japanese shamanism, which was closely related to Korean shamanism, often played a role in political disturbances and still does. Among old Japanese Buddhist sects is that founded by Nichiren (13th century AD), a prophetic enthusiast, religious revivalist, and zealous nationalist who taught that the Japanese people were the chosen people of God. In the Shintō revival movements of the 18th and 19th centuries, inspired persons with eschatological concepts founded movements that became messianic in character, and drew many of their followers from among the farmers, many of whom had practiced a Buddhist folk piety.

Prophetic movements and figures in the primitive religions. In many primitive religions, especially those of Africa, shamans, seers, and prophets are quite common. The same distinction between technical divination and charismatic prophecy is to be found in these cultures as in the ancient Near East. When it is possible to trace the history of prophetic activity in Africa, scholars usually find that it arises in times of confrontations with foreign cultures and with the advent of new religions. A sharp distinction between the diviner and the prophet cannot always be maintained, for diviners sometimes appear as prophets. A diviner may hear the voice of a god or spirit in his dreams and visions (in Zulu he is called a "dreamhouse") and receive a message. Some prophets, avowing a call, deny any training in prophecy. There are many parallels with the "rebel" prophets of India. Ecstatic prophets have played an important role not only in chiliastic and messianic movements but also in those movements opposing imperialism and European colonization of Africa. Their goal was and is a return to the old African culture and religion. Eschatological motifs have often been used in the prophetic preaching of tribal and national movements aspiring for freedom. Many of these prophets took up Christian ideas. Nxele, a 19th-century prophet of the South African Xhosas, preached the return of the dead on a certain day, and his successor, Mlandsheni, claimed to be the reincarnation of Nxele. He and others like him were healers and miracle workers.

Some of the prophetic founders of reform movements, which often were more political than religious, became messianic figures. Other prophets started out as Christian converts but came to a strong awareness that God had destined them to separate from their churches and lead syncretistic movements (fusions of various sources), all of which incorporate aspects of old African religion and, often, allow polygamy. In all these movements, syncretistic or not, there are also many prophetesses.

Prophets also have been found among American Indians. In 1675, a medicine man, Popé, arose as a prophetic leader among the Pueblo Indians. He preached the end of Spanish tyranny and a restoration of Indian sovereignty. At the height of the movement, several massacres took place, along with the burning of various church buildings (see also NATIVISTIC RELIGIOUS MOVEMENTS).

BIBLIOGRAPHY

General: G. HOLSCHER, *Die Propheten* (1914), a classic; A.J. HESCHEL, *The Prophets* (1962), a theological comparison between Israelite and non-Israelite prophets; J. LINDBLOM, *Prophecy in Ancient Israel* (1962), a good introduction to the phenomenological, psychological and theological problems of prophecy; G. WIDENGREN, *Religionsphänomenologie* (1969); R.B.Y. SCOTT, *The Relevance of the Prophets,* 2nd ed. (1968).

Prophecy in the ancient Near East and Israel: H.B. HUFFMON, "Prophecy in the Mari Letters," *Biblical Archaeologist,* 31:101–124 (1968); A. GUILLAUME, *Prophecy and Divinatiori Among the Hebrews and Other Semites* (1938), a standard work; D.R. HILLERS, *Treaty-Curses and Old Testament Prophets* (1964); W.L. MORAN, "New Evidence from Mari on the History of Prophecy," *Biblica,* 50:15–56 (1969); A.L. OPPENHEIM, *Ancient Mesopotamia: Portrait of a Dead Civilization* (1964); G. WIDENGREN, *Die Religionen Irans* (1965); for a collection of the available prophetical texts from Mari thus far, see F. ELLERMEIER, *Prophetie in Mari und Israel* (1968); G.W. AHLSTROM, "Some Remarks on Prophets and Cult," in J.C. RYLAARSDAM (ed.), *Transitions in Biblical Scholarship* (1968); R.E. CLEMENTS, *Prophecy and Covenant* (1965), a valuable study; I. ENGNELL, *A Rigid Scrutiny,* ed. and trans. by J.T. WILLIS (1969), a somewhat unorthodox treatment of prophecy; N.K. GOTTWALD, *All the Kingdoms of the Earth* (1964), on prophets and politics; E. HAMMERSHAIMB, *Some Aspects of Old Testament Prophecy from Isaiah to Malachi* (1966), deals with the Canaanite, cultic, and historical background; B.W. ANDERSON and W. HARRELSON (eds.), *Israel's Prophetic Heritage* (1962); A.R. JOHNSON, *The Cultic Prophet in Ancient Israel,* 2nd ed. (1962); J. PEDERSEN, *Israel,* 4 vol. (1926-40), a classic on religious life and institutions; H. RINGGREN, *Israelite Religion* (1966); H.H. ROWLEY, "Ritual and the Hebrew Prophets," in S.H. HOOKE (ed.), *Myth, Ritual, and Kingship* (1958); G. WIDENGREN, *Literary and Psychological Aspects of the Hebrew Prophets* (1948); W. ZIMMERLI, *The Law and the Prophets* (Eng. trans. 1965); S.B. FROST, *Old Testament Apocalyptic: Its Origins and Growth* (1952); H.H. ROWLEY, *The Relevance of Apocalyptic,* 3rd ed. (1963); B. VAWTER, "Apocalyptic: Its Relation to Prophecy," *Catholic Biblical Quarterly,* 22:33–46 (1960); P.D. HANSON, "Jewish Apocalyptic against its Near Eastern Environment," *Revue Biblique* 78:31–58 (1971); R. MEYER, "Prophecy and Prophets in the Judaism of the Hellenistic-Roman Period," in G. FRIEDRICH (ed.), *Theological Dictionary of the New Testament,* vol. 6, pp. 812–828 (1968); see also R. RENDTORFF, *ibid.,* pp. 783–812.

Prophecy in Christianity: L. HARTMAN, *Prophecy Interpreted* (1966); H.A. GUY, *New Testament Prophecy* (1947); G FRIEDRICH, "Prophets and Prophecies in the New Testament," *Theological Dictionary of the New Testament,* vol. 6, pp. 828–861 (1968); S. UMEN, *Pharisaism and Jesus* (1963).

Prophecy in Islām: T. ANDRAE, *Mohammed: The Man and His Faith* (Eng. trans. 1956); S. FUCHS, *Rebellious Prophets* (1965); H.A.R. GIBB, *Modern Trends in Islam* (1947); A. GUILLAUME, *Islam,* new ed. (1963); P.K. HITTI, *Islam: A Way of Life* (1970); W. MONTGOMERY WATT, "Muhammad," *The Cambridge History of Islam,* 1:30–56 (1970).

Prophetic movements and figures in Eastern and primitive religions: I. HORI, *Folk Religion in Japan: Continuity and Change,* ed. by J.M. KITAGAWA and A.L. MILLER (1968); E.R. and K. HUGHES, *Religion in China* (1950); J. MOONEY, *The Ghost-Dance Religion and the Sioux Outbreak of 1890* (1965); B.G.M. SUNDKLER, *Bantu Prophets in South Africa,* 2nd ed. (1961); M. WEBER, *The Religion of China,* trans. by H.H. GERTH (1968). See also S. FUCHS, *Rebellious Prophets* (1965).

(G.W.A.)

Prosody

As it has come to be defined in modern criticism, the term prosody encompasses the study of all of the elements of language that contribute toward acoustic and rhythmic effects, chiefly in poetry but also in prose. The term derived from an ancient Greek word that originally meant a song accompanied by music or the particular tone or accent given to an individual syllable. Greek and Latin literary critics generally regarded prosody as part of grammar; it concerned itself with the rules determining the length or shortness of a syllable, with syllabic quantity, and with how the various combinations of short and long syllables formed the metres (*i.e.*, the rhythmic patterns) of Greek and Latin poetry. Prosody was the study of metre and its uses in lyric, epic, and dramatic verse. In sophisticated modern criticism, however, the scope of prosodic study has been expanded until it now concerns itself with what the 20th-century poet Ezra Pound called "the articulation of the total sound of a poem."

Prosody and prose rhythm

Prose as well as verse reveals the use of rhythm and sound effects; however, critics do not speak of "the prosody of prose" but of prose rhythm. The English critic George Saintsbury wrote *A History of English Prosody from the Twelfth Century to the Present* (3 vol., 1906–10), which treats English poetry from its origins to the end of the 19th century; but he dealt with prose rhythm in an entirely separate work, *A History of English Prose Rhythm* (1912). Many prosodic elements such as the rhythmic repetition of consonants (alliteration) or of vowel sounds (assonance) occur in prose; the repetition of syntactical and grammatical patterns also generates rhythmic effect. Traditional rhetoric, the study of how words work, dealt with acoustic and rhythmic techniques in Classical oratory and literary prose. But although prosody and rhetoric intersected, rhetoric dealt more exactly with verbal meaning than with verbal surface. Rhetoric dealt with grammatical and syntactical manipulations and with figures of speech; it categorized the kinds of metaphor. Modern critics, especially those who practice the New Criticism, might be considered rhetoricians in their detailed concern with such devices as irony, paradox, and ambiguity. These subjects are discussed at greater length in the articles RHETORIC; LITERARY CRITICISM.

This article will discuss prosody chiefly in terms of the English language—the only language that all of the readers of this article may be assumed to know. Some examples are given in other languages to illustrate particular points about the development of prosody in those languages; because these examples are pertinent only for their rhythm and sound, and not at all for their meaning, no translations are given. A further general discussion of the development of prosodic elements will be found in the article POETRY; prosodic development in specific literatures is discussed in articles on the history of literature, such as LITERATURE, WESTERN; LITERATURE, EAST ASIAN; and in the literature sections of articles on the arts of various peoples, such as AFRICAN PEOPLES, ARTS OF; AMERICAN INDIAN PEOPLES, ARTS OF; etc.

ELEMENTS OF PROSODY

As a part of modem literary criticism, prosody is concerned with the study of rhythm and sound effects as they occur in verse and with the various descriptive, historical, and theoretical approaches to the study of these structures.

Scansion. The various elements of prosody may be examined in the aesthetic structure of prose. The celebrated opening passage of Charles Dickens' novel *Bleak House* (1853) affords a compelling example of prose made vivid through the devices of rhythm and sound:

> Fog everywhere. Fog up the river, where it flows among green aits and meadows; fog down the river, where it rolls defiled among the tiers of shipping, and the waterside pollutions of a great (and dirty) city. Fog on the Essex Marshes, fog on the Kentish heights. Fog creeping into the cabooses of collier-brigs; fog lying out on the yards, and hovering in the rigging of great ships; fog drooping on the gunwales of barges and small boats. Fog in the

eyes and throatc of ancient Greenwich pensioners, wheezing by the firesides of their wards; fog in the stem and bowl of the afternoon pipe of the wrathful skipper. . . .

Two phrases of five syllables each ("Fog everywhere"; "Fog up the river") establish a powerful rhythmic expectation that is clinched in repetition:

. . . fog down the river . . . . Fog on the Essex . . . , fog on the Kentish . . . . Fog creeping into . . . ; . . . fog drooping on the . . . .

This phrase pattern can be scanned; that is, its structure of stressed and unstressed syllables might be translated into visual symbols:

´ ˘ ˘ ´ ˘
Fog down the ri ver.

(This scansion notation uses the following symbols: the acute accent [´] to mark metrically stressed syllables; the breve [˘] to mark metrically weak syllables; a single line [|] to mark the divisions between feet [*i.e.*, basic combinations of stressed and unstressed syllables]; a double line [‖] to mark the caesura, or pause in the line; a rest [A] to mark a syllable metrically expected but not actually occurring.) Such a grouping constitutes a rhythmic constant, or cadence, a pattern binding together the separate sentences and sentence fragments into a long surge of feeling. At one point in the passage, the rhythm sharpens into metre; a pattern of stressed and unstressed syllables falls into a regular sequence:

Cadence and metre

´ ˘ ˘ ´ ˘ ´ ˘ ´ ˘ ˘
Fog on the|Es sex|mar shes, ‖ fog on the|

´ ˘ ´
Ken tish|heights.

The line is a hexameter (*i.e.*, it comprises six feet), and each foot is either a dactyl (´ ˘ ˘) or a trochee (´ ˘).

The passage from Dickens is strongly characterized by alliteration, the repetition of stressed consonantal sounds:

Fog creeping into the cabooses of collier-brigs;

and by assonance, the patterned repetition of vowel sounds:

. . . fog down the river, where it rolls defiled among . . . .

Here the vowel sounds are symmetrically distributed: short, long and long, short. Thus, it is clear that Dickens uses loosely structured rhythms, or cadences, an occasional lapse into metre, and both alliteration and assonance.

The rhythm and sound of all prose are subject to analysis; but compared with even the simplest verse, the "prosodic" structure of prose seems haphazard, unconsidered. The poet organizes his structures of sound and rhythm into rhyme, stanzaic form, and, most importantly, metre. Indeed, the largest part of prosodical study is concerned with the varieties of metre, the nature and function of rhyme, and the ways in which lines of verse fall into regular patterns or stanzas. An analysis of "Vertue" by the 17th-century English poet George Herbert reveals how the elements of prosody combine into a complex organism, a life sustained by the technical means available to the poet. When the metre is scanned with the symbols, it can be seen (and heard) how metre in this poem consists of the regular recurrence of feet, how each foot is a pattern of phonetically stressed and unstressed syllables.

Scansion

˘ ´ ˘ ´ ˘ ´ ˘ ´
1 Sweet day,|so cool,|so calm,|so bright,
˘ ´ ˘ ´ ˘ ´ ˘ ´
2 The bri|dall of|the earth|and skie:
˘ ´ ˘ ´ ˘ ´ ˘ ´
3 The dew|shall weep|thy fall|to-night;
˘ ´ ˘ ´
4 For thou|must die.

˘ ´ ˘ ´ ´ ˘ ˘ ´
5 Sweet rose,|whose hue|an grie|and brave
´ ˘ ˘ ´ ˘ ´ ˘ ´
6 Bids the|rash ga|zer wipe|his eye:
˘ ´ ˘ ´ ˘ ´ ˘ ´
7 Thy root|is ev|er in|its grave,
˘ ´ ˘ ´
8 And thou (must die.

˘ ´ ´ ˘ ˘ ´ ˘ ´ ˘
9 Sweet spring,|full of|sweet dayes|and ro|ses,
˘ ´ ˘ ´ ˘ ´ ˘ ´
10 A box|where sweets|com pac|ted lie;
˘ ´ ˘ ´ ˘ ´ ˘ ´ ˘
11 My mu|sick shows|ye have|your clo|ses,
˘ ´ ˘ ´
12 And ail|must die.

´ ˘ ˘ ´ ˘ ´ ˘ ´
13 Onely|a sweet|and ver|tuous soul,
˘ ´ ˘ ´ ˘ ´ ˘ ´
14 Like sea|son'd tim|ber, ne|ver gives;
˘ ´ ˘ ´ ˘ ´ ˘ ´
15 But though|the whole|world turn|to coal,
˘ ´ ˘ ´
16 Then chief|ly lives.

The basic prosodic units are the foot, the line, and the stanza. The recurrence of similar feet in a line determines the metre; here there are three lines consisting of four iambic feet (*i.e.*, of four units in which the common pattern is the iamb—an unstressed syllable followed by a stressed syllable), which are followed by a line consisting of two iambic feet. Thus the stanza or recurring set of lines consists of three iambic tetrameters followed by one iambic dimeter. The stanzaic form is clinched by the use of rhyme; in "Vertue" the first and third and second and fourth lines end with the same sequence of vowels and consonants: bright/night, skie/die, brave/grave, eye/die, etc. It should be observed that the iambic pattern (˘´) is not invariable; the third foot of line 5, the first foot of line 6, the second foot of line 9, and the first foot of line **13** are reversals of the iambic foot or trochees (´˘). These reversals are called substitutions; they provide tension between metrical pattern and meaning, as they do in these celebrated examples from Shakespeare:

˘ ´ ˘ ´ ˘ ´ ´ ˘ ˘ ´ ˘
To be,|or not|to be, ‖ that is|the ques tion . . . .

Hamlet

˘ ´ ˘ ´ ´ ˘ ˘ ´ ˘ ´
His sil|ver skin|laced with|his gol|den blood . . .

Macbeth

Meaning, pace, and sound. Scansion reveals the basic metrical pattern of the poem; it does not, however, tell everything about its prosody. The metre combines with other elements, notably propositional sense or meaning, pace or tempo, and such sound effects as alliteration, assonance, and rhyme. In the fifth line of "Vertue," the reversed third foot occurring at "angry" brings that word into particular prominence; the disturbance of the metre combines with semantic re-enforcement to generate a powerful surge of feeling. Thus, the metre here is expressive. The pace of the lines is controlled by the length of vowel sounds; although lines 5 and 6 contain the same number of syllables and feet, line 5 obviously takes longer to read or recite. The line contains more long vowel sounds:

Sweet rose, whose hue angrie and brave. . .

Quantity

Vowel length is called quantity. In English verse, quantity cannot by itself form metre although a number of English poets have experimented with quantitative verse. Generally speaking, quantity is a rhythmical but not a metrical feature of English poetry; it can be felt but it cannot be precisely determined. The vowel sounds in "Sweet rose" may be lengthened or shortened at will. No such options are available, however, with the stress patterns of words; the word ´˘ an-gry cannot be read ˘´ an-gry.

Assonance takes into account the length and distribution of vowel sounds. A variety of vowel sounds can be noted in this line:

Sweet day, so cool, so calm, so bright . . .

To borrow a term from music, the line modulates from ēē, through ā, 66, ă, to i. Alliteration takes into account the recurrence and distribution of consonants:

so cool, so calm.. .
Sweet spring . . .

Rhyme normally occurs at the ends of lines; "Virtue" reveals, however, a notable example of interior rhyme, or rhyme within the line:

> My musick *shows* ye have your *closes*. . .

Types of metre. *Syllable-stress metres.* It has been shown that the metre of "Virtue" is determined by a pattern of stressed and unstressed syllables arranged into feet and that a precise number of feet determines the measure of the line. Such verse is called syllable-stress verse (in some terminologies accentual-syllabic) and was the norm for English poetry from the beginning of the 16th century to the end of the 19th century. A line of syllable-stress verse is made up of either two-syllable (disyllabic) or three-syllable (trisyllabic) feet. The disyllabic feet are the iamb and the trochee (noted in the scansion of "Virtue"); the trisyllabic feet are the dactyl (´∪∪) and anapest (∪∪´).

Following are illustrations of the four principal feet found in English verse:

iambic	be hold
trochaic	ti ger
dactylic	des per ate
anapestic	un der stand

Some theorists also admit the spondaic foot (´´) and pyrrhic foot (∪∪) into their scansions; however, spondees and pyrrhics occur only as substitutions for other feet, never as determinants of a metrical pattern:

> When to | the ses | sions of | sweet si | lent thought . . .

It has been noted that four iambic feet make up a line of tetrameter verse; a line consisting of one iambic foot is called monometer, of two dimeter, of three trimeter, of five pentameter, of six hexameter, and of seven heptameter. Lines containing more than seven feet rarely occur in English poetry.

The following examples illustrate the principal varieties of syllable-stress metres and their scansions:

Iambic (pentameter)

Then say | not Man's | im per | fect, ‖ Heaven | in fault;
Say ra | ther, ‖ Man's | as per | fect as | he ought:
His know | ledge meas | ured ‖ to | his state | and place,
His time | a mo | ment, ‖ and | a point | his space.

Alexander Pope, *An Essay on Man* (1733–34)

Trochaic (dimeter)

Could I | catch that
Nim ble | trai tor
Scorn ful | Lau ra,
Swift-foot | Lau ra,
Soon then | would I
Seek a | venge ment.

Thomas Campion (1602)

Dactylic (tetrameter)

Af ter the | pangs of a | des per ate | Lover, ^
When day and | night I have | sigh'd all in | vain, ^ ^
Ah what a | pleas ure it | is to dis | co ver ^
In her eyes | pi ty, who | cau ses my | pain! ^ ^

John Dryden, *An Evening's Love* (1671)

Anapestic (tetrameter)

The As syr | ian came down | like a wolf | on the fold,
And his co | horts were gleam | ing in pur | ple and gold;
And the sheen / of their spears | was like stars | on the sea,
When the blue | wave rolls night | ly on deep | Ga li lee.

Lord Byron, "The Destruction of Sennacherib" (1815)

Syllable stress became more or less established in the poetry of Geoffrey Chaucer (c. 1340–1400). In the century that intervened between Chaucer and the early Tudor poets, syllable-stress metres were either ignored or misconstrued. By the end of the 16th century, however, the now-familiar iambic, trochaic, dactylic, and anapestic metres became the traditional prosody for English verse.

Challenges to traditional prosody in the 19th century

Strong-stress metres. In the middle of the 19th century, with Walt Whitman's free verse and Gerard Manley Hopkins' extensive metrical innovations, the traditional prosody was challenged. Antecedent to the syllable-stress metres was the strong-stress metre of Old English and Middle English poetry. Strong-stress verse is measured by count of stresses alone; the strong stresses are usually constant, but the number of unstressed syllables may vary considerably.

Strong-stress verse survives in nursery rhymes and children's counting songs:

> One, two, ‖ buckle my shoe;
>
> Three, four, ‖ knock at the door;
>
> Five, six, ‖ pick up sticks . . .

The systematic employment of strong-stress metre can be observed in the Old English epic poem *Beowulf* (c. 1000) and in William Langland's vision-poem, *Piers Plowman* ('A' Text, *c.* 1362):

> In a somer sesun ‖ whon softe was the sonne,
>
> I schop me in-to a schroud ‖ a scheep as I were;
>
> In habite of an hermite ‖ un-holy of werkes,
>
> Wende I wydene in this world ‖ wondres to here.

These lines illustrate the structural pattern of strong-stress metre. Each line divides sharply at the caesura (‖), or medial pause; on each side of the caesura are two stressed syllables strongly marked by alliteration.

Strong-stress verse is indigenous to the Germanic languages with their wide-ranging levels of stressed syllables and opportunities for alliteration. Strong-stress metre was normative to Old English and Old Germanic heroic poetry, as well as to Old English lyric poetry. With the rising influence of French literature in the 12th and 13th centuries, rhyme replaced alliteration and stanzaic forms replaced the four-stress lines. But the strong-stress rhythm persisted; it can be felt in the anonymous love lyrics of the 14th century and in the popular ballads of the 15th century.

"Lord Randal" can be comfortably scanned to show a line of mixed iambic and anapestic feet; it clearly reveals, however, a four-stress structure:

> 'O where ha' you been, ‖ Lord Randal, my son?
>
> And where ha' you been, ‖ my handsome young man?'
>
> 'I ha' been at the greenwood ; ‖ mother, mak my bed soon,
>
> For I'm wearied wi' huntin', ‖ and fain wad lie down.'

A number of 20th-century poets, including Ezra Pound, T.S. Eliot, and W.H. Auden, have revived strong-stress metre. The versification of Pound's *Cantos* and Eliot's *Four Quartets* (1943) shows the vitality of the strong-stress, or, as they are often called, "native," metres.

Syllabic metres. The greatest part of English poetry is carried by the strong-stress and syllable-stress metres.

The dominant alexandrine

Two other kinds of metres must be mentioned: the purely syllabic metres and the quantitative metres. The count of syllables determines the metres of French, Italian, and Spanish verse. In French poetry (both dramatic and lyric) the alexandrine, or 12-syllabled line, is a dominant metrical form:

> O toi, qui vois la honte oh je suis descendue,
> Implacable Vbnus, suis-je assez confondue?
> Tu ne saurais plus loin pousser ta cruautb.
> Ton triomphe est parfait; tous tes traits ont porté.
> Racine, *Phèdre* (1677)

Stress and pause in these lines are variable; only the count of syllables is fixed. English poets have experimented with syllabic metres; the Tudor poet Thomas Wyat's translations from Petrarch's Italian poems of the 14th century attempted to establish a metrical form based on a decasyllabic or ten-syllabled line:

> The long love that in my thought doth harbor,
> And in my heart doth keep his residence,
> Into my face presseth with bold pretense
> And there encampeth, spreading his banner.
> "The Lover for Shamefastness Hideth . . ." **(1557)**

Most ears can detect that these lines waver between syllabic and syllable-stress metre; the second line falls into a pattern of iambic feet. Most ears also discover that the count of syllables alone does not produce any pronounced rhythmic interest; syllabic metres in English generate a prosody more interesting to the eye than to the ear.

Quantitative metres. Quantitative metres determine the prosody of Greek and Latin verse. Renaissance theorists and critics initiated a confused and complicated argument that tried to explain European poetry by the rules of Classical prosody and to draft laws of quantity by which European verse might move in the hexameters of the ancient Roman poets Virgil or Horace. Confusion was compounded because both poets and theorists used the traditional terminology of Greek and Latin prosody to describe the elements of the already existing syllable-stress metres; iambic, trochaic, dactylic, and anapestic originally named the strictly quantitative feet of Greek and Latin poetry. Poets themselves adapted the metres and stanzas of Classical poetry to their own languages; whereas it is not possible here to trace the history of Classical metres in European poetry, it is instructive to analyze some attempts to make English and German syllables move to Greek and Latin music. Because neither English nor German has fixed rules of quantity, the poets were forced to revise the formal schemes of the Classical paradigms in accordance with the phonetic structure of their own language.

A metrical paradigm much used by both Greek and Latin poets was the so-called Sapphic stanza. It consisted of three quantitative lines that scanned

-u---uu-u-u,

followed by a shorter line, called an Adonic,

–◡◡––.

Sapphics in English and German

"Sapphics" by the 19th-century English poet Algernon Charles Swinburne shows the Sapphic metre and stanza in English:

> **All** the night sleep came not upon my eyelids,
> Shed not dew, nor shook nor unclosed a feather,
> Yet with lips shut close and with eyes of iron
>   Stood and beheld me.. .
> Saw the white implacable Aphrodite,
> Saw the hair unbound and the feet unsandalled
> Shine as fire of sunset on western waters;
>   Saw the reluctant. . . .

The same metre and stanza in German are found in "Sapphische Ode," by the 19th-century poet Hans Schmidt, which was beautifully set to music by Johannes Brahms (Opus 94, No. 4):

> Rosen brach ich nachts mu am dunklen Hage;
> süsser hauchten Duft sie, als je am Tage;
> doch verstreuten reich die bewegten Äste
>   Tau den mich nasste.
> Auch der Küsse Duft mich wie nie beriickte,
> die ich nachts vom Strauch deiner Lippen pfluckte:
> doch auch dir, bewegt im Gernut gleich jenem,
>   tauten die Tranen.

Quantitative metres originated in Greek, a language in which the parts of speech appear in a variety of inflected forms (*i.e.*, changes of form to indicate distinctions in case, tense, mood, number, voice, and others). Complicated metrical patterns and long, slow-paced lines developed because the language was hospitable to polysyllabic metrical feet and to the alternation of the longer vowels characterizing the root syllables and the shorter vowels characterizing the inflected case-endings. The Classical metres can be more successfully adapted to German than to English because English lost most of its inflected forms in the 15th century, while German is still a highly inflected language. Thus Swinburne's "Sapphics" does not move as gracefully, as "naturally" as Schmidt's. A number of German poets, notably Goethe and Friedrich Holderlin, both of the early 19th century, made highly successful use of the Classical metres. English poets, however, have never been able to make English syllables move in the ancient metres with any degree of comfort or with any sense of vital rhythmic force.

The American poet Henry Wadsworth Longfellow adapted the Classical hexameter for his Evangeline (1847):

´ ◡ ◡ ´ ◡ ◡ ´ ◡ ◡ ´ ◡ ◡ ´ ◡
This is the | for est pri | me val. The | mur mu ring | pines and
◡ ´ ◡
the | hem locks . . .

In Virgil's Aeneid, Longfellow's Classical model, the opening line scans:

– ◡ ◡ – u u - – – ◡ – u u - -
Ar ma vi rum que ca no, Troj ae qui pri mus ab or is

The rules determining length of syllable in Classical Greek and Latin poetry are numerous and complicated; they were established by precise grammatical and phonetic conventions. No such rules and conventions obtain in English; Robert Bridges, the British poet laureate and an authority on prosody, remarked in his Poetical Works (1912) that the difficulty of adapting English syllables to the Greek rules is "very great, and even deterrent." Longfellow's hexameter is in reality a syllable-stress line of five dactyis and a final trochee; syllabic quantity plays no part in determining the metre.

PROSODIC STYLE

The analysis of prosodic style begins with recognizing the metrical form the poet uses. Is he writing syllable-stress, strong-stress, syllabic, or quantitative metre? Or is he using a nonmetrical prosody? Again, some theorists would not allow that poetry can be written without metre; the examples of Whitman and many 20th-century innovators, however, have convinced most modern critics that a nonmetrical prosody is not a contradiction in terms but an obvious feature of modern poetry. Metre has not disappeared as an important element of prosody; indeed, some of the greatest poets of the modern period—William Butler Yeats, T.S. Eliot, Ezra Pound, Wallace Stevens—revealed themselves as masters of the traditional metres. They also experimented with newer prosodies based on prose cadences, on expansions of the blank-verse line, and revivals of old forms—such as strong-stress and ballad metres. Also noteworthy are the "visual" prosodies fostered by the poets of the Imagist movement and by such experimenters as E.E. Cummings. Cummings revived the practice of certain 17th-century poets (notably George Herbert) of "shaping" the poem by typographic arrangements.

Non-metrical prosody

The prosodic practice of poets has varied enormously with the historical period, the poetic genre, and the poet's individual style. In English poetry, for example, during the Old English period (to 1100), the strong-stress metres carried both lyric and narrative verse. In the Middle English period (from *c.* 1100 to *c.* 1500), stanzaic forms developed for both lyric and narrative verse. The influence of French syllable counting pushed the older stress lines into newer rhythms; Chaucer developed for The Canterbury Tales a line of ten syllables with alternating accent and regular end rhyme—an ancestor of the heroic couplet. The period of the English Renaissance (from c. 1500 to 1660) marks the fixing of syllable-stress metre as normative for English poetry. Iambic metre carried three major prosodic forms: the sonnet, the rhyming couplet, and blank verse. The sonnet was the most important of the fixed stanzaic forms. The iambic pentameter rhyming couplet (later known as the heroic couplet) was used by Christopher Marlowe for his narrative poem *Hero* and Leander (1598); by John Donne in the early 17th century for his satires, his elegies, and his longer meditative poems. Blank verse (unrhymed iambic pentameter), first introduced into English in a translation by Henry Howard, earl of Surrey, published in 1557, became the metrical norm for Elizabethan drama. The period of the Renaissance also saw the refinement of a host of lyric and song forms; the rapid development of English music during the second half of the 16th century had a salutary effect on the expressive capabilities of poetic rhythms.

The personal element. A poet's choice of a prosody obviously depends on what his language and tradition afford; these are primary considerations. The anonymous author of the Old English poem *Deor* used the conventional four-stress metric available to him; but he punctuated groups of lines with a refrain:

þaes ofereode: þisses swa maeg!
(that passed away: this also may!)

The refrain adds something to the prosodic conventions of regulated stress, alliteration, and medial pause: a sense of a smaller and sharper rhythmic unit within the larger rhythms of the given metre. While the poet accepts from history his language and from poetic convention the structure of his metre, he shapes his own style through individual modifications of the carrying rhythms. When critics speak of a poet's "voice," his personal tone, they are also speaking of his prosodic style.

Style through tension

Prosodic style must be achieved through a sense of tension; it is no accident that the great masters of poetic rhythm work against the discipline of a given metrical form. In his sonnets, Shakespeare may proceed in solemn iambic regularity, creating an effect of measured progression through time and its legacy of suffering and despair:

No longer mourn for me when I am dead
Than you shall hear the surly sullen bell
Give warning to the world that I am fled. . . .
"Sonnet 71"

Or he may wrench the metre and allow the reader to feel the sudden violence of his feelings, the power of a conviction raised to a command:

´ ∪ ´ ∪ ∪ ´ ∪ ∪ ´
Let me|not to|the mar|riage of|true minds
∪ ´ ∪ ´ ∪ ´ • ∪ ∪ ´
Ad mit|im pe|di ments.|Love is|not love . . . .
"Sonnet 116"

The first two feet of the first line are trochaic reversals; the last two feet comprise a characteristic pyrrhic-spondaic formation. A trochaic substitution is auite normal in the first foot of an iambic pentameter line–a trochaic substitution in the second foot, however, creates a marked disturbance in the rhythm. There is only one "normal" iambic foot in the first line; this line runs over (or is enjambed) to the second line with its three consecutive iambic feet followed by a strong caesura and reversed fourth foot. These lines are, in Gerard Manley Hopkins' term, metrically "counter-pointed"; trochees, spondees, and pyrrhics are heard against a ground rhythm of regular iambics. Without the ground rhythm, Shakespeare's expressive departures would not be possible.

Whitman's innovations

A poet's prosodic style may show all of the earmarks of revolt against prevailing metrical practice. Whitman's celebrated "free verse" marks a dramatic break with the syllable-stress tradition; he normally does not count syllables, stresses, or feet in his long sweeping lines. Much of his prosody is rhetorical; that is, Whitman urges his language into rhythm by such means as anaphora (*i.e.*, repetition at the beginning of successive verses) and the repetition of syntactical units. He derives many of his techniques from the example of biblical verses, with their line of various types of parallelism. But he often moves toward traditional rhythms; lines fall into conventional parameters:

O past! O happy life! O songs of joy!
"Out of the Cradle Endlessly Rocking" (1859)

Or they fall more often into dissyllabic hexameters:

Borne through the smoke of the battles and pierc'd
with missiles I saw them . . . .
"When Lilacs Last in the Door-Yard Bloom'd" (1865–66)

Despite the frequent appearance of regular metrical sequences, Whitman's lines cannot be scanned by the usual graphic method of marking syllables and feet; his prosody, however, is fully available to analysis. The shape on the page of the lines below (they comprise a single strophe or verse unit) should be noted, specifically the gradual elongation and diminution of line length. Equally noteworthy are the repetition of the key word *carols*, the alliteration of the *s* sounds, and the use of words in falling (trochaic) rhythm, "lagging," "yellow," "waning":

Shake out carols!
Solitary here, the night's carols!
Carols of lonesome love! death's carols!
Carols under that lagging, yellow, waning moon!
O under that moon where she droops almost down into the sea!
O reckless despairing carols.
"Out of the Cradle"

No regular metre moves these lines; but a clearly articulated rhythm—produced by shape, thematic repetitions, sound effects, and patterns of stress and pause—defines a prosody.

Whitman's prosody marks a clear break with previous metrical practices. Often a new prosody modifies an existing metrical form or revives an obsolete one. In "Gerontion" (1920), T.S. Eliot adjusted the blank-verse line to the emotionally charged, prophetic utterance of his persona, a spiritually arid old man:

After such knowledge, what forgiveness? Think now
History has many cunning passages, contrived corridors
And issues, deceives with whispering ambitions,
Guides us by vanities. Think now. . .
(From T.S. Eliot, Collected Poems *1909–1962*,
Harcourt Brace Jovanovich, Inc.)

The stress prosodies of Pound and Eliot

The first three lines expand the pentameter line beyond its normal complement of stressed and unstressed syllables; the fourth line contracts, intensifying the arc of feeling. Both Pound and Eliot used stress prosodies. Pound counted out four strong beats and used alliteration in his brilliant adaptation of the old English poem "The Seafarer" (1912):

Chill its chains are; chafing sighs
Hew my heart round and hunger begot
Mere-weary mood. Lest man know not
That he on dry land loveliest liveth . . .
(From Ezra Pound, Personae, Copyright 1926 by Ezra Pound. Reprinted by permission of New Directions Publishing Corporation.)

He uses a similar metric for the energetic opening of his "Canto I." Eliot mutes the obvious elements of the form in the celebrated opening of The *Waste Land* (1922) :

April is the cruellest month, breeding
Lilacs out of the dead land, mixing
Memory and desire, stirring
Dull roots with spring rain.
(From T.S. Eliot, Collected Poems *1909–1962*,
Harcourt Brace Jovanovich, Inc.)

Here is the "native metre" with its falling rhythm, elegiac tone, strong pauses, and variably placed stresses. If this is free verse, its freedoms are most carefully controlled. "No verse is free," said Eliot, "for the man who wants to do a good job."

The prosodic styles of Whitman, Pound, and Eliot—though clearly linked to various historical antecedents—are innovative expressions of their individual talents. In a sense, the prosody of every poet of genius is unique; rhythm is perhaps the most personal element of the poet's expressive equipment. Alfred Lord Tennyson and Robert Browning, English poets who shared the intellectual and spiritual concerns of the Victorian age, are miles apart in their prosodies. Both used blank verse for their dramatic lyrics, poems that purport to render the accents of real men speaking. The blank verse of Tennyson's "Ulysses"

The prosody of Tennyson and Browning

(1842) offers smoothly modulated vowel music, carefully spaced spondaic substitutions, and unambiguous pentameter regularity:

The long day wanes; the slow moon climbs; the deep
Moans round with many voices. Come, my friends,
'Tis not too late to seek a newer world.

Browning's blank verse aims at colloquial vigour; its "irregularity" is a function not of any gross metrical violation—it always obeys the letter of the metrical law—but of the adjustment of abstract metrical pattern to the rhythms of dramatic speech. If Tennyson's ultimate model is Milton's Baroque prosody with its oratorical rhythms, Browning's model was the quick and nervous blank verse of the later Elizabethan dramatists. Characteristic of Browning's blank verse are the strong accents, involuted syntax, pregnant caesuras, and headlong energy in "The Bishop Orders His Tomb at St. Praxed's Church" (1845):

Vanity, saith the preacher, vanity!
Draw round my bed: is Anselm keeping back?
Nephews—sons mine . . . ah God, I know not! Well—
She, men would have to be your mother once,
Old Gandolf envied me, so fair she was!

Influence of period and genre. In the lyric genres, the rhythms of the individual poet—or, in the words of the 20th-century American poet Robert Lowell, "the person himself"—can be heard in the prosody. In the long poem, the dramatic, narrative, and didactic genres, a period style is more likely to be heard in prosody. The blank-verse tragedy of the Elizabethan and Jacobean dramatists, the blank verse of Milton's Paradise Lost (1667) and its imitators in the 18th century (James Thomson and William Cowper), and the heroic couplet of Neoclassical satiric and didactic verse, each, in different ways, defines the age in which these prosodies flourished. The flexibility and energy of the dramatic verse of Marlowe, Shakespeare, and John Webster reflect the later Renaissance with its nervous open-mindedness, its obsessions with power and domination, and its lapses into despair. Miltonic blank verse, based on Latin syntax and adaptations of the rules of Latin prosody, moved away from the looseness of the Elizabethans and Jacobeans toward a more ceremonial style. It is a Baroque style in that it exploits the musical qualities of sounds for their ornamental values. The heroic couplet, dominating the poetry of the entire 18th century, was unequivocally a prosodic period style; its elegance and epigrammatic precision entirely suited an age that valued critical judgment, satiric wit, and the powers of rationality.

Prosody in dramatic verse

It is in dramatic verse, perhaps, that a prosody shows its greatest vitality and clarity. Dramatic verse must make a direct impression not on an individual reader able to reconsider and meditate on what he has read but on an audience that must immediately respond to a declaiming actor or a singing chorus. The ancient Greek dramatists developed two distinct kinds of metres: "stichic" forms (*i.e.*, consisting of "stichs," or lines, as metrical units) such as the iambic trimeter for the spoken dialogues; and lyric, or strophic, forms (*i.e.*, consisting of stanzas), of great metrical intricacy, for the singing and chanting of choruses. Certain of the Greek metres developed a particular ethos; characters of low social standing never were assigned metres of the lyric variety. Similar distinctions obtained in Elizabethan-drama. Shakespeare's kings and noblemen speak blank verse: comic characters, servants, and country bumpkins discburse in prose; clowns, romantic heroines, and supernatural creatures sing songs. In the early tragedy Romeo and Juliet, the chorus speaks in "excellent conceited" sonnets: in what was one of the most popular and easily recognized lyric forms of the period.

The metrical forms used by ancient and Renaissance dramatists were determined by principles of decorum. The use or non-use of a metrical form (or the use of prose) was a matter of propriety: was the metre suitable to the social status and ethos of the individual character; was the metre suitable to the emotional intensity of the particular situation? Decorum, in turn, was a function of the dominant Classical and Neoclassical theories of imitation.

THEORIES OF PROSODY

Ancient critics like Aristotle and Horace insisted that certain metres were natural to the specific poetic genres; thus, Aristotle (in the Poetics) noted that "Nature herself, as we have said, teaches the choice of the proper measure." In epic verse the poet should use the heroic measure (dactylic hexameter) because this metre most effectively represents or imitates such qualities as grandeur, dignity, and high passion. Horace narrowed the theory of metrical decorum, making the choice of metre prescriptive; only an ill-bred and ignorant poet would treat comic material in metres appropriate to tragedy. Horace prepared the way for the legalisms of the Renaissance theorists who were quite willing to inform practicing poets that they used "feete without joyntes," in the words of Roger Ascham, Queen Elizabeth's tutor, and should use the quantitative metres of Classical prosody.

Middle Ages. During the Middle Ages little of importance was added to actual prosodic theory; in poetic practice, however, crucial developments were to have important ramifications for later theorists. From about the second half of the 6th century to the end of the 8th century, Latin verse was written that no longer observed the rules of quantity but was clearly structured on accentual and syllabic bases. This change was aided by the invention of the musical sequence; it became necessary to fit a musical phrase to a fixed number of syllables, and the older, highly complex system of quantitative prosody could not be adapted to simple melodies that must be sung in sequential patterns. In the musical sequence lies the origin of the modem lyric form.

The musical sequence

The 9th-century hymn "Ave maris stella" is a striking instance of the change from quantitative to accentual-syllabic prosody; each line contains three trochaic feet determined not by length of syllable but by syllabic intensity or stress:

Ave maris stella
Dei mater alma
atque semper virgo,
felix caeli porta.

Sumens illud Ave
Gabrielis ore,
funda nos in pace,
mutans nomen Evae

The rules of quantity have been disregarded or forgotten; rhyme and stanza and a strongly felt stress rhythm have taken their place. In the subsequent emergence of the European vernacular literatures, poetic forms follow the example of the later Latin hymns. The earliest art lyrics, those of the Provençal troubadours of the 12th and 13th centuries, show the most intricate and ingenious stanzaic forms. Similarly, the Goliardic songs of the Carmina Burana (13th century) reveal a rich variety of prosodic techniques; this "Spring-song" embodies varying lines of trochees and iambs and an ababcdccd rhyme scheme:

Ver redit optatum
  Cum gaudio,
Flore decoratum
  Purpureo;
Aves edunt cantus
  Quam dulciter!
Revirescit nemus,
Cantus est amoenus
  Totaliter.

Renaissance. Renaissance prosodic theory had to face the fact of an accomplished poetry in the vernacular that was not written in metres determined by "rules" handed down from the practice of Homer and Virgil. Nevertheless, the classicizing theorists of the 16th century made a determined attempt to explain existing poetry by the rules of short and long and to draft "laws" by which modem verse might move in Classical metres. Roger Ascham, in *The* Scholemaster (1570), attacked "the Gothic . . . barbarous and rude Ryming" of the early Tudor poets. He admitted that Henry Howard, earl of Surrey, did passably well as a poet but complained that Surrey did not understand "perfite and trewe versifying"; that is, Surrey did not compose his English verses according to the principles of Latin and Greek quantitative prosody.

Rules of "trewe versifying"

Ascham instigated a lengthy argument, continued by succeeding theorists and poets, on the nature of English prosody. Sir Philip Sidney, Gabriel Harvey, Edmund Spenser, and Thomas Campion all (to use Saintsbury's phrase) committed whoredom with the enchantress of quantitative metric. While this hanky-panky had no adverse effect on poetry itself (English poets went on writing verses in syllable-stress, the prosody most suitable to the language), it produced misbegotten twins of confusion and discord, whose heirs, however named, are still apparent today. Thus, those who still talk about "long and short" (instead of stressed and unstressed), those who perpetuate a punitive prosodic legalism, and those who regard prosody as an account of what poets should have done and did not, trace their ancestry back to Elizabethan dalliance and illicit classicizing.

Although Renaissance prosodic theory produced scarcely anything of value to either literary criticism or poetic technique—indeed, it did not even develop a rational scheme for scanning existing poetry—it raised a number of important questions. What were the structural principles animating the metres of English verse? What were the aesthetic nature of prosody and the functions of metre? What were the connections between poetry and music? Was poetry an art of imitation (as Aristotle and all of the Neoclassical theorists had maintained), and was its sister art painting; or was poetry (as Romantic theory maintained) an art of expression, and prosody the element that produced (in Coleridge's words) the sense of musical delight originating (in T.S. Eliot's words) in the auditory imagination?

Pope's doctrine of imitation

The 18th century. Early in the 18th century, Pope affirmed, in his ***Essay on Criticism*** (1711), the classic doctrine of imitation. Prosody was to be more nearly onomatopoetic; the movement of sound and metre should represent the actions they carry:

> 'Tis not enough no harshness gives offence,
> The sound must seem an Echo to the sense:
> Soft is the strain when Zephyr gently blows,
> And the smooth stream in smoother numbers flows;
> But when loud surges lash the sounding shoar,
> The hoarse, rough verse should like the torrent roar.
> When Ajax strives some rock's vast weight to throw,
> The line too labours, and the words move slow;
> Not so, when swift Camilla scours the plain,
> Flies o'er th' unbending corn, and skims along the main.

In 18th-century theory the doctrine of imitation was joined to numerous strictures on "smoothness," or metrical regularity. Theorists advocated a rigid regularity; minor poets composed in a strictly regular syllable-stress verse devoid of expressive variations. This regularity itself expressed the rationalism of the period. The prevailing dogmas on regularity made it impossible for Samuel Johnson to hear the beauties of Milton's versification; he characterized the metrically subtle lines of "Lycidas" as "harsh" and without concern for "numbers." Certain crosscurrents of metrical opinion in the 18th century, however, moved toward new theoretical stances. Joshua Steele's ***Prosodia Rationalis*** (1779) is an early attempt to scan English verse by means of musical notation. (A later attempt was made by the American poet Sidney Lanier in his ***Science of English Verse,*** 1880.) Steele's method is highly personal, depending on an idiosyncratic assigning of such musical qualities as pitch and duration to syllabic values; but he recognized that a prosodic theory must take into account not merely metre but "all properties or accidents belonging to language." His work foreshadows the current concerns of the structural linguists who attempt an analysis of the entire range of acoustic elements contributing to prosodic effect. Steele is also the first "timer" among metrists; that is, he bases his scansions on musical pulse and claims that English verse moves in either common or triple time. Modern critics of musical scanners have pointed out that musical scansion constitutes a performance, not an analysis of the metre, that it allows arbitrary readings, and that it levels out distinctions between poets and schools of poetry.

The 19th century. With the Romantic movement and its revolutionary shift in literary sensibility, prosodic theory became deeply influenced by early 19th-century speculation on the nature of imagination, on poetry as expression—"the spontaneous overflow of powerful feelings," in Wordsworth's famous phrase—and on the concept of the poem as organic form. The discussion between Wordsworth and Coleridge on the nature and function of metre illuminates the crucial transition from Neoclassical to modern theories. Wordsworth (in his "Preface" to the ***Lyrical Ballads,*** 1800) followed 18th-century theory and saw metre as "superadded" to poetry; its function is more nearly ornamental, a grace of style and not an essential quality. Coleridge saw metre as being organic; it functions together with all of the other parts of a poem and is not merely an echo to the sense or an artifice of style. Coleridge also examined the psychologic effects of metre, the way it sets up patterns of expectation that are either fulfilled or disappointed:

The poem as organic form

> As far as metre acts in and for itself, it tends to increase the vivacity and susceptibility both of the general feelings and of the attention. This effect it produces by the continued excitement of surprize, and by the quick reciprocations of curiosity still gratified and still re-excited, which are too slight indeed to be at any one moment objects of distinct consciousness, yet become considerable in their aggregate influence. As a medicated atmosphere, or as wine during animated conversation; they act powerfully, though themselves unnoticed. Where, therefore, correspondent food and appropriate matter are not provided for the attention and feelings thus roused, there must needs be a disappointment felt; like that of leaping in the dark from the last step of a staircase, when we had prepared our muscles for a leap of three or four.
>
> ***Biographia Literaria, XVIII (1817)***

Romantic literary theory, although vastly influential in poetic practice, had little to say about actual metrical structure. Coleridge described the subtle relationships between metre and meaning and the effects of metre on the reader's unconscious mind; he devoted little attention to metrical analysis. Two developments in 19th-century poetic techniques, however, had greater impact than any prosodic theory formulated during the period. Walt Whitman's nonmetrical prosody and Gerard Manley Hopkins' far-ranging metrical experiments mounted an assault on the traditional syllable-stress metric. Both Whitman and Hopkins were at first bitterly denounced, but, as is often the case, the heresies of a previous age become the orthodoxies of the next. Hopkins' "sprung rhythm"—a rhythm imitating natural speech, using mixed types of feet and counterpointed verse-emerged as viable techniques in the poetry of Dylan Thomas and W.H. Auden. It is virtually impossible to assess Whitman's influence on the various prosodies of modem poetry. Such American poets as Hart Crane, William Carlos Williams, and Theodore Roethke all have used Whitman's long line, extended rhythms, and "shaped" strophes.

The 20th century. Since 1900 the study of prosody has emerged as an important and respectable part of literary study. George Saintsbury published his great ***History of English Prosody*** during the years 1906–10. Sometime later, a number of linguists and aestheticians turned their attention to prosodic structure and the nature of poetic rhythm. Graphic prosody (the traditional syllable and foot scansion of syllable-stress metre) was placed on a securer theoretical footing. A number of prosodists, taking their lead from the work of Joshua Steele and Sidney Lanier, have recently attempted to use musical notation to scan English verse. For the convenience of synoptic discussion, modern prosodic theorists may be divided into four groups: the linguists who examine verse rhythm as a function of phonetic structures; the aestheticians who examine the psychologic effects, the formal properties, and the phenomenology of rhythm; the musical scanners, or "timers," who try to adapt the procedures of musical notation to metrical analysis; and the traditionalists who rely on the graphic description of syllable and stress to uncover metrical paradigms. It is necessary to point out that only the traditionalists concern themselves specifically with metrical form; aestheticians, linguists, and timers all examine prosody in its larger dimensions.

Linguists, aestheticians, timers, and traditionalists

Modern structural linguistics has placed the study of language on a scientific basis. Linguists have measured the varied intensities of syllabic stress and pitch and the durations of junctures or the pauses between syllables. These techniques of objective measurement have been applied to prosodic study. The Danish philologist Otto Jespersen's early essay "Notes on Metre" (1900) made a number of significant discoveries. He established the principles of English metre on a demonstrably accurate structural basis; he recognized metre as a gestalt phenomenon (*i.e.*, with emphasis on the configurational whole); he saw metrics as descriptive science rather than proscriptive regulation. Jespersen's essay was written before the burgeoning interest in linguistics; since World War II numerous attempts have been made to formulate a descriptive science of metrics.

It has been noted that Coleridge defined metrical form as a pattern of expectation, fulfillment, and surprise. Taking his cue from Coleridge, the British aesthetician I.A. Richards in *Principles of Literary Criticism* (1924) developed a closely reasoned theory of the mind's response to rhythm and metre. His theory is organic and contextual; the sound effects of prosody have little psychologic effect by themselves. It is prosody in conjunction with "its contemporaneous other effects^x—chiefly meaning or propositional sense—that produces its characteristic impact on our neural structures. Richards insists that everything that happens in a poem depends on the organic environment; in his *Practical Criticism* (1929) he constructed a celebrated "metrical dummy" to "support [an] argument against anyone who affirms that the mere sound of verse has *independently* any considerable aesthetic virtue." For Richards the most important function of metre is to provide aesthetic framing and control; metre makes possible, by its stimulation and release of tensions, "the most difficult and delicate utterances."

Rhythmic structure as a symbolic form

Other critics, following the Neo-Kantian theories of the philosophers Ernst Cassirer and Susanne Langer, have suggested that rhythmic structure is a species of symbolic form. Harvey Gross in *Sound and Form in Modern Poetry* (1964) saw rhythmic structure as a symbolic form, signifying ways of experiencing organic processes and the phenomena of nature. The function of prosody, in his view, is to image life in a rich and complex way. Gross's theory is also expressive; prosody articulates the movement of feeling in a poem. The unproved assumption behind Gross's expressive and symbolic theory is that rhythm is in some way iconic to human feeling: that a particular rhythm or metre symbolizes, as a map locates the features of an actual terrain, a particular kind of feeling.

The most sophisticated argument for musical scansion is given by Northrop Frye in his influential *Anatomy of Criticism* (1957). He differentiates between verse that shows unmistakable musical quality and verse written according to the imitative doctrines current in the Renaissance and Neoclassic periods. All of the poetry written in the older strong-stress metric, or poetry showing its basic structure, i3 musical poetry, and its structure resembles the music contemporary with it.

The most convincing case for traditional "graphic prosody" has been made by the American critics W.K. Wimsatt and Monroe C. Beardsley. Their essay "The Concept of Meter" (1965) argues that both the linguists and musical scanners do not analyze the abstract metrical pattern of poems but only interpret an individual performance of the poem. Poetic metre is not generated by any combination of stresses and pauses capable of precise scientific measurement; rather, metre is generated by an abstract pattern of syllables standing in positions of relative stress to each other. In a line of iambic pentameter

Preserved in Milton's or in Shakespeare's name

the "or" of the third foot is only slightly stronger than the preceding syllable "-ton's," but this very slight difference makes the line recognizable as iambic metre. Wimsatt and Beardsley underline the paradigmatic nature of metre; as an element in poetic structure, it is capable of exact abstraction.

Non-Western theories of prosody. The metres of the verse of ancient India were constructed on a quantitative basis. A system of long and short syllables, as in Greek, determined the variety of complicated metrical forms that are found in poetry of post-Vedic times—that is, after the 5th century BC.

Chinese prosody is based on the intricate tonal system of the language. In the T'ang dynasty (AD 618–907) the metrical system for classical verse was fixed. The various tones of the language were subsumed under two large groups, even tones and oblique tones. Patterned arrangements of tones and the use of pauses, or caesuras, along with rhyme determine the Chinese prosodic forms.

Japanese prosody

Japanese poetry is without rhyme or marked metrical structure; it is purely syllabic. The two main forms of syllabic verses are the tanka and the haiku. Tanka is written in a stanza of 31 syllables divided into alternating lines of five and seven syllables. Haiku is an extremely concentrated form of only 17 syllables. Longer poems of 40 to 50 lines are also written; however, alternate lines must contain either five or seven syllables. The haiku form has been adapted to English verse and has become in recent years a popular form. Other experimenters in English syllabic verse show the influence of Japanese prosody. Syllabic metre in English, however, is limited in its rhythmic effects; it is incapable of expressing the range of feeling available in the traditional stress and syllable-stress metres.

BIBLIOGRAPHY

Greek and Latin prosody: PAUL MAAS, *Greek Metre,* trans. by HUGH LLOYD-JONES (1962); ULRICH VON WILAMOWITZ-MOLLENDORFF, *Griechische Verskunst* (1921), the definitive work on the subject, but difficult for beginners.

Prose rhythm: MORRIS W. CROLL, *Style, Rhetoric, and Rhythm* (1966), contains classic essays on the period styles of prose and on musical scansion of verse; GEORGE SAINTSBURY, *A History of English Prose Rhythm* (1912, reprinted 1965).

History and uses of English prosody: WILLIAM BEARE, *Latin Verse and European Song: A Study in Accent and Rhythm* (1957), an excellent account of the transition from quantitative to accentual Latin verse and the concomitant musical influences; ROBERT BRIDGES, *Milton's Prosody,* rev. ed. (1921); HARVEY S. GROSS, *Sound and Form in Modern Poetry: A Study of Prosody from Thomas Hardy to Robert Lowell* (1964); T.S. OMOND, *English Metrists* (1921, reprinted 1968); GEORGE SAINTSBURY, *A History of English Prosody from the Twelfth Century to the Present,* 3 vol. (1906–10), a history of the subject from its beginnings to the close of the 19th century; JOHN THOMPSON, *The Founding of English Metre* (1961).

Theories of prosody: SEYMOUR B. CHATMAN, *A Theory of Meter* (1965); OTTO JESPERSEN, "Notes on Metre," in *The Selected Writings of Otto Jespersen* (1962); I.A. RICHARDS, *Principles of Literary Criticism* (1924, reprinted 1961); WILLIAM K. WIMSATT, JR., and MONROE C. BEARDSLEY, "The Concept of Metre: An Exercise in Abstraction," in WILLIAM K. WIMSATT, JR., *Hateful Contraries* (1965); YVOR WINTERS, "The Audible Reading of Poetry," in *The Function of Criticism* (1957).

Non-Western prosody: ROBERT H. BROWER and EARL MINER, *Japanese Court Poetry* (1961); JAMES LEGGE (ed. and trans.), *The Chinese Classics,* vol. 4 (1960).

General works: manuals, handbooks, and anthologies: PAUL FUSSELL, JR., *Poetic Meter and Poetic Form* (1965); HARVEY S. GROSS (ed.), *The Structure of Verse: Modern Essays on Prosody* (1966); JOSEPH MALOF, *A Manual of English Meters* (1970).

(Ha.G.)

Prostitution

Prostitution is described most precisely by emphasizing two essential elements: (1) the exchange of money or valuable materials in return for sexual activity and (2) the relatively indiscriminate availability of such a transaction to individuals other than spouses or friends. This description specifically limits the exchange to money or valuable materials; sexual activity to earn good will or subsequent favours is not properly construed as prostitution. Acceptance of money or gifts in exchange for sexual activity may be found among mistresses, gigolos, friends, and spouses; the economic criterion alone does not suffice. Similarly, sexual activity with strangers or with persons for whom there is no affectional feeling does not in itself constitute prostitution if the economic element is absent. Lastly, sexual activity denotes some physical contact. Sexual gratification solely from visual or auditory stimuli cannot ordinarily be included as prostitution; if it could, a substantial part of the entertainment world would be inadvertently included under the term.

Distinctions between prostitutes and other sexual partners

It is most expedient to adopt an operational definition: a prostitute is a person who for immediate payment in money or valuables will engage in sexual activity with any other person, known or unknown, who meets minimal requirements as to gender, age, cleanliness, sobriety, ethnic group, and health.

Even this definition does not wholly suffice, for the concept of prostitution is based on culturally determined values that differ in various societies in the world. Such a definition represents only one end of a continuum ranging from the socially accepted arrangement of marriage, wherein one male is morally and legally entitled to sexual gratification in exchange for support, to the other extreme where the arrangement is of very brief duration and involves numerous males. Borderline cases are not uncommon. A female who makes herself sexually available only to suitors who bestow gifts is a case in point: how many suitors in what span of time suffice to change her status from that of girl friend to mistress to prostitute? The opinion varies according to the attitudes and values of the culture or subculture in which the individuals live. In modern Western society the fact that a sexual relationship frequently involves an important economic element is often not faced up to: in many cases a female would not continue with a male who did not periodically give her gifts of at least some monetary value, which are most often understood as symbolic of affection or esteem rather than as payment.

CROSS-CULTURE COMPARISONS

Cross-species comparisons. Some primates use sexual behaviour for nonsexual purposes: female chimpanzees and baboons will sometimes present themselves sexually to a male in order to avoid attack or to distract the male while the female purloins his food. It is a small step from such primate behaviour to accepting coitus in exchange for food. Consequently it seems likely that the exchanging of food for coitus began in the transitional period between man and ape; and that with the subsequent development of more elaborate rules of social behaviour, the sexual restrictions of marriage, and the concept of parenthood, prostitution was eventually defined in some form and set apart as an entity to be accepted or condemned.

Historical comparisons. Prostitution has not, insofar as is known, been a cultural universal. In sexually permissive societies it is often rare because it is unnecessary, whereas in other societies it has been largely suppressed. Complete suppression seems virtually impossible in large urban centres where anonymity is easily achieved and where many persons are transients. But in a small community where secrecy is difficult and where life depends on communal cooperation, social sanctions—chiefly in the forms of ridicule, economic retribution, pressure from relatives, and ostracism—are extremely effective. In such small groups prostitution can be and has been prevented. Human societies are so labile and diverse that virtually every form of sexual behaviour, even those that are generally assumed to be socially disruptive, has somewhere at some time been regarded as either normal or, under special circumstances, permissible. This is true of prostitution. In addition to the societies in which it is absent, there are many (possibly the majority) in which it is tolerated, accepted, or encouraged. Toleration with some degree of stigma seems a common societal posture; in such societies prostitution is often the resort of disadvantaged females: slaves, captives, divorcees, widows, outcasts, and the unmarriageable. In brief it is, in part, the solution to the economic problem faced by females without husbands.

Motivations for prostitution

Acceptance or encouragement of prostitution under special conditions seems chiefly a matter of economics or religion. In a number of societies girls earn their dowries through prostitution, chiefly away from home, and return enriched and eminently marriageable. This custom occurred in some, but not many, preliterate groups in the New World, among certain ancient Mediterranean cultures such as the Lydians and Cyprians, and, within modern times, among the *ouled-naïl* of Algeria. Sometimes, in societies that ordinarily denigrate prostitution and even make some effort to preserve female virginity, it is encouraged when the economic reward is sufficiently great. Thus Marquesan parents may encourage daughters to barter coitus for valuable goods brought by sailors.

Religions have sometimes incorporated prostitution as a transitory rite to be done once or, more commonly, as a continuing religious obligation of a particular class of priestesses. An example of the former is the "Myletta" rite of ancient Babylonia, wherein every female was required to sit in the temple of the goddess Ishtar and accept coitus from the first male who threw a silver coin in her lap. Similar rites involving other goddesses are known to have existed elsewhere in the Middle East. Obligatory prostitution by certain priestesses was also a custom in this area but rare elsewhere, although examples are known in India and western Africa.

In societies of sufficient technological achievement and urbanization to warrant the label "civilizations," there has usually been an attempt to make secular prostitutes identifiable by requiring them to dress in some distinctive manner or live in a restricted area or both. Efforts of this kind were made in ancient Israel, Greece, and Rome and continued through medieval times into the 20th century. In other areas distinctive garb and locale were typical of prostitutes, but it seems that this was generally a voluntary pattern of behaviour to facilitate the clients' search rather than a regulation imposed by society. It should be noted, however, that informal social pressures have always served the same purpose as codified laws and ordinances; even in modern times prostitutes tolerated in one section of a city find it difficult to establish residence in a "decent" neighbourhood.

Early attempts at government control

Attempts to confine a group as inherently mobile as prostitutes have met with only temporary success. Prostitution always tends to fan out beyond the demarcated borders. The Oriental civilizations were no more successful than those of the West. During the T'ang dynasty in China, prostitutes (as well as merchants) were required to operate in specified areas; but in the following Sung Dynasty (AD 960–1126), cafes employing prostitutes infiltrated other districts, and combination entertainment places and brothels known as *wa-tzu* proliferated; there were over a score in the city of Hangchow alone. It should not surprise readers of modern Western newspapers to learn that prostitutes have been discovered operating from residences in upper middle class neighbourhoods.

Nevertheless, sufficient identification and localization were often obtained to make prostitutes liable to the inevitable by-products of civilization, licensing and taxation, which were imposed as early as Roman times. The degree of governmental control of prostitution has varied from nation to nation. Localization of prostitution also rendered it more susceptible to exploitation, often by criminal groups.

Historical data as to classes of prostitutes are scanty, but some hierarchy seems inescapable except in small societies. Aside from religious prostitution, prostitutes automatically graduate into classes according to their age, beauty, intelligence, and health; and society treats them accordingly. The independent courtesans who congregated around men of political power—such as the hetaerae of ancient Greece; Theodora, who subsequently married the Byzantine emperor Justinian; and certain of the Japanese geisha—were accorded considerable respect. Females of lesser wit or beauty who catered to the general public were generally denigrated and ill-treated.

The Industrial Revolution brought with it markedly increased urbanization and economic exploitation of large numbers of persons—both factors conducive to prostitution. At the same time the spread of humanist ideas, concern over public health, and changes in the status of women caused an increasing concern over prostitution, which in Western civilization was identified as a social evil and a problem. Colonization of other parts of the world spread that attitude widely.

Varieties of social control. All human societies exercise varying degrees of control over the sexual behaviour of their members, and in the more urban and technologically advanced cultures prostitution has been subjected to particular control because it embodies two major societal interests: sex and money. Before the 19th century the emphasis was upon control or, at least, the localization and identification of the prostitute. The motivation seems to have been partly fiscal (taxes, license fees, graft) and partly moral (to recognize and isolate the sinful). Such isolation served other useful ends; for example, one's favourite prostitute would be unlikely to be admitted to the social groups containing one's wife, fiancee, or employer. Life could be rigorously compartmentalized and the double standard of sexual morality maintained.

Effects of 19th-century social changes

In the 19th century, at least in Western nations, the emphasis tended to shift from control to attempted eradication. The previous laws and municipal ordinances concerned with regulation were superseded by statutes either prohibiting prostitution outright or so restricting its practice as to drive it underground. This change seems to have come about through the fusion of several social developments: the emancipation of women, the growth of humanitarian movements involved in social action, and the optimistic rationale that social problems could be cured by suitably worded laws.

The attempts at public control have, of course, varied in form and severity from nation to nation and so have the responses to these attempted controls. One natural problem is that in all countries the problem of enforcing legislation so closely related to variable moral codes of behaviour is fraught with difficulties; as soon as the law enters the sphere of morals, both its interpretation and its execution may be called in question as unjust or indeed immoral, at least by some members of the community. In Great Britain, for example, the Contagious Diseases Acts of 1864, 1866, and 1869, which introduced a measure of state regulation and inspection of prostitutes, had to be suspended in 1883 and repealed in 1886 as a result of the campaign initiated against them in 1869 by Josephine Elizabeth Butler, leader of the international movement to abolish state regulation (and therefore acceptance) of prostitution. There remained little that the law could do in relation to the prostitutes themselves except to see that the police maintained a certain standard of order.

In some countries, on the other hand, special "morals police" were instituted, with more power than the English "vice squad" and with the function, among others, of observing the registered prostitutes and watching for any who were unregistered; but the arbitrary power that these morals police could exercise caused considerable misgiving. Yet again, some countries, for instance Belgium, favoured the employment of women police to deal specifically with prostitutes. It may be stated, however, that the laws controlling prostitution and kindred offenses in the different countries of the world are very similar, once the fundamental distinction has been made between those of countries in which houses of prostitution have been abolished by law and those of countries that permit brothels and require prostitutes to be registered. (Some countries, of course, cannot be classified in this way.)

In Great Britain, legislation—apart from laws dealing with soliciting and offenses against decency in public places—is directed not at the prostitute or her client but against those third parties—brothel-keepers, procurers, pimps—who find her commerce profitable and easy to exploit. Thus while prostitution itself still is not a penal offense, a series of enactments were developed during the 19th century to prevent procuration, to close brothels (defined as premises used by at least two women for purposes of prostitution), to penalize landlords who rented their buildings for such purposes, and to prohibit a man's living on a prostitute's earnings. The Street Offenses Act of 1959, following recommendations of the Wolfenden Report (the ***Report of the Committee on Homosexual Offensesand Prostitution),*** for the first time prohibited all street solicitation and loitering by prostitutes, with increasing fines for repeated offenses. The act also raised the fines for permitting solicitation in refreshment houses and increased maximum penalties for procuring and for living on the earnings of prostitution.

Problems entailed in outlawing brothels and street-walking

The difficulties that attend the abolition of houses of prostitution in a country where they have been long established are illustrated by the course of events in France. In 1946 state registration was ended and the *maisons* de *tolérance* were closed; then a subsequent law, in the same year, established a card-index system for all prostitutes and regulated their registration for purposes of hygiene (*i.e.*, for contact tracing and the treatment of venereal disease); and a furtherenactment (1948) seemed to make prostitutes subject to examination even before their occupation was proved, the effect being practically a return to the regime of state registration. The law of 1946 had indeed resulted in the closing of the brothels in France, but little was done for the rehabilitation of their inmates, and it soon appeared that the latter were pursuing their old trade under worse conditions than before. Marthe Richard, a national heroine of World Wars I and II, who had been among the first to urge the closing of brothels in Paris and indeed was held to be primarily responsible for the enactment of the law, showed that she had changed her mind about the rightness of the measure in her book ***L'Appel des sexes*** (1951). To make a compromise between the abolitionists and those who wanted to repeal the law of 1946, it was proposed that brothels should be allowed in the neighbourhood of army camps at least.

In the United States, houses of prostitution are illegal in most states, as is procuring, and the majority of the states penalize soliciting and living on the earnings of prostitution. The federal Mann Act (the White-Slave Traffic Act, 1910) prohibits the transporting of a woman across state lines for immoral purposes; and interstate travel or transportation in aid of illegal business activities, including prostitution, was banned by anti-racketeering laws passed in 1961.

The effect of all these laws, however, has been in part to drive prostitution underground, where it can become more disruptive to society than when it is allowed to exist, to some extent, openly. Such a development is to be expected on the closing of brothels when at the same time no fundamental alteration of the social pattern is effected: the removal of the controls exercised by the brothel system clears the way for a "vice racket" in which both prostitute and client are exploited by third parties. In the United States strictly criminal organizations largely took control of prostitution, managing it on a national or regional basis.

B-girls and call girls

With the closing down or repression of the more flamboyant operations of streetwalkers and houses of prostitution, other categories emerged. One was the bar girl, or B-girl, employed to entertain men customers of a bar, tavern, or nightclub and induce them to spend money, often by offering to perform lewd or indecent acts or to engage in prostitution. The B-girl is in the tradition of the dance hall "hostess" of, for example, the cow towns and mining towns of the American West. And a new upper class among prostitutes emerged—the call girl, who may be quite selective in her choice of customers, at least limiting her favours to men who are willing and able to pay high fees. In addition to the call-girl system, massage parlours and other disguisesalso have been adopted. Overall, prostitution was decreased, though not drastically, through the loss of the efficient production-line activity of the brothel. A more serious blow to prostitution appears to be the growing increase in the number of females willing to engage in sexual intercourse without financial recompznse.

In the U.S.S.R., the bold claim was made in the 1930s and 1940s that no prostitution and no brothels or trafficking in the streets existed any longer. The legislation and administrative orders promulgated indicated that the problem had been treated very seriously. Before the Revolution, prostitution existed on a very large scale in Russia. The desire of the revolutionaries to abolish it would seem perhaps to emanate from the urge toward female emancipation and equality of the sexes. With regard to prostitution, the Soviet authorities repeatedly emphasized that the war against it should on no account degenerate into a war against prostitutes. It was maintained that the disappearance of prostitution in the U.S.S.R. was due, in effect, to the changing of the cultural patterns in Russian and Soviet society. The exact meaning of this asseveration is difficult to assess: the U.S.S.R. is a large country comprising a number of republics, each with its separate codes of law and in a varying stage of development; and in some of these republics tribal customs, in particular those connected with child marriage and with the bride price, would seem to have survived for a long time. Occasional reports of clandestine prostitution continued to appear.

Prostitution has been widespread and widely accepted in Asia, with casual sexual relations between both married and unmarried men and prostitutes regarded as a matter of course. Until recent times it was almost the only occupation open to dependent and unattached women. That prostitution was long accepted as a way of life throughout the Middle East and the Far East was held to be related to the inferior position of females. The surge of rising aspirations in the nations of Asia after World War II encompassed strong movements to improve the status of women. These found support in the United Nations in the Universal Declaration of Human Rights, in the Economic and Social Council's efforts to combat and reduce prostitution as incompatible with the dignity and worth of a human being, and in its Convention for the Suppression of the Traffic in Persons and of the Exploitation of the Prostitution of Others.

Prostitution in East Asia

In connection with its attempts to abolish prostitution as a practice imposing indignities on human beings, the UN made a number of surveys in member countries. Surveys of countries in Asia and the Far East covered Burma, Ceylon (now Sri Lanka), India, Indonesia, Japan, the Philippines, and Thailand. With the exception of Thailand, all of these nations had adopted laws before 1960 either to abolish or to prohibit prostitution. In Thailand prostitution was regulated and all prostitutes and brothels were registered. No new prostitutes were permitted a license after 1956, and no new brothels were allowed to open after 1948; the government reported it expected to abolish prostitution eventually.

Despite laws aiming to prohibit or suppress prostitution, it was found still to exist in the other countries mentioned. In Burma, Ceylon, Indonesia, and the Philippines, it was basically an urban phenomenon. In India, on the other hand, the majority of prostitutes (66.5 percent) were from rural areas. In some parts of India prostitution was still associated with girls dedicated to temples, and in some communities certain girls were dedicated to prostitution, although both forms of religious prostitution were forbidden by law. In northern India many dancing girls still became prostitutes.

In Japan licensed prostitution was abolished between 1946 and 1948. Public brothels were closed and offices were set up throughout the country to rehabilitate former prostitutes and to protect girls from becoming engaged in prostitution. Many geisha girls meanwhile have been prostitutes, although others do not participate in this profession.

In the Philippines prostitution was reported in the larger cities, particularly in Manila, where the military reservations and air bases afforded many customers.

In Indonesia prostitution was reported to be common in the villages. This was attributed largely to the fact that many men, who were permitted four wives, divorced one or more wives with little concern for what happened to them; such women often resorted to prostitution. In Cambodia brothels were permitted but were subject both to licensing and to special regulations. In Ceylon prostitution, though it was technically an offense, was not punished as such. It could be punished, however, under the Vagrants Ordinance. Private institutions, including the Salvation Army, tried to rehabilitate prostitutes and help them in securing respectable employment.

Additional surveys in Malaysia, Hong Kong, and Singapore found that measures to prevent prostitution were chiefly incumbent upon the police and immigration officials, who cooperated in checking the immigration of prostitutes and in the repatriation or deportation of prostitutes. Much preventive work was required because of the large number of Chinese refugees, 'and casework services were developed in governmental homes and shelters and by private organizations.

Except for a small group recruited from the upper classes, economic conditions apparently have been a major factor in prostitution in all Asian countries. Nevertheless, the influx of refugees into some areas, the traditionally rigid class restrictions on marriage and the exclusion of women from social life have all been important factors in promoting prostitution. In Japan and the Philippines especially, military installations have resulted in considerable demand for prostitutes despite restrictive legislation.

In mainland China, the Communist government, after it came into power in 1949, abolished prostitution and closed numerous brothels, converting some of the largest into youth centres. A program of re-education for prostitutes was announced, with additional re-education including assignment to hard labour for any backsliders.

(P.H.Ge./Ed.)

SOCIAL AND PSYCHOLOGICAL CHARACTERISTICS

The complexity of modern life, with accelerated cross-cultural diffusion and rapid transportation, makes a brief description of the exact character of prostitution in any given nation impossible. In an Asian country, for example, there may be villages sharing a few local prostitutes; inland towns with organized brothels staffed by indigenous females; and coastal cities with a potpourri of brothels and independent prostitutes of varying races, nationalities, residential stability, and status. All of these situations, as suggested in the survey above, are generally subject to rapid change depending upon political, economic, and social changes and the establishment or removal of military bases, factories, and other important sources of clients.

Types or forms of prostitution. Speaking generally, and with special regard for the Western nations, prostitution operates in three guises. First, there is the brothel—an establishment wherein the prostitutes generally reside, supervised usually by an older female who has acquired sufficient money and made enough social contacts to establish a brothel. One or more males may be on the premises to deal with unruly or homosexual clients. Brothels are quite varied, some catering to particular socioeconomic and ethnic groups, but they are generally confined to particular districts in a city or else relegated to its outskirts just beyond municipal authority. In some cities a given area may be almost wholly given over to prostitution, as was the case in Las Vegas, Nevada; the Calle Amistad of Havana, Cuba; and the Herbertstrasse of Hamburg, Germany. In provinces or nations where prostitution is illegal, brothels tend to be confined to locales where law enforcement is ineffective due to political corruption or to geographic factors such as an adjacent provincial or national border or an ocean. Since clients prefer novelty, prostitutes not infrequently move from one brothel to another in a different city. The brothel owner or supervisor takes a percentage of the prostitutes' earnings, a percentage that can range from 20 percent up to a nearly intolerable proportion.

Second, there is the "call girl," a prostitute who has her own residence (sometimes shared with another prostitute). In return for a percentage of her earnings she is notified by the head of the operation where to meet each client, generally the hotel room of the client and less often his home. The call girl is ordinarily expected to limit her clientele to persons obtained through the calling system and is discouraged from developing contacts on her own initiative since she might not share the revenue from these. The call girl has higher status and price than the average brothel inmate, who is known as a "house girl." Because there is often some screening of new clients, chiefly through recommendation by known persons, the call girl is protected to a considerable degree from undesirable customers and disguised police.

Third, there is the independent prostitute who shares her earnings with no one except hotel employees, taxi drivers, bartenders, her procurer, or others who direct customers to her. For security and companionship, pairs of independent prostitutes sometimes live together. The independent girl is more apt than other prostitutes to pursue this occupation sporadically and in conjunction with some other employment, almost always of low status and pay.

Variations on these three basic forms of prostitution are manifold. A brothel may pass as a massage parlour, a restaurant or bar may have an adjacent room, and escort services and modelling agencies may be disguises. To add to the complexity, there is some specialization among prostitutes in terms of techniques as, for example, those serving sadomasochists. Some prostitutes may largely or wholly avoid coitus: in certain lower class cafes and bars there are females who induce a customer to buy numerous drinks at inflated prices and when he has purchased enough will then lead him to some dimly lit, isolated booth or corner and masturbate him to orgasm. This practice seems commoner in western Europe than in the United States. **A** small number of females eschew physical contact entirely and thereby escape the definition of prostitution by simply engaging in obscene conversation in exchange for drinks (a percentage of the cost of which reverts to them). On occasion prostitutes may be employed to give exhibitions of sexual activity before audiences. In Latin America and parts of Europe such exhibitions are frequently a standard entertainment in certain taverns and brothels; in the United States they have become much less common and are largely confined to occasional acts for men's organizations.

Entry into prostitution

The career and life of female prostitutes. Entry into a career of prostitution may once have involved parents casting out a seduced daughter or "white slavers" exerting duress upon some hapless girl, but this is no longer true in the second half of the 20th century. Entry has become voluntary and often involves little trauma aside from transient pangs of shame or guilt. Some mentor is almost invariably involved—another girl with experience as a prostitute or a male friend who points out the financial advantages and offers to help the neophyte on her new career if she chooses it. A moment of clearly defined decision is generally lacking and the entry is a transitional phase buffered by rationalizations. A typical example would be that of a girl who is accustomed to presents from suitors and who by degrees comes to accept gifts in the form of cash and whose suitors may become more numerous as they become less known to her. Or another common mode of entry is the financial emergency that can be remedied by "being nice" to someone a friend happens to know, the price being disguised as a gift or loan. Subsequent fiscal emergencies arise and ultimately lead to an acknowledged career of prostitution. In any case, the motivation is financial. Nymphomania giving rise to prostitution is extremely rare.

Entry into prostitution generally involves entry into the subculture, and the neophyte must go through a sometimes stressful period of learning a new set of interpersonal relationships, a new system of patterned behaviour, and a new argot. More importantly, she must develop a new self-image. This change necessarily involves cutting most, if not all, ties with her former life, and the prostitute becomes more exclusively involved with and dependent upon the subculture of prostitution. The term "subculture" is appropriate because the individuals think of themselves as a group apart from the rest of society, establish their own mores, have their own distinctive linguistic terms, and share similar ideas and values.

Although one can find embittered individuals who dislike their work and despise their clients, the majority of prostitutes maintain an attitude similar to that of anyone providing services to a diverse clientele. Some clients prove troublesome and are consequently resented; others may be pleasant or unexpectedly generous and hence engender some regard or even affection; but most are regarded with objectivity and neutrality.

Psychological detachment and compartmentalization

The prostitute often simulates passion and sometimes affection for business reasons, yet in reality she is expected to preserve a rigorous attitude of psychological noninvolvement, precisely such as is required of other professionals who deal with human beings. Although the client should be sufficiently pleased to induce him to return, he should be processed efficiently, quickly, and dispassionately. Experiencing sexual arousal or orgasm is regarded not only as unprofessional but also as fatiguing and hence inefficient. Allowing affection to develop for a client is thought of as foolishly making oneself vulnerable to emotional hurt and possible fiscal exploitation.

Having severed her ties with her former life and preserving an emotional aloofness from her clients, the prostitute is prone to be starved for love. This deficiency is remedied by a lover who is frequently also a procurer, or pimp. The pimp is someone whom she can love and who she can imagine returns her love. He gives the sexual pleasure absent in her professional activity. More importantly, the pimp proves that she is needed—if only for mercenary reasons. Many prostitutes take pride in providing well for their "man," whose ostentatious affluence attests to their success. To love, to be loved, and to be needed are psychological necessities for most human beings, and the pimp provides these for the prostitute. In addition, he has other useful functions: he may find clients, protect her from mistreatment, pay her bail, safeguard her savings, and perhaps even join in her professional work if a client wishes homosexual activity or an exhibition. It is easy to see why the pimp is valued even if he is shared, as is often the case, with several other prostitutes, or even if he is sometimes exploitative or brutal.

The average prostitute usually is able to compartmentalize her life successfully: in her work she is seldom consciously sexually aroused and seldom or never reaches orgasm, but in her private life she is as responsive as, or perhaps somewhat more so than, the average housewife. The speculations as to prostitutes' being either sexually frigid or nymphomaniac are not supported by evidence. The compartmentalization may extend to sexual techniques and positions, some of these being employed solely in business relations. This compartmentalization serves a vital buffering protective function and permits the prostitute to keep a tolerable self-image and to engage in the emotional and social relations that are important to human well-being. The prostitute can say that her work is no measure of herself as a person, that it is a thing apart, simply an economic matter. Nevertheless, the defense mechanism cannot wholly negate reality, and the prostitute is to a considerable degree alienated from normal social relations and functions.

Hazards of pregnancy, disease, and criminality

Pregnancy and venereal disease are ordinarily thought of as inescapable occupational hazards for prostitutes. The matter of conception is enormously complicated by varying methods of contraception, the postulated antagonistic effects of one male's semen upon another's, chronic pelvic congestion, consequences of venereal disease, and other factors. It appears, however, that although prostitutes' high coital frequencies do not result in a high incidence of pregnancy, neither does it seem that prostitutes are particularly infertile. The majority have been impregnated by male friends or husbands, but few by clients. On the other hand, venereal disease is an indisputable occupational hazard, and the incidence of having had such a disease at some time or another is far higher among prostitutes than among the general population.

It has been thought that, through overexposure to males and through occasional but repeated unpleasant experiences, the prostitute would turn to other females for affection and sexual gratification. This speculation seems largely unfounded; although prostitutes may be more permissive toward any sexual behaviour, and although they may have had some homosexual experience as a part of their work, the incidence of extensive homosexuality does not appear to be substantially greater among prostitutes.

Because the prostitute is somewhat isolated from normal society, confined to her own subculture, and her professional life labelled as criminal in most states and nations, it is not surprising that criminality or, more commonly, affiliation with criminals often develops. Any lucrative illegal business invites the intrusion of organized crime.

This situation is not wholly a case of prostitutes' being victimized; it is to a great extent a mutually beneficial arrangement because the criminal organization can provide a greater protection from law and social action than could any individual or small group. In addition, more efficient management is often provided. The price for such benefits is, of course, loss of autonomy and having to adhere to financial agreements that are probably rather onerous and that draw severe punishment if broken.

The prostitute herself is not ordinarily criminal in a narrow sense of the word. Any business operator must maintain a good reputation in order to prosper, and hence only the most desperate or disreputable prostitutes rob or blackmail. The independent, lower class prostitute catering to transients is more likely to rob or defraud. The prostitute, however, is rather prone to be indirectly involved as an accomplice or accessory to a crime simply because her lover, pimp, or management organization is engaged in criminal activity. The police are well aware of this and sometimes attempt to coerce prostitutes into serving as informers.

Exiting from prostitution is, unlike entry, not always voluntary. Loss of physical beauty ultimately forces out of "the life" all those except a few who become administrators or supervisors. The more foresighted prostitutes exit before being compelled to do so, some via marriage (generally to someone who is fully aware of their profession), or the already married may simply become housewives precisely as a conventional woman may give up employment after some years of marriage and devote herself to the home. Others intelligent enough to have saved money may simply buy into some legitimate business. The improvident and less intelligent prostitute faces a grim future since she has neither funds nor skills to save her from eventual unemployment. Such an unfortunate generally ends on public relief or working at the most menial tasks. Actually little is known of the lives of former prostitutes because those who have adequately coped with the transition back into society are anxious to conceal their past; one tends to see the failures rather than the successes, although a few of the latter become famous.

The extent of female prostitution is difficult to measure even in those nations that require licensing, but it is no small social phenomenon in all densely populated areas. In the United States the survey data accumulated by Alfred C. Kinsey and his associates indicated that around mid-century roughly one-third of the college-educated and three-quarters of the less-educated white males eventually had had sexual intercourse with a prostitute. There is evidence, however, that in the second half of the 20th century, with the increasing permissiveness toward nonmarital sexual behaviour of women, prostitution is becoming correspondingly less important: the percentage of males having experience with prostitutes seemed to be diminishing, and the frequency of contact was markedly lower.

Male prostitution. Male prostitution has been largely disregarded in treatises on prostitution, and the public knows little of it. Heterosexual male prostitution—that is, males hired by or for females—is extremely rare. Almost any female desirous of sexual activity can, with little difficulty, find a male to oblige her, providing she does not set her standards too high. The gigolo, the male counterpart of the mistress, does not qualify as a prostitute under the definition here employed because the elements of promiscuity and immediate payment are absent.

Homosexual male prostitution, on the other hand, was common in the second half of the 20th century in large cities; and in some cities the male "hustlers" perhaps rivalled the female prostitutes in numbers. Most homosexual male prostitution was of the independent type: "callboy" systems existed but were not commonplace, and male brothels were extremely rare. Some basically heterosexual brothels, however, kept a male around for the convenience of a homosexually inclined client.

In Western civilizations male homosexual prostitution differs radically in many respects from female heterosexual prostitution. There is no pimp, large-scale organization is absent, the price is lower, and the sexual relation with the client is quite different. In female prostitution the prostitute rarely or never reaches orgasm and the client almost invariably does; in male homosexual prostitution the prostitute almost invariably reaches orgasm, but the client frequently does not. This paradox is the result of a curious mythology, which the male hustler and his client feel compelled to enact. The homosexual male ideally seeks a masculine-appearing heterosexual male, and the prostitute attempts to fit this image. Consequently the prostitute can do little or nothing for or to the homosexual client lest he betray a homosexual inclination of his own and ruin the illusion. So the prostitute plays an essentially passive role and has orgasm (this is regarded as a necessary part of the bargain), while the client must ordinarily content himself with psychological arousal, self-masturbation, and body contact. This arrangement is reinforced by the male prostitute's protective image of himself as a "real" and heterosexual man who tolerates homosexual activity solely for financial reasons. In actuality, of course, the hustler has a substantial homosexual component that is necessary or he could not achieve erection and orgasm; and many of them are predominately homosexual in orientation, though loath to admit it. One might regard this as the reverse of female prostitution: the female simulates a passion she does not feel, whereas the male prostitute conceals a passion he does feel. There is some evidence that this curious pattern of feigned indifference is gradually breaking down and that more male prostitutes are taking an active role in the sexual relation while maintaining a masculine image. In societies other than those of Western civilizations, the homosexual prostitute does not disguise his interest and is often as active as, or even more active than, the client (see also SEXUAL DEVIATIONS: Homosexuality).

Differences between male and female prostitution

The socioeconomic demand. Because prostitution is basically an economic matter, it is destined to follow the "laws" of economics. It will grow or diminish according to the number of individuals who find it difficult or impossible to obtain sexual contact without paying for it. These individuals include not only those with physical or cosmetic handicaps and those few with sexual desires too bizarre to be satisfied by most partners but also numerous normal persons who experience temporary difficulty in obtaining a sexual relationship. Travellers and military personnel are good examples of the latter group. Added to these are persons who could establish sexual relationships with nonprostitutes but do not wish to make the effort or do not wish to become emotionally involved. Finally, some individuals seek prostitutes to enjoy sexual techniques that their wives or customary partners refuse them—mouth-genital contact being the prime example.

Prostitution also is reinforced by the double standard of sexual morality and by the affiliations with drinking, dancing, and entertainment. These recreational aspects have been, and in many areas of the world have continued to be, an important component of prostitution.

This almost inescapable and reasonably steady demand for prostitutes is in many societies complemented by a socio-economic treatment of females that is conducive to prostitution. If most females are taught to be financially dependent upon males and discriminated against in employment and salary, a substantial number of them either will be economically forced toward prostitution or will turn to it as an easy and more lucrative alternative. Even in a utopian society, however, there would be some individuals whose skills, personality traits, or intelligence would limit their earnings, or whose real or imagined needs would exceed the beneficence of the state, and some of these persons would prostitute.

It appears impossible to eradicate prostitution in a complex society, particularly one in which sexual gratification is made difficult by mores and law. Moreover, the trend of legalistic and humanitarian thinking has been toward the idea that what consenting adults do sexually in private should not be subject to law, and that view presumably would include prostitution.

BIBLIOGRAPHY. The two major comprehensive works are PAUL LACROIX, ***Histoire de la prostitution*** . . . (1861; Eng. trans., ***History of Prostitution Among All the Peoples of the***

World from the Most Remote Antiquity to the Present Day, 3 vol. 1926); and L. FERNANDO HENRIQUES, *Prostitution and Society:* vol. *1, Primitive, Classical and Oriental* (1962); vol. 2, *Prostitution in Europe and the New World* (1963); and vol. 3, *Modern Sexuality* (1968). Perhaps the best single volume on the subject is that of HARRY BENJAMIN and R.E. MASTERS, *Prostitution and Morality: A Definitive Report* (1964). A broad sociological survey is represented by the BRITISH SOCIAL BIOLOGY COUNCIL, *Women of the Streets: A Sociological Study of the Common Prostitute,* ed. by C.H. ROLPH (1955); while a higher status variety of prostitution is dealt with by HAROLD GREENWALD in *The Call Girl: A Social and Psychoanalytic Study* (1958). The legal aspects of prostitution are the subject of THOMAS E. JAMES, *Prostitution and the Law* (1951). Important data derived from interviews with prostitutes are presented in three journal articles: JAMES H. BRYAN, "Apprenticeships in Prostitution," *Social Problems,* 12:287–297 (1965); WARDELL B. POMEROY, "Some Aspects of Prostitution," *Journal* of *Sex Research,* 1:177–187 (1965); and PAUL H. GEBHARD, "Misconceptions About Female Prostitutes," *Medical Aspects of Human Sexuality,* 3:24–30 (1969). Lastly, the extent to which males in the United States relied upon prostitution is detailed at various points in *Sexual Behavior in the Human Male,* by ALFRED C. KINSEY *et al.* (1948).

(P.H.Ge.)

Protein

Proteins are highly complex substances that are present in all living organisms. They are of great nutritional value, have numerous industrial uses, and are directly involved in the chemical processes necessary to maintain life. Proteins are both species-specific and organ-specific; for instance, muscle proteins differ from those of the brain and liver.

A protein molecule is very large compared to molecules of sugar or salt and consists of many amino acids joined together to form long chains, much as beads are arranged on a string. There are about 20 different amino acids that occur naturally in proteins. Proteins of similar function have similar amino acid composition and sequence. Although it is not yet possible to explain all of the functions of a protein from its amino acid sequence, established correlations between structure and function can be attributed to the properties of the amino acids of which proteins are composed.

Amino acid production and requirements by plants and animals

Plants can synthesize all of the amino acids; animals cannot, even though all of them are essential for life. Plants can grow in a medium containing inorganic nutrients that provide nitrogen, potassium, and other substances essential for growth. They utilize the carbon dioxide in the air during the process of photosynthesis to form organic compounds such as carbohydrates. Animals, however, must obtain organic nutrients from outside sources. Because the protein content of most plants is low, very large amounts of plant material are required by animals, such as ruminants (*e.g.*, cows), that eat only plant material to meet their amino acid requirements. Nonruminant animals, including man, obtain proteins principally from animals and their products—*e.g.*, meat, milk, and eggs. The seeds of legumes are increasingly being used to prepare inexpensive protein-rich food (see NUTRITION AND DIET, HUMAN).

The protein content of animal organs is usually much higher than that of the blood plasma. Muscles, for example, contain about 30 percent protein, the liver 20 to 30 percent, and red blood cells 30 percent. Higher percentages of protein are found in hair, bones, and other organs and tissues with a low water content. The quantity of free amino acids and peptides in animals is much smaller than the amount of protein. Evidently, protein molecules are produced in cells by the stepwise alignment of amino acids and are released into the body fluids only after synthesis is complete.

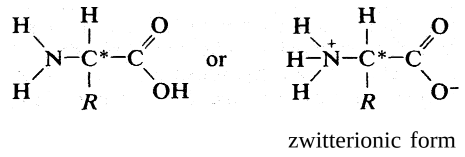
The high protein content of some organs does not mean that the importance of proteins is related to their amount in an organism or tissue; some of the most important proteins, such as enzymes and hormones, occur in extremely small amounts. The importance of proteins is related principally to their function. All enzymes identified thus far are proteins. Enzymes, which are the catalysts of all metabolic reactions, enable an organism to build up the chemical substances necessary for life—proteins, nucleic acids, carbohydrates, and lipids (fats)—to convert them into other substances, and to degrade them. Life without enzymes is not possible. There are several protein hormones. Hormones are regulatory substances formed in endocrine glands and secreted directly into the bloodstream. In all vertebrates, the respiratory protein hemoglobin acts as oxygen carrier in the blood, transporting oxygen from the lung to body organs and tissues. A large group of structural proteins maintains and protects the structure of the animal body.

This article emphasizes the basic structural and functional aspects of proteins as a class of chemical substances. For further information on specialized subjects in which proteins play a major role see BLOOD AND LYMPH; ENZYME; HORMONE; IMMUNITY; GENE.

I. General structure and properties of proteins

THE AMINO ACID COMPOSITION OF PROTEINS

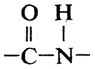
The common property of all proteins is that they consist of long chains of a-amino (alpha amino) acids. The general structure of a-amino acids is shown in Formula 1.



Formula 1: Generalized structure of all a-amino acids. C stands for a carbon atom; C* stands for the α-carbon. H is hydrogen, O is oxygen, and N is nitrogen. R is a general term for any of several different chemical structures that range from one hydrogen atom to large and complex molecular units containing many different atoms. The + and − signs represent electrical charges that exist when the molecule takes the configuration shown at the right.

The a-amino acids are so called because the a-carbon atom in the molecule (shown by an asterisk [*] in Formula 1) carries an amino group $(—NH_2)$; the a-carbon atom also carries a carboxyl group $(-COOH)$. In acidic solutions, when the pH is less than 4, the $-COO-$ groups combine with hydrogen ions (H^+) and are thus converted into the uncharged form $(-COOH)$. In alkaline solutions, at pH above 9, the ammonium groups $(-NH^+_3)$ lose a hydrogen ion and are converted into amino groups $(-NH_2)$. In the pH range between 4 and 8, the amino acids exist almost exclusively in the structure shown at the right side of Formula 1. Because in this form they carry both a positive and a negative charge, they do not migrate in an electrical field. Such structures have been designated as dipolar ions, or zwitterions (*i.e.*, hybrid ions).

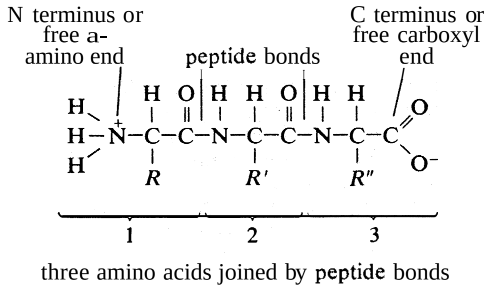
Although more than 100 amino acids occur in nature, particularly in plants, only 20 types are commonly found in most proteins (see Figure 1). In protein molecules the a-amino acids are linked to each other by peptide bonds between the amino group of one amino acid and the carboxyl group of its neighbour; the structure of the peptide bond is given in Formula 2. The condensation (joining) of three amino acids yields the tripeptide shown in Formula 3.



Formula 2: The peptide bond.

Peptide structure

It is customary to write the structure of peptides in such a way that the free a-amino group (also called the N terminus of the peptide) is at the left side and the free carboxyl group (the C terminus) at the right side. Proteins are macromolecular polypeptides—*i.e.*, very large molecules composed of many peptide-bonded amino acids. Most of the common ones contain more than 100 amino acids linked to each other in a long peptide chain. The average molecular weight (based on the weight of a hydrogen atom as 1) of each amino acid is approximately


N terminus or free a-amino end
peptide bonds
C terminus or free carboxyl end
1
2
3
three amino acids joined by peptide bonds

Formula 3: A tripeptide. *R′* and *R″* represent the possibility that the three *R* groups (side chains) could be different.

100 to 125; thus, the molecular weights of proteins are usually in the range of 10,000 to 100,000 daltons (one dalton is the weight of one hydrogen atom). The species-specificity and organ-specificity of proteins result from differences in the number and sequences of amino acids. Twenty different amino acids in a chain 100 amino acids long can be arranged in far more than 10^{100} ways (10^{100} is the number one followed by 100 zeroes).

Structures of common amino acids. The amino acids present in proteins differ from each other in the structure of their side (R) chains. The simplest amino acid is glycine, in which R is a hydrogen atom (see Figure 1). In a number of amino acids, R represents straight or branched carbon chains. One of these amino acids is alanine, in which R is the methyl group ($—CH_3$). Valine, leucine, and isoleucine, with longer R groups, complete the alkyl side-chain series. The alkyl side chains (R groups) of these amino acids are nonpolar; this means that they have no affinity for water but some affinity for each other. Although plants can form all of the alkyl amino acids, animals can synthesize only alanine and glycine; thus valine, leucine, and isoleucine must be supplied in the diet.

Two amino acids, each containing three carbon atoms, are derived from alanine; they are serine and cysteine. Serine contains an alcohol group ($—CH_2OH$) instead of the methyl group of alanine, and cysteine contains a mercapto group ($—CH_2SH$). Animals can synthesize serine but not cysteine nor cystine. Cysteine occurs in proteins predominantly in its oxidized form (oxidation in this sense meaning the removal of hydrogen atoms), called cystine. Cystine consists of two cysteine molecules linked by the disulfide bond (—S—S—) that results when a hydrogen atom is removed from the mercapto group of each of the cysteines (see Figure 1). Disulfide bonds are important in protein structure because they allow the linkage of two different parts of a protein molecule to—and thus the formation of loops in—the otherwise straight chains. Some proteins contain small amounts of cysteine with free sulfhydryl (—SH) groups.

Four amino acids, each consisting of four carbon atoms, occur in proteins; they are aspartic acid, asparagine, threonine, and methionine. Aspartic acid and asparagine, which occur in large amounts, can be synthesized by animals. Threonine and methionine cannot be synthesized and thus are essential amino acids—*i.e.*, they must be supplied in the diet. Most proteins contain only small amounts of methionine.

Most abundant amino acids

Proteins also contain an amino acid with five carbon atoms (glutamic acid) and an imino acid (proline), which is a structure with the amino group ($—NH_2$) bonded to the alkyl side chain, forming a ring. Glutamic acid and aspartic acid are dicarboxylic acids--that is, they have two carboxyl groups (—COOH). Glutamine is similar to asparagine in that both are the amides of their corresponding dicarboxylic acid forms; *i.e.*, they have an amide group ($—CONH_2$) in place of the carboxyl (—COOH) of the side chain (see Figure 1). Glutamic acid and glutamine are abundant in most proteins; in plant proteins, for example, they sometimes comprise more than one third of the amino acids present. Both glutamic acid and glutamine can be synthesized by animals. The imino acids proline and hydroxyproline occur in large amounts in collagen, the protein of the connective tissue of animals (see Table 1). Proline and hydroxyproline lack free amino ($—NH_2$) groups because the amino group is enclosed in a ring structure with the side chain; they thus cannot exist in a zwitterion form. Although the imino group ($>NH$) of these amino acids can form a peptide bond with the carboxyl group of another amino acid, the bond so formed gives rise to a kink in the otherwise straight peptide chain—that is, the imino ring structure alters the regular bond angle of normal peptide bonds.

Proteins usually are almost neutral molecules; that is, they have neither acidic nor basic properties. This means that the acidic carboxyl (—COO-) groups of aspartic and glutamic acid are about equal in number to the amino acids with basic side chains. Three such basic amino acids, each containing six carbon atoms, occur in proteins. The one with the simplest structure, lysine, is synthesized by plants but not by animals. Even some plants have a low lysine content. Arginine is found in all proteins; it occurs in particularly high amounts in the strongly basic protamines (simple proteins composed of relatively few amino acids) of fish sperm. The third basic amino acid is histidine. Both arginine and histidine can be synthesized by animals. Histidine is a weaker base than either lysine or arginine. The imidazole ring, a five-membered ring structure containing two nitrogen atoms in the side chain of histidine (see Figure 1), acts as a buffer (*i.e.*, a stabilizer of hydrogen ion concentration) by binding hydrogen ions (H^+) to the nitrogen atoms of the imidazole ring.

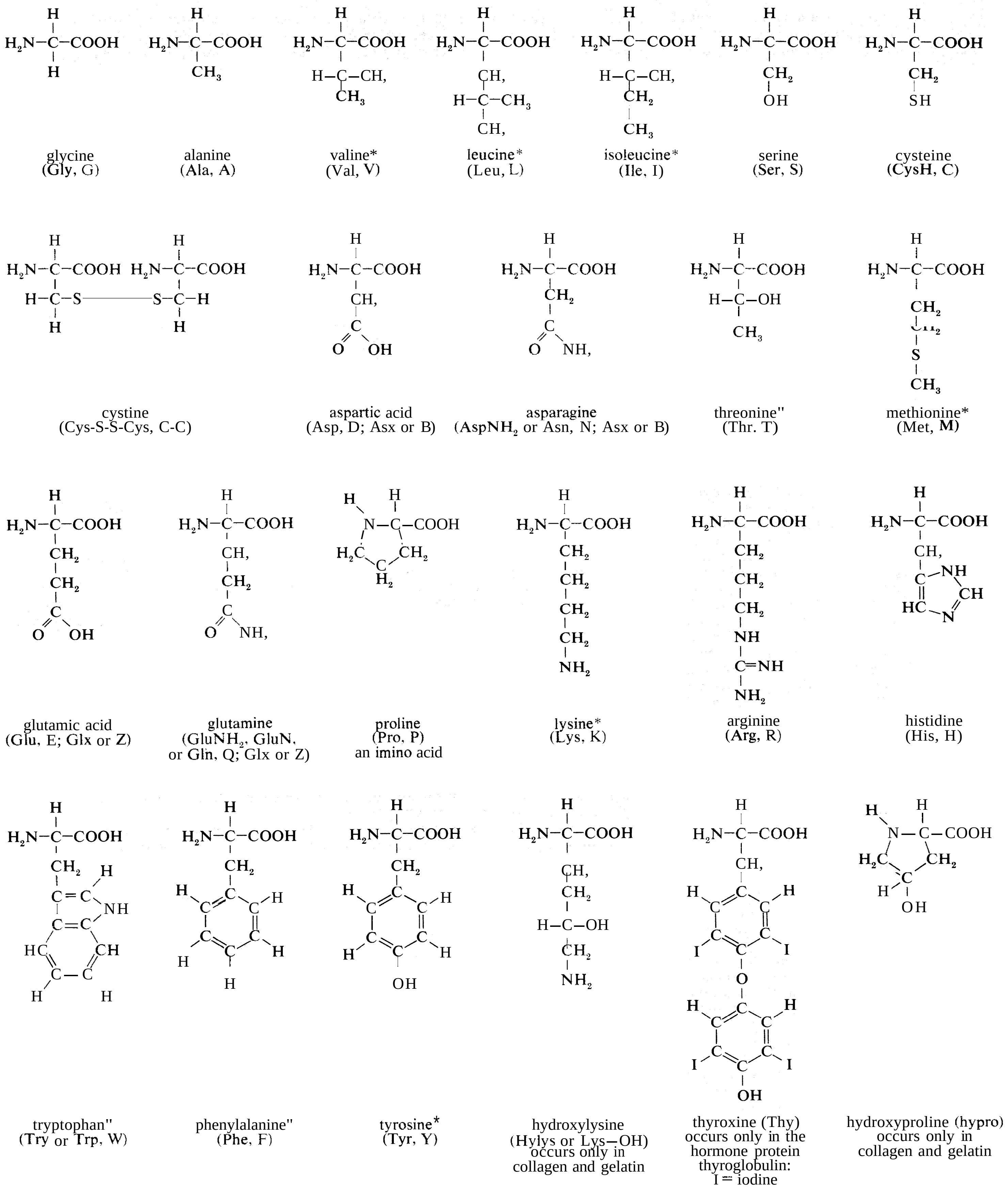
The remaining amino acids—phenylalanine, tyrosine, and tryptophan—have in common an aromatic structure; *i.e.*, a benzene ring is present (see Figure 1). Animals cannot synthesize the benzene ring; therefore these three amino acids are essential ones. Animals can convert phenylalanine to tyrosine, however. Because these three amino acids contain benzene rings, they can absorb ultraviolet light at wavelengths between 270 and 290 nanometres (nm; one nanometre $= 10^{-9}$ metre $=$ ten angstrom units). Phenylalanine absorbs very little ultraviolet light; tyrosine and tryptophan, however, absorb it strongly and are responsible for the absorption band most proteins exhibit at 280–290 nanometres. This absorption is frequently used to determine the quantity of protein present in protein samples.

Most proteins contain only the amino acids described above; however, other amino acids occur in proteins in small amounts. Thyroglobulin, the hormone of the thyroid gland, for example, contains thyroxine, which is an iodine-containing compound derived from tyrosine. The collagen found in connective tissue contains, in addition to hydroxyproline, small amounts of hydroxylysine. Other proteins contain some monomethyl- , dimethyl-, or trimethyllysine—*i.e.*, lysine derivatives containing one, two, or three methyl groups (—CH,). The amount of these unusual amino acids in proteins, however, rarely exceeds 1 or 2 percent of the total amino acids.

Physicochemical properties of the amino acids. The physicochemical properties of a protein are determined by the analogous properties of the amino acids in it.

The α-carbon atom of all amino acids, with the exception of glycine, is asymmetric; this means that four different chemical entities (atoms or groups of atoms) are attached to it. As a result, each of the amino acids, except glycine, can exist in two different spatial, or geometric, arrangements (*i.e.*, isomers), which are mirror images akin to right and left hands (see Formula 4). These isomers exhibit the property of optical rotation.

Optical rotation is the rotation of the plane of polarized light, which is composed of light waves that vibrate in one plane, or direction, only. Solutions of substances that rotate the plane of polarization are said to be optically active, and the degree of rotation is called the optical rotation of the solution. The direction in which the light is rotated is generally designed as plus, or d, for dextrorotatory (to the right), or as minus, or *l*, for levorotatory (to the left). Some amino acids are dextrorotatory; others are levorotatory. With the exception of a few small proteins (peptides) that occur in



glycine (Gly, G)

alanine (Ala, A)

valine* (Val, V)

leucine* (Leu, L)

isoleucine* (Ile, I)

serine (Ser, S)

cysteine (CysH, C)

cystine (Cys-S-S-Cys, C-C)

aspartic acid (Asp, D; Asx or B)

asparagine (AspNH$_2$ or Asn, N; Asx or B)

threonine" (Thr. T)

methionine* (Met, M)

glutamic acid (Glu, E; Glx or Z)

glutamine (GluNH$_2$, GluN, or Gln, Q; Glx or Z)

proline (Pro, P) an imino acid

lysine* (Lys, K)

arginine (Arg, R)

histidine (His, H)

tryptophan" (Try or Trp, W)

phenylalanine" (Phe, F)

tyrosine* (Tyr, Y)

hydroxylysine (Hylys or Lys—OH) occurs only in collagen and gelatin

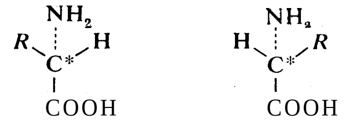
thyroxine (Thy) occurs only in the hormone protein thyroglobulin: I = iodine

hydroxyproline (hypro) occurs only in collagen and gelatin

Figure 1: Structures of amino acids found in proteins. Those amino acids marked with an asterisk (*) must be supplied in the diet of animals, which cannot synthesize them. The abbreviations in parentheses represent the shorthand notations(in three-letter codes and one-letter codes) used when indicating protein structures. The one-letter symbol for an unknown amino acid is X

Configuration of amino acids

bacteria, the amino acids that occur in proteins have the configuration shown on the left of Formula 4. For this reason all the amino acids found in proteins are designed as L-amino acids.



L-amino acid D-amino acid

Formula 4: The tetrahedral (four-faced) arrangement of the bonds around the a-carbon (C*). The solid lines represent bonds that slant upward from the plane of the drawing (*i.e.*, toward the reader). The broken lines represent bonds that recede from the plane of the drawing (*i.e.*, away from the reader).

In bacteria, D-alanine and some other D-amino acids have been found as components of gramicidin and bacitracin. These peptides are toxic to other bacteria and are used in medicine as antibiotics. The D-alanine has also been found in some peptides of bacterial membranes.

In contrast to most organic acids and amines, the amino acids are insoluble in organic solvents. In aqueous solutions they are dipolar ions (zwitterions, or hybrid ions) that react with strong acids or bases in a way that leads to the neutralization of the negatively or positively charged ends, respectively. Because of their reactions with strong acids and strong bases, the amino acids act as buffers—stabilizers of hydrogen ion (H^+) or hydroxide ion (OH–) concentrations. In fact, glycine is frequently used as a buffer in the pH range from 1 to 3 (acid solutions) and from 9 to 12 (basic solutions). In acid solutions, glycine has a positive charge and therefore migrates to the cathode (negative electrode of a direct-current electrical circuit with terminals in the solution). Its charge, however, is negative in alkaline solutions, in which it migrates to the anode (positive electrode). At pH 6.1 glycine does not migrate, because each molecule has one positive and one negative charge. The pH at which an amino acid does not migrate in an electrical field is called the isoelectric point. Most of the monoamino acids (*i.e.*, those with only one amino group) have isoelectric points similar to that of glycine. The isoelectric points of aspartic and glutamic acids, however, are close to pH 3; and those of histidine, lysine, and arginine are at pH 7.6, 9.7, and 10.8, respectively.

Amino acid sequence in protein molecules. Since each protein molecule consists of a long chain of amino acid residues, linked to each other by peptide bonds, the hydrolytic cleavage of all peptide bonds is a prerequisite for the quantitative determination of the amino acid residues. Hydrolysis is most frequently accomplished by boiling the protein with concentrated hydrochloric acid. The quantitative determination of the amino acids is based on the discovery that amino acids can be separated from each other by chromatography on filter paper and made visible by spraying the paper with ninhydrin. The amino acids of the protein hydrolysate are separated from each other by passing the hydrolysate through a column of adsorbents which adsorb the amino acids with different affinities and, on washing the column with buffer solutions, release them in a definite order. The amount of each of the amino acids can be determined by the intensity of the colour reaction with ninhydrin.

To obtain information about the sequence of the amino acid residues in the protein, the protein is degraded stepwise, one amino acid being split off in each step. This is accomplished by coupling the free a-amino group (—NH,) of the N-terminal amino acid with phenyl isothiocyanate; subsequent mild hydrolysis does not affect the peptide bonds; the procedure, called the Edman degradation, can be applied repeatedly; it thus reveals the sequence of the amino acids in the peptide chain.

Unavoidable small losses that occur during each step make it impossible to determine the sequence of more than about 30 to 50 amino acids by this procedure. For this reason the protein is usually first hydrolyzed by exposure to the enzyme trypsin (see below, Catalytic ***proteins:*** Enzymes), which cleaves only peptide bonds formed by the carboxyl groups of lysine and arginine. The Edman degradation is then applied to each of the few resulting peptides produced by the action of trypsin. Further information can be gained by hydrolyzing another portion of the protein with another enzyme, for instance with chymotrypsin, which splits predominantly peptide bonds formed by the amino acids tyrosine, phenylalanine, and tryptophan. The combination of results obtained with two or more different proteolytic (protein degrading) enzymes was first applied by the English biochemist Frederick Sanger, and it enabled him to elucidate the amino acid sequence of insulin. The amino acid sequences shown in formulas 7 to 11 and those of many other proteins have been determined in this manner.

LEVELS OF STRUCTURAL ORGANIZATION IN PROTEINS

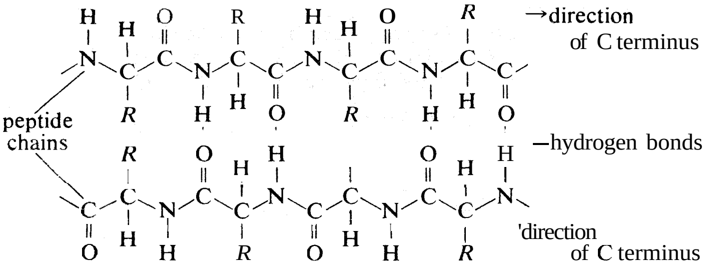
Primary structure. Analytical and synthetic procedures reveal only the primary structure of the proteins—that is, the amino acid sequence of the peptide chains. They do not reveal information about the conformation (arrangement in space) of the peptide chain—that is, whether the peptide chain is present as a long straight thread or is irregularly coiled and folded into a globule. The configuration, or conformation, of a protein is determined by mutual attraction or repulsion of polar or nonpolar groups in the side chains (R groups) of the amino acids. The former have positive or negative charges in their side chains; the latter repel water but attract each other. Some parts of a peptide chain containing 100 to 200 amino acids may form a loop, or helix; others may be straight or form irregular coils.

Definitions of protein structural terms

The terms secondary, tertiary, and quaternary structure are frequently applied to the configuration of the peptide chain of a protein. A nomenclature committee of the International Union of Biochemistry (IUB) has defined these terms as follows: The primary structure of a protein is determined by its amino acid sequence without any regard for the arrangement of the peptide chain in space. The secondary structure is determined by the spatial arrangement of the main peptide chain without any regard for the conformation of side chains or other segments of the main chain. The tertiary structure is determined by both the side chains and other adjacent segments of the main chain, without regard for neighbouring peptide chains. Finally, the term quaternary structure is used for the arrangement of identical or different subunits of a large protein in which each subunit is a separate peptide chain.

Secondary structure. The nitrogen and carbon atoms of a peptide chain cannot lie on a straight line because of the magnitude of the bond angles between adjacent atoms of the chain; the bond angle is about 110". Each of the nitrogen and carbon atoms can rotate to a certain extent, however, so that the chain has a limited flexibility. Because all of the amino acids, except glycine, are asymmetric L-amino acids, the peptide chain tends to assume an asymmetric helical shape; some of the fibrous proteins consist of elongated helices around a straight screw axis. Such structural features result from properties common to all peptide chains. The product of their effects is the secondary structure of the protein.

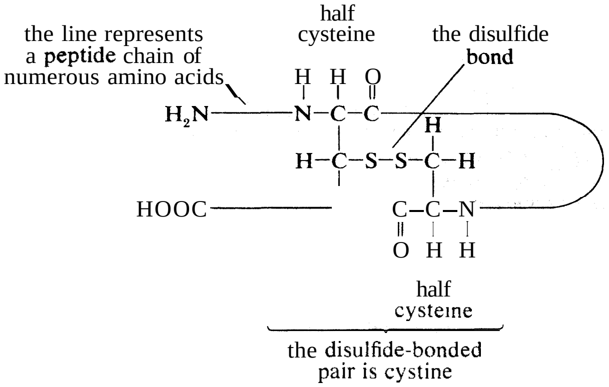
Tertiary structure. The tertiary structure is the product of the interaction between the side chains (R) of the amino acids comprising the protein. Some of them contain positively or negatively charged groups, others are polar, and still others are nonpolar. The number of carbon atoms in the side chain varies from zero in glycine to nine in tryptophan (see Figure 1). Positively and negatively charged side chains have the tendency to attract each other; side chains with identical charges repel each other. The bonds formed by the forces between the negatively charged side chains of aspartic or glutamic acid on the one hand, and the positively charged side chains of lysine or arginine on the other hand, are called salt bridges. Mutual attraction of adjacent peptide chains also results from the formation of numerous hydrogen bonds. They are shown by dotted lines in the diagram of an antiparallel pleated sheet protein structure (see Formula 5). Hydrogen bonds form as a result of the attraction between the nitrogen-bound hydrogen atom (the imide hydrogen) and the unshared pair of electrons of the oxygen atom in the double bonded carbon-oxygen group (the carbonyl group) ($>C{=}O$). The result is a slight displacement of the imide hydrogen toward the oxygen atom of the carbonyl group. Although the hydrogen bond is much weaker than a covalent bond (*i.e.*, the type of bond between two carbon atoms, which equally share the pair of bonding electrons between them), the large number of imide and carbonyl groups in peptide chains results in the formation of numerous hydrogen bonds. Another type of attraction is that between nonpolar side chains of valine, leucine, isoleucine, and phenylalanine; the attraction results in the displacement of water molecules and is called hydrophobic interaction.



Formula 5: The antiparallel pleated sheet structure.

Hydrophobic interaction defined

In proteins rich in cystine, the conformation of the peptide chain is determined to a considerable extent by the disulfide bonds (—S—S—) of cystine. The halves of cystine may be located in different parts of the peptide chain and thus may form a loop closed by the disulfide bond, as shown in Formula 6. If the disulfide bond is reduced (*i.e.*, hydrogen is added) to two sulfhydryl (—SH) groups, the tertiary structure of the protein undergoes a drastic change--closed loops are broken and adjacent disulfide-bonded peptide chains separate.



Formula 6: The disulfide bridge between two cystine halves in an amino acid chain showing how loops in the chain are formed by this amino acid.

Quaternary structure. The nature of the quaternary structure is demonstrated by the structure of hemoglobin. Each molecule of human hemoglobin consists of four peptide chains, two a-chains and two β-chains; *i.e.*, it is a tetramer. The four subunits are linked to each other by hydrogen bonds and hydrophobic interaction. Because the four subunits are so closely linked, the hemoglobin tetramer is called a molecule, even though no covalent bonds occur between the peptide chains of the four subunits. In other proteins, the subunits are bound to each other by covalent bonds (disulfide bridges; see below the structure of insulin in Formula 8).

THE ISOLATION AND DETERMINATION OF PROTEINS

Animal material usually contains large amounts of protein and lipids (fats) and small amounts of carbohydrate; in plants, the bulk of the dry matter is usually carbohydrate. No general method exists for the isolation of proteins from organs or tissues. If it is necessary to determine the amount of protein in a mixture of animal foodstuffs, a sample is converted to ammonium salts by boiling with sulfuric acid and a suitable inorganic catalyst, such as copper sulfate (Kjeldahl method). The method is based on the assumption that proteins contain 16 percent nitrogen, and that nonprotein nitrogen is present in very small amounts. The assumption is justified for most tissues from higher animals but not for insects and crustaceans, in which a considerable portion of the body nitrogen is present in the form of chitin, a carbohydrate. Large amounts of nonprotein nitrogen are also found in the sap of many plants. In such cases, the precise quantitative analyses are made after the proteins have been separated from other biological compounds.

Concentration of protein molecules by dialysis

Proteins are sensitive to heat, acids, bases, organic solvents, and radiation exposure; for this reason, the chemical methods employed to purify organic compounds cannot be applied to proteins. Salts and molecules of small size are removed from protein solutions by dialysis; *i.e.*, by placing the solution into a sac of semipermeable material, such as cellulose or acetylcellulose, which will allow small molecules to pass through but not large protein molecules, and immersing the sac in water or a salt solution. Small molecules can also be removed either by passing the protein solution through a column of resin that adsorbs only the protein or by gel filtration. In gel filtration, the large protein molecules pass through the column, and the small molecules are adsorbed to the gel.

Groups of proteins are separated from each other by salting out—*i.e.*, the stepwise addition of sodium sulfate or ammonium sulfate to a protein solution. Some proteins, called globulins, become insoluble and precipitate when the solution is half-saturated with ammonium sulfate or when its sodium sulfate content exceeds about 12 percent. Other proteins, the albumins, can be precipitated from the supernatant solution (*i.e.*, the solution remaining after a precipitation has taken place) by saturation with ammonium sulfate. Water-soluble proteins can be obtained in a dry state by freeze-drying (lyophilization), in which the protein solution is deep-frozen by lowering the temperature below —15" C (5" F) and removing the water; the protein is obtained as a dry powder.

Most proteins are insoluble in boiling water and are denatured by it—*i.e.*, irreversibly converted into an insoluble material. Heat denaturation cannot be used with connective tissue because the principal structural protein, collagen, is converted by boiling water into water-soluble gelatin.

Fractionation (separation into components) of a mixture of proteins of different molecular weight can be accomplished by gel filtration. The size of the proteins retained by the gel depends upon the properties of the gel. The proteins retained in the gel are removed from the column by solutions of a suitable concentration of salts and hydrogen ions.

Criteria of purity

Although many proteins were originally obtained in crystalline form, crystallinity is not proof of purity; in fact, many crystalline protein preparations contain other substances. Various tests are used to determine whether a protein preparation contains only one protein. The purity of a protein solution can be determined by such techniques as chromatography and gel filtration. In addition, a solution of pure protein will yield one peak when spun in a centrifuge at very high speeds (ultracentrifugation) and will migrate as a single band in electrophoresis (migration of the protein in an electrical field). Only after all of these methods (and others, such as amino acid analysis) indicate that the protein solution is pure, can it be considered so. Because chromatography, ultracentrifugation, and electrophoresis cannot be applied to insoluble proteins, very little is known about them; they may be mixtures of many similar proteins.

Very small (microheterogeneous) differences in some of the apparently pure proteins are known to occur; they are differences in the amino acid composition of otherwise identical proteins and are transmitted from generation to generation; *i.e.*, they are genetically determined; for example, some humans have two hemoglobins, hemoglobin A and hemoglobin S, which differ in one amino acid at a specific site in the molecule. In hemoglobin A the site is occupied by glutamic acid, and in hemoglobin S by valine. Refinement of the techniques of protein analysis has resulted in the discovery of other instances of "microheterogeneity."

The quantity of a pure protein can be determined by weighing or by measuring the ultraviolet absorbancy at 280 nanometres. The absorbency at 280 nanometres depends on the content of tyrosine and tryptophan in the protein (see above The amino acid composition of proteins). Sometimes the slightly less sensitive biuret reaction, a purple colour given by alkaline protein solutions upon the addition of copper sulfate, is used; its intensity depends only on the number of peptide bonds per gram, which is similar in all proteins.

PHYSICOCHEMICAL PROPERTIES OF PROTEINS

The molecular weight of proteins. The molecular weight of proteins cannot be determined by the methods of classical chemistry (*e.g.*, freezing-point depression) because they require solutions of a higher concentration of protein than can be prepared.

If a protein contains only one molecule of one of the amino acids or one atom of iron, copper, or another element, the minimum molecular weight of the protein or a subunit can be calculated; for example, the protein myoglobin contains 0.34 gram of iron in 100 grams of protein. The atomic weight of iron is 56; thus the minimum molecular weight of myoglobin is $(56 \times 100)/0.34 =$ about 16,500. Direct measurements of the molecular weight of myoglobin yield the same value. The molecular weight of hemoglobin, however, which also contains 0.34 percent iron, has been found to be 66,000 or $4 \times 16{,}500$; thus hemoglobin contains four atoms of iron.

Determination of molecular weight by ultracentrifugation

The method most frequently used to determine the molecular weight of proteins is ultracentrifugation; *i.e.*, spinning in a centrifuge at velocities up to about 60,000 revolutions per minute. Centrifugal forces of more than 200,000 times the gravitational force on the surface of the Earth are achieved at such velocities. The first ultracentrifuges, built in 1920, were used to determine the molecular weight of proteins. The molecular weights of a large number of proteins have been determined. Most consist of several subunits, the molecular weight of which is usually less than 100,000 and frequently ranges from 20,000 to 30,000. Proteins of very high molecular weights are found among hemocyanins, the copper-containing respiratory proteins of invertebrates; some range as high as several million. Although there is no definite lower limit for the molecular weight of proteins, short amino acid sequences are usually called peptides.

The shape of protein molecules. In the technique of X-ray diffraction the X-rays are allowed to strike a protein crystal; the X-rays, diffracted (bent) by the crystal, impinge on a photographic plate, forming a pattern of spots. This method reveals that peptide chains can assume very complicated, apparently irregular shapes. Two extremes in shape include the closely folded structure of the globular proteins and the elongated, unidimensional structure of the threadlike fibrous proteins; both were recognized many years before the technique of X-ray diffraction was developed. Solutions of fibrous proteins are extremely viscous (*i.e.*, sticky); those of the globular proteins have low viscosity (*i.e.*, they flow easily). A 5 percent solution of a globular protein—ovalbumin, for example—easily flows through a narrow glass tube; a 5 percent solution of gelatin, a fibrous protein, however, does not flow through the tube because it is liquid only at high temperatures and solidifies at room temperature. Even solutions containing only 1 or 2 percent of gelatin are highly viscous and flow through a narrow tube either very slowly or only under pressure. The elongated peptide chains of the fibrous proteins can be imagined to become entangled not only mechanically but also by mutual attraction of their side chains; in this way they incorporate large amounts of water. Most of the hydrophilic (water-attracting) groups of the globular proteins, however, lie on the surface of the molecules; as a result, globular proteins incorporate only a few water molecules. If a solution of a fibrous protein flows through a narrow tube, the elongated molecules become oriented parallel to the direction of the flow (see Figure 2), and the solution thus becomes birefringent like a crystal; *i.e.*, it splits a light ray into two components that travel at different velocities and are polarized at right angles to each other. Globular proteins do not show this phenomenon, which is called flow birefringence. Solutions of myosin, the contractile protein of muscles, show very high flow birefringence; other proteins with very high flow birefringence include solutions of fibrinogen, the clotting material of blood plasma, and solutions of tobacco mosaic virus. The gamma-globulins of the blood plasma show low flow birefringence; and none can be observed in solutions of serum albumin and ovalbumin.

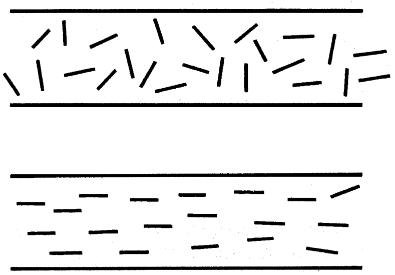


Figure 2: Flow birefringence. The upper diagram shows a solution containing elongated, rodlike macromolecules that, in the resting solution, are randomly oriented. The lower diagram shows the same solution during flow through a horizontal tube.

Hydration of proteins. When dry proteins are exposed to air of high water content, they rapidly bind water up to a maximum quantity, which differs for different proteins; usually it is 10 to 20 percent of the weight of the protein. The hydrophilic groups of a protein are chiefly the positively charged groups in the side chains of lysine and arginine and the negatively charged groups of aspartic and glutamic acid. Hydration (*i.e.*, the binding of water) may also occur at the hydroxyl ($-OH$) groups of serine and threonine or at the amide ($-CONH_2$) groups of asparagine and glutamine.

Polar structure of water molecule

The binding of water molecules to either charged or polar (partly charged) groups is explained by the dipolar structure of the water molecule; that is, the two positively charged hydrogen atoms form an angle of about 105°, with the negatively charged oxygen atom at the apex. The centre of the positive charges is located between the two hydrogen atoms; the centre of the negative charge of the oxygen atom is at the apex of the angle. The negative pole of the dipolar water molecule binds to positively charged groups; the positive pole binds negatively charged ones. The negative pole of the water molecule also binds to the hydroxyl and amino groups of the protein.

The water of hydration is essential to the structure of protein crystals; when they are completely dehydrated, the crystalline structure disintegrates. In some proteins this process is accompanied by denaturation and loss of the biological function.

In aqueous solutions, proteins bind some of the water molecules very firmly; others are either very loosely bound or form islands of water molecules between loops of folded peptide chains. Because the water molecules in such an island are thought to be oriented as in ice, which is crystalline water, the islands of water in proteins are called icebergs. Water molecules may also form bridges between the carbonyl ($>C=O$) and imino ($>NH$) groups of adjacent peptide chains, resulting in structures similar to those of the pleated sheet (see Formula 5) but with a water molecule in the position of the hydrogen bonds of that configuration. The extent of hydration of protein molecules in aqueous solutions is important, because some of the methods used to determine the molecular weight of proteins yield the molecular weight of the hydrated protein. The amount of water bound to one gram of a globular protein in solution varies from 0.2 to 0.5 gram. Much larger amounts of water are mechanically immobilized between the elongated peptide chains of fibrous proteins; for example, one gram of gelatin can immobilize at room temperature 25 to 30 grams of water.

The salting-out Process

Hydration of proteins is necessary for their solubility in water. If the water of hydration of a protein dissolved in water is reduced by the addition of a salt such as ammonium sulfate, the protein is no longer soluble and is salted out, or precipitated. The salting-out process is reversible because the protein is not denatured (*i.e.*, irreversibly converted to an insoluble material) by the addition of such salts as sodium chloride, sodium sulfate, or ammonium sulfate. Some globulins, called euglobulins, are insoluble in water in the absence of salts; their insolubility is attributed to the mutual interaction of polar groups on the surface of adjacent molecules, a process that results in the formation of large aggregates of molecules. Addition of small amounts of salt causes the euglobulins to become soluble. This process, called salting in, results from a combination between anions (negatively charged ions) and cations (positively charged ions) of the salt and positively and negatively charged side chains of the euglobulins. The combination prevents the aggregation of euglobulin molecules by preventing the formation of salt bridges between them. The addition of more sodium or ammonium sulfate causes the euglobulins to salt out again and to precipitate.

Electrochemistry of proteins. Because the α-amino group and a-carboxyl group of amino acids are converted into peptide bonds (see Formula 2) in the protein molecule, there is only one a-amino group (at the N terminus) and one a-carboxyl group (at the C terminus). The electrochemical character of a protein is affected very little by these two groups. Of importance, however, are the numerous positively charged ammonium groups ($-NH_3^+$) of lysine and arginine and the negatively charged carboxyl groups (—COO-) of aspartic acid and glutamic acid. In most proteins, the number of positively and negatively charged groups varies from 10 to 20 per 100 amino acids.

Electrometric titration. When measured volumes of hydrochloric acid are added to a solution of protein in salt-free water, the pH decreases in proportion to the amount of hydrogen ions added until it is about 4. Further addition of acid causes much less decrease in pH because the protein acts as a buffer at pH values of 3 to 4. The reaction that takes place in this pH range is the protonation of the carboxyl group—*i.e.*, the conversion of —COO- into —COOH. Electrometric titration of an isoelectric protein with potassium hydroxide causes a very slow increase in pH and a weak buffering action of the protein at pH 7; a very strong buffering action occurs in the pH range from 9 to 10 (see Figure 3). The buffering action at pH 7, which is caused by loss of protons (positively charged hydrogen) from the imidazolium groups (*i.e.*, the five-member ring structure in the side chain; see Figure 1) of histidine, is weak because the histidine content of proteins is usually low. The much stronger buffering action at pH values from 9 to 10 is caused by the loss of protons from the hydroxyl group of tyrosine and from the ammonium groups of lysine. Finally, protons are lost from the guanidinium groups (*i.e.*, the nitrogen-containing terminal portion of the arginine side chains; see Figure 1) of arginine at pH 12. A curve of the electrometric titration of glycine is shown in Figure 3. Electrometric titrations of proteins yield similar curves. Electrometric titration makes possible the determination of the approximate number of carboxyl groups, ammonium groups, histidines, and tyrosines per molecule of protein.

Electrophoresis. The positively and negatively charged side chains of proteins cause them to behave like amino acids in an electrical field; that is, they migrate during electrophoresis at low pH values to the cathode (negative terminal) and at high pH values to the anode (positive terminal). The isoelectric point, the pH value at which the protein molecule does not migrate, is in the range of pH 5 to 7 for many proteins. Proteins such as lysozyme, cytochrome *c*, histone, and others rich in lysine and arginine (see Table 2), however, have isoelectric points in the pH range between 8 and 10. The isoelectric point of pepsin, which contains very few basic amino acids, is close to 1.

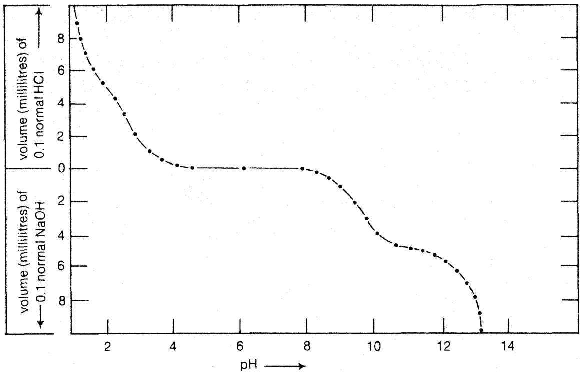


Figure 3: Electrometric titration of glycine. Addition of measured quantities of known-concentration hvdrochloric acid (HCl) to glycine is shown in the upper half of'the diagram; sodium' hydroxide (NaOH) in the lower half. The dots indicate the experimental results. The addition of a trace of acid or base to pure glycine, the isoelectric point of which is close to pH 6.1, causes a strong change in pH. Glycine acts as a buffer, however, in the acidic pH range below 3 and in the alkaline pH range above 8.

Zone electrophoresis

Free-boundary electrophoresis, the original method of determining electrophoretic migration, has been replaced in many instances by zone electrophoresis, in which the protein is placed in either a gel of starch, agar, or polyacrylamide or in a porous medium such as paper or cellulose acetate. The migration of hemoglobin and other coloured proteins can be followed visually. Colourless proteins are made visible after the completion of electrophoresis by staining them with a suitable dye.

CONFORMATION OF GLOBULAR PROTEINS

Results of X-ray diffraction studies. Most knowledge concerning secondary and tertiary structure of globular proteins has been obtained by the examination of their crystals using X-ray diffraction. In this technique X-rays are allowed to strike the crystal; the X-rays are diffracted by the crystal and impinge on a photographic plate, forming a pattern of spots. The measured intensity of the diffraction pattern, as recorded on a photographic film, depends particularly on the electron density of the atoms in the protein crystal. This density is lowest in hydrogen atoms, and they do not give a visible diffraction pattern. Although carbon, oxygen, and nitrogen atoms yield visible diffraction patterns, they are present in such great number—about 700 or 800 per 100 amino acids—that the resolution of the structure of a protein containing more than 100 amino acids is almost impossible. Resolution is considerably improved by substituting into the side chains of certain amino acids very heavy atoms, particularly those of heavy metals. Mercury ions, for example, bind to the sulfhydryl (—SH) groups of cysteine. Platinum chloride has been used in other proteins. In the iron-containing proteins, the iron atom already in the molecule is adequate.

Although the X-ray diffraction technique cannot resolve the complete three-dimensional conformation (that is, the secondary and tertiary structure of the peptide chain), complete resolution has been obtained by combination of the results of X-ray diffraction with those of amino acid sequence analysis. In this way the complete conformation of such proteins as myoglobin, chymotrypsinogen, lysozyme, and ribonuclease has been resolved.

The X-ray diffraction method has revealed regular structural arrangements in proteins; one is an extended form of antiparallel peptide chains that are linked to each other by hydrogen bonds between the carbonyl ($>C=O$) and imino ($>NH$) groups (shown in Formula 5). This conformation, called the pleated sheet, or β-structure, is found in some fibrous proteins. Short strands of the β-structure have also been detected in some globular proteins.

The a-helix

A second important structural arrangement is the α-helix (see Figure 4); it is formed by a sequence of amino acids wound around a straight axis in either a right-handed or a left-handed spiral. Each turn of the helix corresponds to a distance of 5.4 angstroms (= 0.54 nanometre) in the direction of the screw axis and contains 3.7 amino acids. Hence, the length of the a-helix per amino acid residue is 5.4 divided by 3.7, or 1.5 angstroms (one angstrom = 0.1 nanometre = 10^{-10} metre). The stability of the a-helix is maintained by hydrogen bonds between the carbonyl and imino groups of neighbouring turns of the helix. It was once thought, based on data from analyses of the myoglobin molecule, more than half of which consists of a-helices, that the a-helix is the predominant structural element of the globular proteins; it is now known that myoglobin is exceptional in this respect. The other globular proteins for which the structures have been resolved by X-ray diffraction contain only small regions of a-helix. In most of them the peptide chains are folded in an apparently random fashion (see Figure 5), for which the term random coil has been used. The term is misleading, however, because the folding is not random; rather, it is dictated by the primary structure and modified by the secondary and tertiary structures.

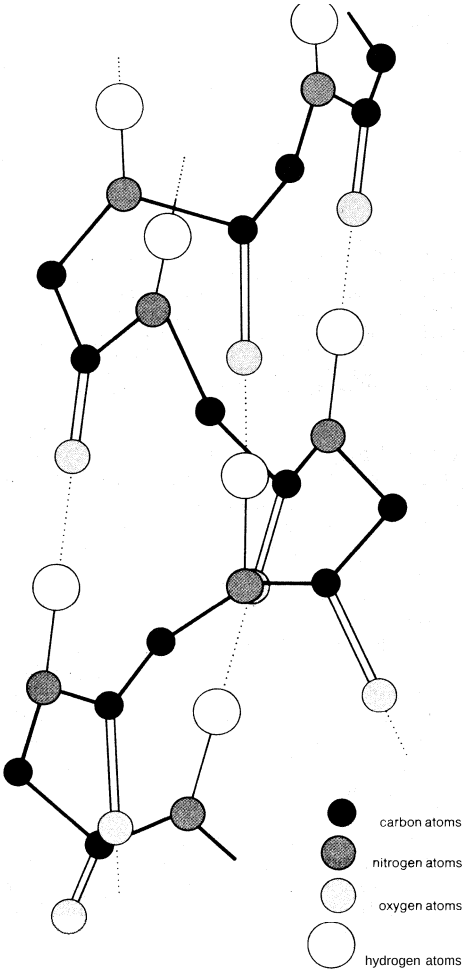


Figure 4: *The a-helix* (see *text).*

The first proteins for which the internal structures were completely resolved are the iron-containing proteins myoglobin and hemoglobin. The investigation of the hydrated crystals of these proteins at Cambridge by Max Perutz and J.C. Kendrew, who won a Nobel Prize for their work, revealed that the folding of the peptide chains is so tight that most of the water is displaced from the centre of the globular molecules. The amino acids that carry the ammonium ($—NH_3^+$) and carboxyl (—COO-) groups were found to be shifted to the surface of the globular molecules, and the nonpolar amino acids were found to be concentrated in the interior.

Other approaches to the determination of protein structure. None of the several other physical methods that have been used to obtain information on the secondary and tertiary structure of proteins provides as much direct information as the X-ray diffraction technique. Most of the techniques, however, are much more simple than X-ray diffraction, which requires, for the resolution of the structure of one protein, many years of work and equipment such as electronic computers. Some of the simpler techniques are based on the optical properties of proteins—refractivity, absorption of light of different wavelengths, rotation of the plane polarized light at different wavelengths, and luminescence.

Spectrophotometric behaviour. Spectrophotometry of protein solutions (the measurement of the degree of absorbance of light by a protein within a specified wavelength) is useful within the range of visible light only with proteins that contain coloured prosthetic groups (the nonprotein components). Examples of such proteins include the red heme proteins of the blood, the purple pigments of the retina of the eye, green and yellow proteins that contain bile pigments, blue copper-containing proteins, and dark brown proteins called melanins. Peptide bonds, because of their carbonyl groups, absorb light energy at very short wavelengths (185–200 nanometres). The aromatic rings of phenylalanine, tyrosine, and tryptophan (see Figure 1), however, absorb ultraviolet light between wavelengths of 280 and 290 nanometres. The absorbance of ultraviolet light by tryptophan is greatest, that of tyrosine is less, and that of phenylalanine is least. If the tyrosine or tryptophan content of the protein is known, therefore, the concentration of the protein solution can be determined by measuring its absorbance between 280 and 290 nanometres.

Optical activity. It will be recalled that the amino acids, with the exception of glycine, exhibit optical activity (rotation of the plane of polarized light; see above, *Physicochemical properties of the amino acids*). It is not surprising, therefore, that proteins also are optically active. They are usually levorotatory (*i.e.*, they rotate the plane of polarization to the left) when polarized light of wavelengths in the visible range is used. Although the specific rotation (a function of the concentration of a protein solution and the distance the light travels in it) of most L-amino acids varies from —30" to +30°, the amino acid cystine has a specific rotation of approximately —300". Although the optical rotation of a protein depends on all of the amino acids, the most important ones are cystine and the aromatic amino acids phenylalanine, tyrosine, and tryptophan. The contribution of the other amino acids to the optical activity of a protein is negligibly small.

Polarized light rotation ranges among proteins

Chemical reactivity *of* proteins. Information on the internal structure of proteins can be obtained with chemical methods that reveal whether certain groups are present on the surface of the protein molecule and thus able to react or whether they are buried inside the closely folded peptide chains and thus are unable to react. The chemical reagents used in such investigations must be mild ones that do not affect the structure of the protein.

The reactivity of tyrosine is of special interest. It has been found, for example, that only three of the six tyrosines found in the naturally occurring enzyme ribonuclease can be iodinated (*i.e.*, reacted to accept an iodine atom). Enzyme-catalyzed breakdown of iodinated ribonuclease is used to identify the peptides in which the iodinated tyrosines are present. The three tyrosines that can be iodinated lie on the surface of ribonuclease; the others, assumed to be inaccessible, are said to be buried in the molecule. Tyrosine can also be identified by using other techniques; *e.g.*, treatment with diazonium compounds or tetranitromethane. Because the compounds formed are coloured, they can easily be detected when the protein is broken down with enzymes.

Cysteine can be detected by coupling with compounds such as iodoacetic acid or iodoacetamide; the reaction results in the formation of carboxymethylcysteine or carbamidomethylcysteine, which can be detected by amino acid determination of the peptides containing them. The imidazole groups of certain histidines can also be located by coupling with the same reagents under different conditions. Unfortunately, few other amino acids can be labelled without changes in the secondary and tertiary structure of the protein.

Association of protein subunits. Many proteins with molecular weights of more than 50,000 occur in aqueous solutions as complexes: dimers, tetramers, and higher polymers—*i.e.*, as chains of two, four, or more repeating basic structural units. The subunits, which are called monomers or protomers, usually are present as an even number. Less than 10 percent of the polymers have been found to have an odd number of monomers. The arrangement of the subunits is thought to be regular and may be cyclic, cubic, or tetrahedral. Some of the small proteins also contain subunits. Insulin, for example, with a molecular weight of about 6,000, consists of two peptide chains linked to each other by disulfide bridges (—S—S—). Similar interchain disulfide bonds have been found in the immunoglobulins. In other proteins, hydrogen bonds and hydrophobic bonds (resulting from the interaction between the amino acid side chains of valine, leucine, isoleucine, and phenylalanine) cause the formation of aggregates of the subunits. The subunits of some proteins are identical; those of others differ. Hemoglobin is a tetramer consisting of two a-chains and two p-chains.

Protein denaturation. When a solution of a protein is boiled, the protein frequently becomes insoluble—*i.e.*, it is denatured—and remains insoluble even when the solution is cooled. The denaturation of the proteins of egg white by heat—as when boiling an egg—is an example of irreversible denaturation. The denatured protein has the same primary structure as the original, or native, protein. The weak forces between charged groups and the weaker forces of mutual attraction of nonpolar groups are disrupted at elevated temperatures, however; as a result, the tertiary structure of the protein is lost. In some instances the original structure of the protein can be regenerated; the process is called renaturation.

Methods of denaturation

Denaturation can be brought about in various ways. Proteins are denatured by treatment with alkaline or acid, oxidizing or reducing agents, and certain organic solvents. Interesting among denaturing agents are those that affect the secondary and tertiary structure without affecting the primary structure. The agents most frequently used for this purpose are urea and guanidinium chloride. These molecules, because of their high affinity for peptide bonds, break the hydrogen bonds and the salt bridges between positive and negative side chains, thereby abolishing the tertiary structure of the peptide chain. When denaturing agents are removed from a protein solution, the native protein re-forms in many cases. Denaturation can also be accomplished by reduction of the disulfide bonds of cystine—*i.e.*, conversion of the disulfide bond (—S—S—) to two sulfhydryl groups (—SH). This, of course, results in the formation of two cysteines. Reoxidation of the cysteines by exposure to air sometimes regenerates the native protein. In other cases, however, the wrong cysteines become bound to each other, resulting in a different protein. Finally, denaturation can also be accomplished by exposing proteins to organic solvents such as ethanol or acetone. It is believed that the organic solvents interfere with the mutual attraction of nonpolar groups.

Some of the smaller proteins, however, are extremely stable, even against heat; for example, solutions of ribonuclease can be exposed for short periods of time to temperatures of 90° C (194" F) without undergoing significant denaturation. Denaturation does not involve identical changes in protein molecules; a common property of denatured proteins, however, is the loss of biological activity—*e.g.*, the ability to act as enzymes or hormones.

Although denaturation had long been considered an all-or-none reaction, it is now thought that many intermediary states exist between native and denatured protein. In some instances, however, the breaking of a key bond could be followed by the complete breakdown of the conformation of the native protein.

Although many native proteins are resistant to the action of the enzyme trypsin, which breaks down proteins during digestion, they are hydrolyzed by the same enzyme after denaturation. Evidently, the peptide bonds that can be split by trypsin are inaccessible in the native proteins but become accessible during denaturation. Similarly, denatured proteins give more intense colour reactions for tyrosine, histidine, and arginine than do the same proteins in the native state. The increased accessibility of reactive groups of denatured proteins is attributed to an unfolding of the peptide chains.

Unfolded peptide chains in denatured proteins

If denaturation can be brought about easily and if renaturation is difficult, how is the native conformation of globular proteins maintained in living organisms, in which they are produced stepwise, by incorporation of one amino acid at a time? Experiments on the biosynthesis of proteins from amino acids containing radioactive carbon or heavy hydrogen reveal that the protein molecule grows stepwise from the N terminus to the C terminus; in each step a single amino acid residue is incorporated. As soon as the growing peptide chain contains six or seven amino acid residues, the side chains interact with each other and thus cause deviations from the straight or p-chain configuration shown in Formula 3. Depending on the nature of the side chains, this may result in the formation of an a-helix (Figure 4) or of loops closed by hydrogen bonds (Formula 5) or disulfide bridges (Formula 6). The final conformation is probably frozen when the peptide chain attains a length of 50 or more amino acid residues.

Conformation of proteins in interfaces. Like many other substances with both hydrophilic and hydrophobic groups, soluble proteins tend to migrate into the interface between air and water or oil and water; the term oil here means a hydrophobic liquid such as benzene or xylene. Within the interface, proteins spread, forming thin films. Measurements of the surface tension, or interfacial tension, of such films indicate that tension is reduced by the protein film. Proteins, when forming an interfacial film, are present as a monomolecular layer; *i.e.*, a layer one molecule in height. Although it was once thought that globular protein molecules unfold completely in the interface, it has now been established that many proteins can be recovered from films in the native state. The application of lateral pressure on a protein film causes it to increase in thickness and finally to form a layer with a height corresponding to the diameter of the native protein molecule. Protein molecules in an interface, because of Brownian motions (molecular vibrations), occupy much more space than do those in the film after the application of pressure. The Brownian motion of compressed molecules is limited to the two dimensions of the interface, since the protein molecules cannot move upward or downward.

The motion of protein molecules at the air–water interface has been used to determine the molecular weight of proteins. The technique involves measuring the force exerted by the protein layer on a barrier.

When a protein solution is vigorously shaken in air, it forms a foam, because the soluble proteins migrate into the air–water interface and persist there, preventing or slowing the reconversibn of the foam into a homogeneous solution. Some of the unstable, easily modified proteins are denatured when spread in the air–water interface. The formation of a permanent foam when egg white is vigorously stirred is an example of irreversible denaturation by spreading in a surface.

CLASSIFICATION OF PROTEINS

Classification by solubility. After two famous German chemists Emil Fischer and Franz Hofmeister independently stated in 1902 that proteins are essentially polypeptides consisting of many amino acids, an attempt was made to classify proteins according to their chemical and physical properties, because the biological function of proteins had not yet been established. (The protein char-

Methods of classification

acter of enzymes was not proved until the 1920s.) Proteins were classified primarily according to their solubility in a number of solvents. This classification is no longer satisfactory, however, because proteins of quite different structure and function sometimes have similar solubilities; conversely, proteins of the same function and similar structure sometimes have different solubilities. The terms associated with the old classification, however, are still widely used. They are defined below.

Albumins are proteins that are soluble in water and in water half-saturated with ammonium sulfate. On the other hand, globulins are salted out (*i.e.*, precipitated) by half-saturation with ammonium sulfate. Globulins that are soluble in salt-free water are called pseudoglobulins; those insoluble in salt-free water are euglobulins. Both prolamins and glutelins, which are plant proteins, are insoluble in water; the prolamins dissolve in 50 to 80 percent ethanol, the glutelins in acidified or alkaline solution. The term protamine is used for a number of proteins in fish sperm that consist of approximately 80 percent arginine and therefore are strongly alkaline. Histones, which are less alkaline, apparently occur only in cell nuclei, where they are bound to nucleic acids. The term scleroproteins has been used for the insoluble proteins of animal organs. They include keratin, the insoluble protein of certain epithelial tissues such as the skin or hair, and collagen, the protein of the connective tissue. A large group of proteins has been called conjugated proteins, becanse they are complex molecules of protein consisting of protein and nonprotein moieties. The nonprotein portion is called the prosthetic group. Conjugated proteins can be subdivided into mucoproteins, which, in addition to protein, contain carbohydrate; lipoproteins, which contain lipids (fats); phosphoproteins, which are rich in phosphate; chromoproteins, which contain pigments such as iron-porphyrins, carotenoids, bile pigments, and melanin; and finally, nucleoproteins, which contain nucleic acid.

The weakness of the above classification lies in the fact that many, if not all, globulins contain small amounts of carbohydrate; thus there is no sharp borderline between globulins and mucoproteins. Moreover, the phosphoproteins do not have a prosthetic group that can be isolated; they are merely proteins in which some of the hydroxyl groups of serine are phosphorylated (*i.e.*, contain phosphate). Finally, the globulins include proteins with quite different roles—enzymes, antibodies, fibrous proteins, and contractile proteins.

Classification by biological functions. In view of the unsatisfactory state of the old classification, it is preferable to classify the proteins according to their biological function. Such a classification is far from ideal, however, because one protein can have more than one function. The contractile protein myosin, for example, also acts as an ATPase (adenosine triphosphatase), an enzyme that hydrolyzes adenosine triphosphate (removes a phosphate group from ATP by introducing a water molecule). In addition, the definite function of a protein frequently is not known. A protein cannot be called an enzyme as long as its substrate (the specific compound upon which it acts) is not known. It cannot even be tested for its enzymatic action when its substrate is not known.

II. Special structure and function of proteins

Despite its weaknesses, a functional classification is used here in order to demonstrate, whenever possible, the correlation between the structure and function of a protein. The structural, fibrous proteins are presented first, because their structure is simpler than that of the globular proteins and more clearly related to their function, which is the maintenance of either a rigid or a flexible structure.

STRUCTURAL PROTEINS

Scleroproteins. Collagen. Collagen is the structural protein of bones, tendons, ligaments, and skin. For many years collagen was considered to be insoluble in water. Part of the collagen of calf skin, however, can be extracted with citrate buffer at pH 3.7. A precursor of collagen called procollagen is converted in the body into collagen.

Structure of collagen

Procollagen has a molecular weight of 120,000. Cleavage of one or a few peptide bonds of procollagen yields collagen, which has three subunits, each with a molecular weight of 95,000; therefore, the molecular weight of collagen is 285,000 (3 × 95,000). The three subunits are wound as spirals around an elongated straight axis. The length of each subunit is 2,900 angstroms, and its diameter is approximately 15 angstroms. The three chains are staggered, so that the trimer has no definite terminal limits.

The amino acid composition of collagen is shown in Table 1. It differs from all other proteins in its high content of proline and hydroxyproline. Hydroxyproline does not occur in significant amounts in any other protein except elastin. Most of the proline in collagen is present in the sequence glycine–proline-*X*, in which X is frequently alanine or hydroxyproline. Collagen does not contain cystine or tryptophan and therefore cannot substitute for other proteins in the diet. The presence of proline causes kinks in the peptide chain and thus reduces the length of the amino acid unit from 3.7 angstroms in the extended chain of the β-structure to 2.86 angstroms in the collagen chain. In the intertwined triple helix, the glycines are inside, close to the axis; the prolines are outside.

Table 1: Amino Acid Content of Some Proteins

amino acid*	protein					
	α-casein	gliadin	edestin	collagen (ox hide)	keratin (wool)	myosin
Lysine	60.9	4.45	19.9	27.4	6.2	85
Histidine	18.7	11.7	18.6	4.5	19.7	15
Arginine	24.7	15.7	99.2	47.1	56.9	41
Aspartic acid†	63.1	10.1	99.4	51.9	51.5	85
Threonine	41.2	17.6	31.2	19.3	55.9	41
Serine	63.1	46.7	55.7	41.0	79.5	41
Glutamic acid†	153.1	311.0	144.9	76.2	99.0	155
Proline	71.3	117.8	32.9	125.2	58.3	22
Glycine	37.3	—	68.0	354.6	78.0	39
Alanine	41.5	23.9	57.7	115.7	43.8	78
Half cystine	3.6	21.3	10.9	0.0	105.0	86
Valine	53.8	22.7	54.6	21.4	46.6	42
Methionine	16.8	11.3	16.4	6.5	4.0	22
Isoleucine	48.8	90.8‡	41.9	14.5	29.0	42
Leucine	60.3		60.0	28.2	59.9	79
Tyrosine	44.7	17.7	26.9	5.5	28.7	18
Phenylalanine	27.9	39.0	38.4	13.9	22.4	27
Tryptophan	7.8	3.2	6.6	0.0	9.6	—
Hydroxyproline	0.0	0.0	0.0	97.5	12.2	—
Hydroxylysine	—	—	—	8.0	1.2	—
Total	839	765	883	1,058	863	832
Average residual weight	119	131	113	95	117	120

'Number of amino acids is given per 100,000 daltons of protein—*i.e.*, the number of gram molecules of amino acid per 100,000 grams of protein.
†The values for aspartic and glutamic acid include asparagine and glutamine, respectively. ‡Isoleucine plus leucine.

Native collagen resists the action of trypsin but is hydrolyzed by the bacterial enzyme collagenase. When collagen is boiled with water, the triple helix is destroyed, and the subunits are partially hydrolyzed; the product is gelatin. The unfolded peptide chains of gelatin trap large amounts of water, resulting in a hydrated molecule.

When collagen is treated with tannic acid or with chromium salts, cross links form between the collagen fibres, and it becomes insoluble; the conversion of hide into leather is based on this tanning process. The tanned material is insoluble in hot water and cannot be converted to gelatin. On exposure to water at 62" to 63° C (144" to 145" F), however, the cross links formed by the tanning agents collapse, and the leather contracts irreversibly to about one-third its original volume.

Collagen seems to undergo an aging process in living organisms that may be caused by the formation of cross links between collagen fibres. They are formed by the conversion of some lysine side chains to aldehydes (compounds with the general structure RCHO), and the combination of the aldehydes with the ε-amino groups of intact lysine side chains. The protein elastin, which occurs in the elastic fibres of connective tissue, contains similar cross links and may result from the combination of collagen fibres with other proteins. When cross-linked collagen or elastin is degraded, products of the cross-linked lysine fragments, called desmosins and isodesmosins, are formed.

Keratin. Keratin, the structural protein of epithelial cells in the outermost layers of the skin, has been isolated from hair, nails, hoofs, and feathers. Keratin is completely insoluble in cold or hot water; it is not attacked by proteolytic enzymes (*i.e.*, enzymes that break apart, or lyse, protein molecules), and therefore cannot replace proteins in the diet. The great stability of keratin results from the numerous disulfide bonds of cystine. The amino acid composition of keratin differs from that of collagen (see Table 1). Cystine may account for 24 percent of the total amino acids. The peptide chains of keratin are arranged in approximately equal amounts of antiparallel and parallel pleated sheets, in which the peptide chains are linked to each other by hydrogen bonds between the carbonyl ($>C{=}O$) and imino ($>NH$) groups.

Pleated-sheet structure of keratin

Reduction of the disulfide bonds to sulfhydryl groups results in dissociation of the peptide chains, the molecular weight of which is 25,000 to 28,000 each. The formation of permanent waves in the beauty treatment of hair is based on partial reduction of the disulfide bonds of hair keratin by thioglycol, or some other mild reducing agent, and subsequent oxidation of the sulfhydryl groups (—SH) in the reoriented hair to disulfide bonds (—S—S—) by exposure to the oxygen of the air.

The length of keratin fibres depends on their water content. They can bind approximately 16 percent of water; this hydration is accompanied by an increase in the length of the fibres of 10 to 12 percent.

The most thoroughly investigated keratin is hair keratin, particularly that of wool. It consists of a mixture of peptides with high and low cystine content. When wool is heated in water to about 90" C (190" F), it shrinks irreversibly. This is attributed to the breakage of hydrogen bonds and other noncovalent bonds; disulfide bonds do not seem to be affected.

Others. The most thoroughly investigated scleroprotein has been fibroin, the insoluble material of silk. The raw silk comprising the cocoon of the silkworm consists of two proteins. One, sericin, is soluble in hot water; the other, fibroin, is not. The amino acid composition of the latter differs from that of all other proteins. It contains large amounts of glycine, alanine, tyrosine, and serine; small amounts of the other amino acids; and no sulfur-containing ones. The peptide chains are arranged in antiparallel β-structures. Fibroin is partly soluble in concentrated solutions of lithium thiocyanate or in mixtures of cupric salts and ethylene diamine. Such solutions contain a protein of molecular weight 170,000, which is a dimer of two subunits.

Little is known about either the scleroproteins of the marine sponges or the insoluble proteins of the cellular membranes of animal cells. Some of the membranes are soluble in detergents; the membrane of the red blood cells contains an insoluble membrane protein that consists of a single peptide chain of molecular weight 200,000.

The muscle proteins. The total amount of muscle proteins in mammals, including man, exceeds that of any other protein. About 40 percent of the body weight of a healthy human adult weighing about 70 kilograms (150 pounds) is muscle, which is composed of about 20 percent muscle protein. Thus, the human body contains about five to six kilograms (11 to 13 pounds) of muscle protein. An albumin-like fraction of these proteins, originally called myogen, contains various enzymes—phosphorylase, aldolase, glyceraldehyde phosphate dehydrogenase, and others; it does not seem to be involved in contraction, The globulin fraction contains myosin, the contractile protein, which also occurs in blood platelets, small bodies found in blood. Similar contractile substances occur in other contractile structures; for example, in the cilia or flagella (whiplike organs of locomotion) of bacteria and protozoans. In contrast to the scleroproteins, the contractile proteins are soluble in salt solutions and susceptible to enzymatic digestion.

Occurrence of contractile proteins

The energy required for muscle contraction is provided by the oxidation of carbohydrates or lipids. The term mechano-chemical reaction has been used for this conversion of chemical into mechanical energy. Although the molecular process underlying the reaction is not yet completely understood, it is known to involve the fibrous muscle proteins, the peptide chains of which undergo a change in conformation during contraction.

Myosin, which can be removed from fresh muscle by adding it to a chilled solution of dilute potassium chloride and sodium bicarbonate, is insoluble in water. Myosin, solutions of which are highly viscous, consists of an elongated—probably double-stranded—peptide chain, which is coiled at both ends in such a way that a terminal globule is formed. The length of the molecule is approximately 160 nanometres and its average diameter 2.6 nanometres. The equivalent weight of each of the two terminal globules is approximately 30,000; the molecular weight of myosin is close to 500,000. Trypsin splits myosin into large fragments called meromyosin. Myosin contains many amino acids with positively and negatively charged side chains (see Table 1); they form 18 and 16 percent, respectively, of the total number of amino acids. Myosin catalyzes the hydrolytic cleavage of ATP (adenosine triphosphate). A smaller protein with properties similar to those of myosin is tropomyosin. It has a molecular weight of 70,000 and dimensions of 45 by 2 nanometres. More than 90 percent of its peptide chains are present in the α-helix form.

Myosin combines easily with another muscle protein called actin, the molecular weight of which is about 50,000; it forms 12 to 15 percent of the muscle proteins. Actin can exist in two forms--one, G-actin, is globular; the other, F-actin, is fibrous. Actomyosin is a complex molecule formed by one molecule of myosin and one or two molecules of actin. In muscle, actin and myosin filaments are oriented parallel to each other and to the long axis of the muscle. The actin filaments are linked to each other lengthwise by fine threads called S filaments. During contraction the S filaments shorten, so that the actin filaments slide toward each other, past the myosin filaments, thus causing a shortening of the muscle (see MUSCLE CONTRACTION for a detailed description of the process).

Fibrinogen and fibrin. Fibrinogen, the protein of the blood plasma, is converted into the insoluble protein fibrin during the clotting process. The fibrinogen-free fluid obtained after removal of the clot, called blood serum, is blood plasma minus fibrinogen. The fibrinogen content of the blood plasma is 0.2 to 0.4 percent.

Fibrinogen can be precipitated from the blood plasma by half-saturation with sodium chloride. Fibrinogen solutions are highly viscous and show strong flow birefringence. In electron micrographs the molecules appear as rods with a length of 47.5 nanometres and a diameter of 1.5 nanometres; in addition, two terminal and a central nodule are visible. The molecular weight is 340,000. An unusually high percentage, about 36 percent, of the amino acid side chains are positively or negatively charged.

The clotting process in blood proteins

The clotting process is initiated by the enzyme thrombin, which catalyzes the breakage of a few peptide bonds of fibrinogen; as a result, two small fibrinopeptides with molecular weights of 1,900 and 2,400 are released. The remainder of the fibrinogen molecule, a monomer, is soluble and stable at pH values less than 6 (*i.e.*, in acid solutions). In neutral solution (pH 7) the monomer is converted into a larger molecule, insoluble fibrin; this results from the formation of new peptide bonds. The newly formed peptide bonds form intermolecular and intramolecular cross links, thus giving rise to a large clot, in which all molecules are linked to each other. Clotting, which takes place only in the presence of calcium ions, can be prevented by compounds such as oxalate or citrate, which have a high affinity for calcium ions.

ALBUMINS, GLOBULINS, AND OTHER SOLUBLE PROTEINS

The blood plasma, the lymph, and other animal fluids usually contain one to seven grams of protein per 100 millilitres of fluid, which includes small amounts of hundreds of enzymes and a large number of protein hormones. The discussion below is limited largely to the proteins that occur in large amounts and can be easily isolated from the body fluids. For further information on enzymes and hormones, see ENZYME and HORMONE.

Proteins of the blood serum. Human blood serum contains about 7 percent protein, two-thirds of which is in the albumin fraction; the other third is in the globulin fraction. Electrophoresis of serum reveals a large albumin peak and three smaller globulin peaks, the alpha-, beta-, and gamma-globulins. The amounts of alpha-, beta-, and gamma-globulin in normal human serum are approximately 1.5, 1.9, and 1.1 percent, respectively. Each globulin fraction is a mixture of many different proteins, as has been demonstrated by immuno-electrophoresis. In this method, the serum of a rabbit injected with human serum is allowed to diffuse into the four protein bands—albumin, alpha-, beta-, and gamma-globulin—obtained from the electrophoresis of human serum. Because the rabbit has previously been injected with human serum, its blood contains antibodies (substances formed in response to a foreign substance introduced into the body) against each of the human serum proteins; each antibody combines with the serum protein (antigen) that caused its formation in the rabbit. The result is the formation of about 20 regions of insoluble antigen–antibody precipitate, which appear as white arcs in the transparent gel of the electrophoresis medium. Each region corresponds to a different human serum protein.

Globulin fractions of serum

Serum albumin is much less heterogeneous (*i.e.*, contains fewer distinct proteins) than are the globulins; in fact, it is one of the few serum proteins that can be obtained in a crystalline form. Serum albumin combines easily with many acidic dyes (*e.g.*, Congo red and methyl orange); with bilirubin, the yellow bile pigment; and with fatty acids. It seems to act, in living organisms, as a carrier for certain biological substances. Present in blood serum in relatively high concentration, serum albumin also acts as a protective colloid, a protein that stabilizes other proteins. Albumin (molecular weight of 68,000) has a single free sulfhydryl (—SH) group, which on oxidation forms a disulfide bond with the sulfhydryl group of another serum albumin molecule, thus forming a dimer. The isoelectric point of serum albumin is pH 4.7.

The alpha-globulin fraction of blood serum is a mixture of several conjugated proteins. The best known are an ***a***-lipoprotein (combination of lipid and protein), and two mucoproteins (combinations of carbohydrate and protein). One mucoprotein is called orosomucoid, or m,-acid glycoprotein; the other is called haptoglobin because it combines specifically with globin, the protein component of hemoglobin. Haptoglobin contains about 20 percent carbohydrate.

The beta-globulin fraction of serum contains, in addition to lipoproteins and mucoproteins, two metal-binding proteins, transferrin and ceruloplasmin, which bind iron and copper, respectively. They are the principal iron and copper carriers of the blood.

The gamma-globulins are the most heretogeneous globulins. Although most have a molecular weight of approximately 150,000, that of some, called macroglobulins, is as high as 800,000. Because typical antibodies are of the same size and exhibit the same electrophoretic behaviour as y-globulins, they are called immunoglobulins. The designation IgM or gamma M (γM) is used for the macroglobulins; the designation IgG or gamma G (γG) is used for y-globulins of molecular weight 150,000.

Milk proteins. Milk contains an albumin, α-lactalbumin; a globulin, beta-lactoglobulin; and a phosphoprotein, casein. If acid is added to milk, casein precipitates. The remaining watery liquid (the supernatant solution), or whey, contains lactalbumin and lactoglobulin. Both have been obtained in crystalline form; their molecular weights are 16,000 and 18,500, respectively. Lactoglobulin also occurs as a dimer of molecular weight 37,000. Small variation? known to occur in the amino acid composition of lactoglobulin result from genetic variations. The amino acid composition and the tertiary structure of lactalbumin resemble that of lysozyme, an egg protein (see below).

Casein is precipitated not only by the addition of acid but also by the action of the enzyme rennin, which is found in gastric juice. Rennin from calf stomachs is used to precipitate casein, from which cheese is made. Milk fat precipitates with casein; milk sugar, however, remains in the supernatant (whey). Casein is a mixture of several similar phosphoproteins, called a-, β-, y-, and κ-casein, all of which contain some serine side chains combined with phosphoric acid. Approximately 75 percent of casein is a-casein (see Table 1). Cystine has been found only in κ-casein. In milk, casein seems to form polymeric globules (micelles) with radially arranged monomers, each with a molecular weight of 24,000; the acidic side chains occur predominantly on the surface of the micelle, rather than inside.

Composition of casein

Egg proteins. About 50 percent of the proteins of egg white are composed of ovalbumin, which is easily obtained in crystals. Its molecular weight is 46,000 and its amino acid composition differs from that of serum albumin. Other proteins of egg white are conalbumin, lysozyme, ovoglobulin, ovomucoid, and avidin. Lysozyme is an enzyme that hydrolyzes the carbohydrates found in the capsules certain bacteria secrete around themselves; it causes lysis (disintegration) of the bacteria. The molecular weight of lysozyme is 14,100; its amino acid composition is shown in Table 2. Its three-dimensional structure, shown in Figure 5, is similar to that of a-lactalbumin, which stimulates the formation of lactose by the enzyme lactose synthetase. Lysozyme has also been found in the urine of patients suffering from leukemia.

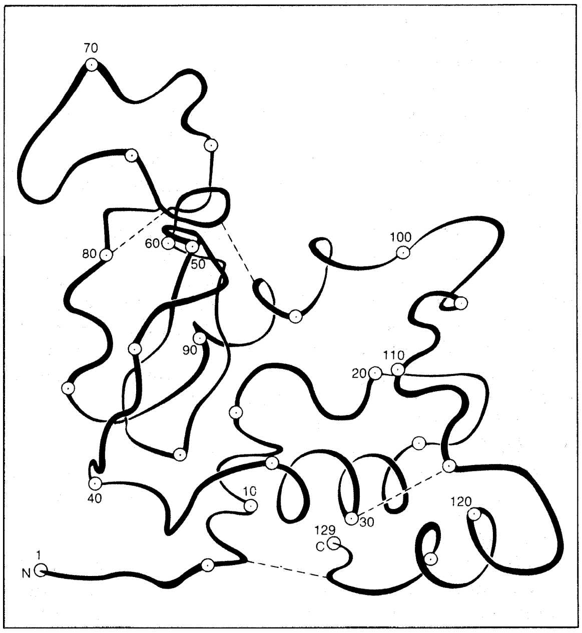


Figure 5: Conformation of lysozyme. Lysozyme from hen's egg white has a single peptide chain of 129 amino acids. The diagram shows the structure in simplified form. The amino acid residues are numbered from the terminal a amino group (N) to the terminal carboxyl group (C). Every fifth residue is shown by a circle, and every tenth residue is numbered. The four disulfide bridges are shown by broken lines. Alpha-helices are visible in the ranges 25 to 35.90 to 100, and 120 to 125.

Avidin is a glycoprotein that combines specifically with biotin, a vitamin. In animals fed large amounts of raw egg white, the action of avidin results in "egg-white injury." The molecular weight of avidin, which forms a tetramer, is 16,200. Its amino acid sequence is known.

Egg-yolk proteins contain a mixture of lipoproteins and livetins. The latter are similar to serum albumin, α-globulin, and β-globulin. The yolk also contains a phosphoprotein, phosvitin. Phosvitin, which has also been found in fish sperm, has a molecular weight of 40,000 and an unusual amino acid composition; one third of its amino acids are phosphoserine.

Protamines and histones. Protamines are found in the sperm cells of fish. The most thoroughly investigated protamines are salmine from salmon sperm and clupeine

Table 2: Number of Amino Acids per Protein Molecule

amino acid	protein*						
	Cyto	Hb a	Hb β	RNase	Lys	Chgen	Fdox
Lysine	18	11	11	10	6	14	4
Histidine	3	10	9	4	1	2	1
Arginine	2	3	3	4	11	4	1
Aspartic acid†	8	12	13	15	21	23	13
Threonine	7	9	7	10	7	23	8
Serine	2	11	5	15	10	28	7
Glutamic acid†	10	5	11	12	5	15	13
Proline	4	7	7	4	2	9	4
Glycine	13	7	13	3	12	23	6
Alanine	6	21	15	12	12	22	9
Half cystine	2	1	2	8	8	10	5
Valine	3	13	18	9	6	23	7
Methionine	3	2	1	4	2	2	0
Isoleucine	8	0	0	3	6	10	4
Leucine	6	18	18	2	8	19	8
Tyrosine	5	3	3	6	3	4	4
Phenylalanine	3	7	8	3	3	6	2
Tryptophan	1	1	2	0	6	8	1
Total	104	141	146	124	129	245	97

*Cyto = human cytochrome c; Hb a = human hemoglobin A, a-chain; Hb β = human hemoglobin A, 0-chain; RNase = bovine ribonuclease; Lys = chicken lysozyme; Chgen = bovine chymotrypsinogen; Fdox = spinach ferredoxin. †The values recorded for aspartic and glutamic acid include asparagine and glutamine.

from herring sperm. The protamines are bound to deoxyribonucleic acid (DNA), forming nucleoprotamines. The amino acid composition of the protamines is simple; they contain, in addition to large amounts of arginine, small amounts of five or six other amino acids. The composition of the salmine molecule, for example, is: Arg_{51}, Ala,, Val,, Ile_1, Pro,, and Ser,, in which the subscript numbers indicate the number of each amino acid in the molecule. Because of the high arginine content, the isoelectric points of the protamines are at pH values of 11 to 12; *i.e.*, the protamines are alkaline. The molecular weights of salmine and clupeine are close to 6,000. All of the protamines investigated thus far are mixtures of several similar proteins.

Protamine composition

The histones are less basic than the protamines. They contain high amounts of either lysine or arginine and small amounts of aspartic acid and glutamic acid. Histones occur in combination with DNA as nucleohistones in the nuclei of the body cells of animals and plants, but not in animal sperm. The molecular weights of histones vary from 10,000 to 22,000. In contrast to the protamines, the histones contain most of the 20 amino acids, with the exception of tryptophan and the sulfur-containing ones. Like the protamines, histone preparations are heterogeneous mixtures. The amino acid sequence of some of the histones has been determined.

Plant proteins. Plant proteins, mostly globulins, have been obtained chiefly from the protein-rich seeds of cereals and leguminous plants (*i.e.*, members of the pea family). Very small amounts of albumins are found in seeds. The best known globulins, insoluble in water, can be extracted from seeds by treatment with 2 to 10 percent solutions of sodium chloride. Many plant globulins have been obtained in crystalline form; they include edestin from hemp, molecular weight 310,000; amandin from almonds, 330,000; concanavalin A (42,000) and B (96,000); and canavalin (113,000) from jack beans. They are polymers of smaller subunits; edestin, for example, is a hexamer of a subunit with a molecular weight of 50,000, and concanavalin B a trimer of a subunit with a molecular weight of 30,000. After extraction of lipids from cereal seeds by ether and alcohol, further extraction with water containing 50 to 80 percent of alcohol yields proteins that are insoluble in water but soluble in water–ethanol mixtures and have been called prolamins. Their solubility in aqueous ethanol may result from their high proline and glutamine content (see Table 1). Gliadin, the prolamin from wheat, contains 14 grams of proline and 46 grams of glutamic acid in 100 grams of protein; most of the glutamic acid is in the form of glutamine. The total amounts of the basic amino acids (arginine, lysine, and histidine) in gliadin are only 5 percent of the weight of gliadin. None of the prolamins has yet been obtained in a pure crystalline state. Because the lysine content is either low or nonexistent, human populations dependent on grain as a sole protein source suffer from lysine deficiency.

CONJUGATED PROTEINS

Combination of proteins with prosthetic groups. The link between a protein molecule and its prosthetic group is a covalent (electron-sharing) bond in the glycoproteins, the biliproteins, and some of the heme proteins. In lipoproteins, nucleoproteins, and some heme proteins, the two components are linked to each other by noncovalent bonds. The noncovalent bonding results from the same forces that are responsible for the tertiary structure of proteins; namely, hydrogen bonds, salt bridges between positively and negatively charged groups, disulfide bonds, and mutual interaction of hydrophobic groups. In the metalloproteins (proteins with a metal element as a prosthetic group), the metal ion usually forms a centre to which various groups are bound.

Chemical bonding in conjugated proteins

Some of the conjugated proteins have been mentioned in preceding sections because they occur in the blood serum, in milk, and in eggs. Other conjugated proteins will be discussed below in sections dealing with respiratory proteins and enzymes.

Mucoproteins and glycoproteins. The prosthetic groups in mucoproteins and glycoproteins are oligosaccharides (carbohydrates consisting of a small number of simple sugar molecules) usually containing from four to 12 sugar molecules; the most common sugars are galactose, mannose, glucosamine, and galactosamine. Xylose, fucose, glucuronic acid, sialic acid, and other simple sugars sometimes also occur. Some mucoproteins contain 20 percent or more of carbohydrate, usually in several oligosaccharides attached to different parts of the peptide chain. The designation mucoprotein is used for proteins with more than 3 to 4 percent carbohydrate; if the carbohydrate content is less than 3 percent, the protein is sometimes called a glycoprotein or simply a protein.

Mucoproteins, highly viscous proteins originally called mucins, are found in saliva, in gastric juice, and in other animal secretions. Mucoproteins occur in large amounts in cartilage, synovial fluid (the lubricating fluid of joints and tendons), and egg white. The mucoprotein of cartilage is formed by the combination of collagen with chondroitinsulfuric acid, which is a polymer of either glucuronic or iduronic acid and acetylhexosamine or acetylgalactosamine. It is not yet clear whether or not chondroitinsulfate is bound to collagen by covalent bonds.

Lipoproteins and proteolipids. The bond between the protein and the lipid portion of lipoproteins and proteolipids is a noncovalent one. It is believed that some of the lipid is enclosed in a meshlike arrangement of peptide chains, and becomes accessible for reaction only after the unfolding of the chains by naturi[illegible]. [illegible]lth[illegible]gh lipoproteins in tl[illegible] α- and [illegible]ɔt[illegible] fraction f[illegible]ɔd serum are soluble in water (but insoluble in organic solvents), some of the brain lipoproteins, because they have a high lipid content, are soluble in organic solvents; they are called proteolipids. The β-lipoprotein of human blood serum is a macroglobulin with a molecular weight of about 1,300,000, 70 percent of which is lipid; of the lipid, about 30 percent is phospholipid and 40 percent cholesterol and compounds derived from it. Because of their lipid content, the lipoproteins have the lowest density (mass per unit volume) of all proteins. They are usually classified as LDL (low density lipoproteins) and HDL (high density lipoproteins).

Coloured lipoproteins are formed by the combination of protein with carotenoids. Crustacyanin, the pigment of lobsters, crayfish, and other crustaceans, contains astaxathin, which is a compound derived from carotene; a similar pigment is found in lobster eggs. Among the most interesting of the coloured lipoproteins are the pigments of the retina of the eye. They contain retinal, which is a compound derived from carotene and which is formed by the oxidation of vitamin A. In rhodopsin. the red pigment of the retina. the aldehyde group (—CHO) of retinal

Pigments of the eye

forms a covalent bond with an amino (—NH,) group of opsin, the protein carrier. Colour vision is mediated by the presence of several visual pigments in the retina that differ from rhodopsin either in the structure of retinal or in that of the protein carrier.

Metalloproteins. Proteins in which heavy metal ions are bound directly to some of the side chains of histidine, cysteine, or some other amino acid are called metalloproteins. Two metalloproteins, transferrin and ceruloplasmin, occur in the globulin fractions of blood serum; they act as carriers of iron and copper, respectively. Transferrin has a molecular weight of 84,000 and consists of two identical subunits, each of which contains one ferric ion (Fe^{3+}) that seems to be bound to tyrosine. Several genetic variants of transferrin are known to occur in man. Another iron protein, ferritin, which contains 20 to 22 percent iron, is the form in which iron is stored in animals; it has been obtained in crystalline form from liver and spleen. A molecule consisting of 20 subunits, its molecular weight is approximately 480,000. The iron can be removed by reduction from the ferric (Fe^{3+}) to the ferrous (Fe^{2+}) state. The iron-free protein, apoferritin, is synthesized in the body before the iron is incorporated.

Green plants and some photosynthetic and nitrogen-fixing bacteria (*i.e.*, bacteria that convert atmospheric nitrogen, N_2, into amino acids and proteins in their own bodies) contain various ferredoxins. They are small proteins containing 50 to 100 amino acids (see Table 2) and a chain of iron and disulfide units (FeS_2), in which some of the sulfur atoms are contributed by cysteine; others are sulfide ions (S^{2-}). The number of FeS_2 units per ferredoxin molecule varies from five in the ferredoxin of spinach to ten in the ferredoxin of certain bacteria. Ferredoxins act as electron carriers in photosynthesis and in nitrogen fixation.

Function of ceruloplasmin

Ceruloplasmin is a copper-containing globulin with a molecular weight of 151,000; the molecule consists of eight subunits, each containing one copper ion. Ceruloplasmin is the principal carrier of copper in organisms, although copper can also be transported by the iron-containing globulin transferrin. Another copper-containing protein, erythrocuprein (molecular weight 64,000), has been isolated from red blood cells; it has also been found in the liver and the brain. The molecule, which consists of four subunits with a molecular weight of 16,000 each, contains four copper and four zinc ions. Because of their copper content, ceruloplasmin and erythrocuprein may have some catalytic activity in oxidation–reduction reactions. Another copper-containing protein, hemocyanin, is described below (see *Respiratory* proteins).

Many animal enzymes contain zinc ions, which are usually bound to the sulfur of cysteine. Horse kidneys contain the protein metallothionein, which, in addition to 2.2 percent zinc, contains 5.9 percent cadmium; both are bound to sulfur. A vanadium-protein complex (homovanadin) has been found in surprisingly high amounts in yellowish-green cells (vanadocytes) of tunicates, which are marine invertebrates.

Heme proteins and other chromoproteins. Although the heme proteins contain iron, they are usually not classified as metalloproteins, because their prosthetic group is an iron-porphyrin complex in which the iron is bound very firmly. The intense red or brown colour of the heme proteins is not caused by iron but by porphyrin, a complex cyclic structure. All porphyrin compounds absorb light intensely at or close to 410 nanometres. Porphyrin consists of four pyrrole rings (five-membered closed structures containing one nitrogen and four carbon atoms) linked to each other by methine groups (—CH=). The iron atom is kept in the centre of the porphyrin ring by interaction with the four nitrogen atoms. The iron atom can combine with two other substituents; in oxyhemoglobin, one substituent is a histidine of the protein carrier, the other is an oxygen molecule. In some heme proteins, the protein is also bound covalently to the side chains of porphyrin. Heme proteins important as respiratory proteins and enzymes are described below (see Respiratory proteins and Oxidoreductases).

Little is known about the structure of the chromoprotein melanin, a pigment found in dark skin, dark hair, and melanotic tumours. It is probably formed by the oxidation of tyrosine, which results in the formation of red, brown, or dark-coloured derivatives.

Green chromoproteins called biliproteins are found in many insects, such as grasshoppers, and also in the eggshells of many birds. The biliproteins are derived from tbe bile pigment biliverdin, which in turn is formed from porphyrin; biliverdin contains four pyrrole rings and three of the four methine groups of porphyrin. Large amounts of biliproteins, the molecular weights of which are about 270,000, have been found in red and blue-green algae; the red protein is called phycoerythrin, the blue one phycocyanobilin. Phycocyanobilin consists of eight subunits with a molecular weight of 28,000 each; about 89 percent of the molecule is protein. It also contains considerable amounts of carbohydrate.

Nucleoproteins. When a protein solution is mixed with a solution of a nucleic acid, the phosphoric acid component of the nucleic acid combines with the positively charged ammonium groups ($—NH_3^+$) of the protein to form a protein–nucleic acid complex. The nucleus of a cell contains predominantly deoxyribonucleic acid (DNA) and the cytoplasm predominantly ribonucleic acid (RNA); both parts of the cell also contain protein. Protein–nucleic acid complexes, therefore, form in living cells. It has not yet been definitely established whether the protein–nucleic acid complexes isolated from biological material are indeed formed during the life of the organism or whether they are artifacts produced during the isolation procedure.

The only nucleoproteins for which some evidence for specificity exists are nucleoprotamines, nucleohistones, and some RNA and DNA viruses. The nucleoprotamines are the form in which protamines occur in the sperm cells of fish; the histones of the thymus and of pea seedlings and other plant material apparently occur predominantly as nucleohistones. Both nucleoprotamines and nucleohistones contain only DNA.

Proteins of the tobacco mosaic virus

Some of the simplest viruses consist of a specific RNA, which is coated by protein. One of the best known RNA viruses, tobacco mosaic virus (TMV), has the shape of a rod. RNA comprises only 5.1 percent of the mass of the virus. The complete sequence of the virus protein, which consists of about 2,130 identical peptide chains, each containing 158 amino acids, has been determined. The protein is arranged in a spiral around the RNA core.

DNA has been found in most bacterial viruses (bacteriophages) and in some animal viruses. As in TMV, the core of DNA is surrounded by protein. Phage protein is a mixture of enzymes and therefore cannot be considered as the protein portion of only one nucleoprotein.

Respiratory proteins. Hemoglobin. Hemoglobin is the oxygen carrier in all vertebrates and some invertebrates. In oxyhemoglobin (HbO_2), which is bright red, the ferrous ion (Fe^{2+}) is bound to the four nitrogen atoms of porphyrin; the other two substituents are an oxygen molecule and the histidine of globin, the protein component of hemoglobin. Deoxyhemoglobin (deoxy-Hb), as its name implies, is oxyhemoglobin minus oxygen (*i.e.*, reduced hemoglobin); it is purple in colour. Oxidation of the ferrous ion of hemoglobin yields a ferric compound, methemoglobin, sometimes called hemiglobin or ferrihemoglobin. The oxygen of oxyhemoglobin can be displaced by carbon monoxide, for which hemoglobin has a much greater affinity, preventing oxygen from reaching the body tissues.

The hemoglobins of all mammals, birds, and many other vertebrates are tetramers of two *a*- and two β-chains (see Table 2). The molecular weight of the tetramer is 64,500; the molecular weight of the α- and β-chains is approximately 16,100. each, and the four subunits are linked to each other by noncovalent interactions. If hemin (the ferric porphyrin component) is removed from globin (the protein component), two molecules of globin, each consisting of one *a*- and one β-chain, are obtained; the molecular weight of globin is 32,200. In contrast to hemoglobin, globin is a rather unstable protein that is easily denatured. If native globin is incubated with a solution of hemin at pH values of 8 to 9, native hemoglobin is reconstituted. Both the hemoglobin of the lamprey (of the vertebrate class Agnatha) and the myoglobin, the red pigment of mammalian muscles, are monomers with a molecular weight of 16,000.

Mammalian hemoglobins

The mammalian hemoglobins differ from each other in their amino acid composition and therefore in their secondary and tertiary structure. Rat and horse hemoglobin cystallize very easily, but those of man, cattle, and sheep, because they are more soluble, are difficult to crystallize. The shape of hemoglobin crystals varies in different species; moreover, decomposition and denaturation occur at different rates in different species. It was also found that the blood of newborn children contains two different hemoglobins, about 20 percent of an adult hemoglobin (hemoglobin A) and 80 percent of a fetal hemoglobin (hemoglobin F). Hemoglobin F persists in the child for the first seven months of life. The same hemoglobin F has also been found in the blood of patients suffering from thalassemia, an anemia that occurs in the countries of southern Europe. Hemoglobin F contains, as does hemoglobin A, two a-chains; the two β-chains, however, have been replaced by two quite different γ-chains. When the technique of electrophoresis was first applied to the hemoglobin of Negroes suffering from sickle cell anemia in 1949, a new hemoglobin (hemoglobin S) was discovered. More than 100 different human hemoglobins now are known. They differ from normal hemoglobin A in the amino acid composition of either the a- or the β-chain.

The hemoglobins of some of the lowest fishes are monomers containing one iron atom per molecule. Hemoglobin-like respiratory proteins have been found in some invertebrates. The red hemoglobin of insects, mollusks, and protozoans is called erythrocruorin. It differs from vertebrate hemoglobin by its high molecular weight.

Although green plants contain no hemoglobin, a red protein, called leg-hemoglobin, has been discovered in the root nodules of leguminous plants. It seems to be produced by the nitrogen-fixing bacteria of the root nodules and may be involved in the reduction of atmospheric nitrogen to ammonia and amino acids.

Other respiratory proteins. A green respiratory protein, chlorocruorin, has been found in the blood of the marine worm Spirographis. It has the same high molecular weight as erythrocruorin, but differs from hemoglobin in its prosthetic group. A red metalloprotein, hemerythrin, acts as a respiratory protein in marine worms of the phylum Sipuncula. The molecule consists of eight subunits with a molecular weight of 13,500 each. Hemerythrin contains no porphyrins and therefore is not a heme protein.

Hemocyanin

A metalloprotein containing copper is the respiratory protein of crustaceans (shrimps, crabs, etc.) and of some gastropods (snails). The protein, called hemocyanin, is pale yellow when not combined with oxygen, and blue when combined with oxygen. The molecular weights of hemocyanins vary from 300,000 to 9,000,000. Each animal investigated thus far apparently has a species-specific hemocyanin.

CATALYTIC PROTEINS: ENZYMES

Crystalline urease and pepsin were the first enzymes shown to be pure proteins, free of any nonprotein material. It is now known that all enzymes are either pure or conjugated proteins. Enzymes can be defined as catalytically active proteins. In many enzymes that are conjugated proteins, the prosthetic group is the catalytically active site. The protein carrier of the prosthetic group is designated the apoenzyme. The International Union of Biochemistry (IUB) recommended in 1964 classification of the enzymes, according to their mode of action, into six classes: (1) oxidoreductases, (2) transferases, (3) hydrolases, (4) lyases, (5) isomerases, and (6) ligases, or synthetases. Each class is further divided in subclasses, sub-sub-classes, etc; for example, the oxidoreductases (class 1) are divided into 14 subclasses according to the chemical group that undergoes oxidation. The first subclass is further subdivided into four categories according to the electron acceptor involved. More than 600 enzymes were listed in 1964, and their number increases each year. Enzymes catalyze chemical reactions in which energy is released (exergonic reactions); they cannot catalyze energy-requiring reactions (endergonic reactions) unless energy is supplied by another reaction. The mechanism of many enzymatic reactions involves an activator; for example, the enzyme amylase, which catalyzes the hydrolysis of starch to smaller carbohydrate units called oligosaccharides, requires chloride ions (Cl^-) as an activator. The activators of oxidoreductases, usually called coenzymes, have complicated structures; they either accept or donate electrons during the catalytic process. The forces involved in the combination of an enzyme with the compound with which it acts (substrate) and with activators or coenzymes are the same as those that determine the conformation of proteins—*e.g.*, hydrogen bonds, electrostatic interaction between positively and negatively charged groups, and mutual attraction of nonpolar groups. Because each enzyme catalyzes a limited type of reaction and frequently reacts with only one substance, its active site is assumed to have a rather rigid conformation, enzyme and substrate fitting each other like a lock and key. Intermolecular forces can then induce a definite orientation of the substrate and render some of its groups more capable of undergoing chemical change; as a result, the reaction converting substrate into product can take place.

Oxidoreductases. Most oxidoreductases, usually called dehydrogenases, either have an iron-porphyrin as the prosthetic group or require a coenzyme. The specificity of enzyme action is determined by the apoenzyme (protein moiety). Almost 90 enzymes act on the alcohol group (—CHOH—); 70 require nicotinamide adenine dinucleotide (NAD^+) or its phosphate ($NADP^+$) as an electron acceptor. The electron acceptors in the others are either oxygen, ferricytochrome *c*, or an as yet unknown substance. Dehydrogenases also catalyze other types of reactions—the oxidation of aldehydes and ketones to carboxylic acids, the dehydrogenation of amino acids, and the oxidation of xanthine.

Some dehydrogenases have been isolated. Glyceraldehydephosphate dehydrogenase, the active site of which contains cysteine, forms about 10 percent of the soluble proteins of muscle. Lactate dehydrogenase contains four subunits in a molecule with a molecular weight of 150,000; the two types of subunits, α and β, exist in five different combinations, called isozymes. Each isozyme occurs in different tissues and has lactate dehydrogenase activity.

Not all dehydrogenations are mediated by the coenzymes NAD^+ and NADP', which are loosely bound to their apoenzymes. Another coenzyme that transfers electrons, riboflavin, is firmly bound to its apoenzyme. Accordingly, the enzymes are called flavoproteins, or yellow enzymes (*flavus* is the Latin word for yellow). The oxidation (dehydrogenation) of the L-amino acids and xanthine, a compound found in most body tissues and some plants, is catalyzed by flavoproteins.

Because tissues contain only limited amounts of coenzymes, a mechanism exists by which the hydrogenated (*i.e.*, reduced) forms can be reoxidized, so that they can accept more hydrogen. The mechanism involves the reduction of oxygen, the principal hydrogen acceptor in all organisms, to water. Oxygen does not react directly with reduced NAD or reduced flavoproteins; rather, hydrogen and electrons are transferred through a series of enzymes called cytochromes to oxygen. Cytochromes are typical heme proteins: the most thoroughly investigated one is cytochrome c from heart muscle. The complete amino acid sequence of a large number of cytochromes *c* of vertebrates, invertebrates, plants, protozoans, and bacteria has been elucidated (see Table 2); as a result, the evolution of the cytochrome c molecule has been clarified. Several other cytochromes are involved in the mechanism by which electrons are transferred from reduced NAD, reduced NADP, and reduced flavoproteins to oxygen (see METABOLISM).

Although most biological oxidations involved in metabolism require the re-oxidation of reduced coenzymes via a series of cytochromes, tyrosinase and similar copper-containing enzymes catalyze the direct transfer of electrons from phenol derivatives to oxygen, frequently forming hydrogen peroxide. The enzyme peroxidase catalyzes the transfer of hydrogen atoms from hydrogen donors to hydrogen peroxide. Catalase catalyzes the splitting of two hydrogen peroxide molecules into oxygen and water. Both peroxidase and catalase are heme proteins.

Function of some copper-containing enzymes

Transferases. Transferases catalyze the transfer of various chemical groups from one compound to another. Phosphate-transferring enzymes, or phosphokinases, for example, catalyze the transfer of phosphate from adenosine triphosphate (ATP) to a sugar or some other compound. One of the most thoroughly investigated transferases, ribonuclease (RNase), has a molecular weight of 13,700; it consists of one peptide chain, the complete amino acid content and sequence of which is known (see Table 2). The chain contains four disulfide bonds (—S—S—); when they are reduced to sulfhydryl groups (— SH), the enzyme is no longer active as a catalyst; activity can be regained by re-oxidation.

Hydrolases. Hydrolases catalyze the hydrolysis of ester bonds (between an acid and a base) in lipids, glycoside bonds (between sugar molecules) in carbohydrates, and amide or peptide bonds (in proteins, peptides, and a few other substrates such as urea). Many hydrolases can be isolated in large amounts and therefore have been investigated thoroughly, particularly enzymes that catalyze the hydrolysis of peptide bonds—trypsin, chymotrypsin, pepsin, and papain. The molecular weights (24,000, 24,300, 32,700, and 21,000, respectively) and the complete amino acid sequences of each have been resolved. Trypsin and chymotrypsin occur in the pancreatic juice in an inactive form called trypsinogen and chymotrypsinogen, respectively. The inactive form of pepsin, pepsinogen, is found in the gastric juice of the stomach. Papain is present in the juice of papaya fruit.

These proteolytic enzymes hydrolyze peptide bonds inside the peptide chain; however, the enzyme carboxypeptidase, which occurs in pancreatic juice, hydrolyzes the C terminal peptide bond; *i.e.*, the bond joining the amino acid with a free carboxyl group (—COOH) to the rest of the protein. The N terminal peptide bond—*i.e.*, the bond joining the amino acid with a free amino group (—NH_2) to the rest of the protein—is hydrolyzed by leucine aminopeptidase, which is found in the intestinal juice. Amide bonds (—CONH,) are hydrolyzed by asparaginase, glutaminase, arginase, and urease (molecular weight 480,000).

Little is known about the chemical structure and the active sites of the enzymes that hydrolyze glycoside bonds such as amylase, which hydrolyzes starch and occurs in saliva and pancreatic juice.

Lysozyme catalyzes the hydrolysis of the glycosidic bond between N-acetylmuramic acid and glucosamine in complex compounds called mucopolysaccharides. The molecular weight of lysozyme is 14,400; like ribonuclease, it contains four disulfide bonds in one peptide chain. Its amino acid sequence is known, and its conformation has been resolved by X-ray diffraction (see Figure 5). Enzymes that hydrolyze ester bonds include lipase, phospholipase A, and cholinesterase.

Structural features of lysozyme

Lyases, isomerases, and ligases. Lyases include a large number of enzymes that catalyze the cleavage of carbon–carbon, carbon–oxygen, and carbon–nitrogen bonds. Carbon–carbon bonds are split by the action of decarboxylases, which remove carbon dioxide from carboxyl groups. Amino acids are converted by bacterial decarboxylases into amines; thus, for example, the group —$CHNH_2$. COOH is converted into —CH_2NH_2 and carbon dioxide is released in the process. Carbonic anhydrase, a zinc protein, catalyzes the release of CO, from bicarbonate and thus enables the removal of CO_2 from man and other vertebrates.

Isomerases catalyze the conversion of various substrates into their isomeric (mirror-image) forms—for example, the conversion of L-alanine into D-alanine or vice versa (see Formula 4). The enzyme that catalyzes this reaction is called alanine racernase. The conversion of a-D-glucose into β-D-glucose is catalyzed by a mutarotase. Numerous other isomerases exist.

Ligases catalyze the formation of carbon–oxygen, carbon–sulfur, carbon–nitrogen, and carbon–carbon bonds. The formation of these bonds is an energy-requiring reaction that cannot take place without the simultaneous occurrence of an energy-releasing reaction. The latter is in most instances the conversion of ATP (adenosine triphosphate) into AMP (adenosine monophosphate).

PROTEIN HORMONES

Some hormones that are products of endocrine glands are proteins or peptides; others are steroids. (The origin of hormones, their physiological role, and their mode of action are dealt with in the article HORMONE.) None of the hormones has any enzymatic activity. Each has a target organ in which it elicits some biological action; *e.g.*, secretion of gastric or pancreatic juice, production of milk, production of steroid hormones, or the production of substances that cause dilation or constriction of blood vessels. The mechanism by which the hormones exert their effects is not yet fully understood. Cyclic adenosine monophosphate is involved in the transmittance of the hormonal stimulus to the cells whose activity is specifically increased by the hormone.

Functional aspects of hormones

Hormones of the thyroid gland. Thyroglobulin, the active groups of which are two molecules of the iodine-containing compound thyroxine (see Figure 1), has a molecular weight of 670,000.

Thyroglobulin also contains thyroxine with two and three iodine atoms instead of the four shown in Figure 1, and tyrosine, with one and two iodine atoms. Injection of the hormone causes an increase in metabolism; lack of it results in a slowdown. Another hormone, calcitonin, which lowers the calcium level of the blood, occurs in the thyroid gland. The structure of human calcitonin is given in Formula 7 (see Figure 1 for structures of amino acids corresponding to the one-letter codes).

The amino acid sequences of calcitonin from pig, beef, and salmon differ from human calcitonin in some amino acids. All of them, however, have the half-cystines and the prolinamide in the same position. Porcine calcitonin has been synthesized in the laboratory.

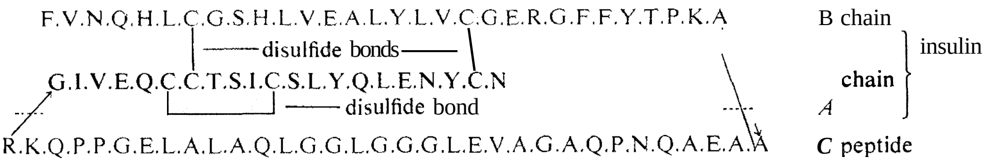
The parathyroid hormone (parathormone), which is produced in small glands that are embedded in or lie behind the thyroid gland, is essential for the maintenance of the calcium level of the blood. Its lack results in the disease hypocalcemia. Bovine parathormone has a molecular weight of 8,500; it contains no cystine or cysteine and is particularly rich in aspartic acid, glutamic acid, or their amides.

Pancreas hormones. Although the structure of insulin has been known since 1949, repeated attempts to synthesize it gave very poor yields because of the failure of the two peptide chains to combine forming the correct disulfide bridge. The ease of the biosynthesis of insulin is explained by the discovery in the pancreas of proinsulin, from which insulin is formed. The single peptide chain of proinsulin loses a peptide consisting of 33 amino acids and called the connecting peptide, or C peptide, during its conversion to insulin. The structure of porcine proinsulin is shown in Formula 8.

In aqueous solutions insulin exists predominantly as a complex of six subunits, each of which contains an *A*- and a B-chain. The insulins of several species have been isolated and analyzed; their amino acid sequences differ

C.S.N.L.S.T.C.V.L.S.A.Y.W.K.D.L.N.N.Y.H.R.F.S.G.M.G.F.G.P.E.T.P($CONH_2$)

Formula 7: The amino acid sequence of human calcitonin. At the left end the line represents the disulfide bond. At the right end ($CONH_2$) indicates that the C terminal proline is present as prolinamide.


F.V.N.Q.H.L.C.G.S.H.L.V.E.A.L.Y.L.V.C.G.E.R.G.F.F.Y.T.P.K.A
disulfide bonds
G.I.V.E.Q.C.C.T.S.I.C.S.L.Y.Q.L.E.N.Y.C.N
disulfide bond
R.K.Q.P.P.G.E.L.A.L.A.Q.L.G.G.L.G.G.G.L.E.V.A.G.A.Q.P.N.Q.A.E.A.A
B chain
insulin
chain
A
C peptide

Formula 8: The amino acid sequence of porcine proinsulin. The arrows indicate the direction from the N terminus of the B chain to the *C* terminus of the A chain. In the C peptide the N terminus is at the right (A = alanine) end and the *C* terminus is at the left end in order to show the connection with the A and *B* chains. Insulin is released when the two peptide bonds at the ends of the *C* peptide (broken lines) are hydrolyzed (broken by the introduction of a water molecule).

somewhat, but all apparently contain the same disulfide bridges between the two chains.

Opposite functions of insulin and glucagon

Although the injection of insulin lowers the blood sugar, administration of glucagon, another pancreas hormone, raises the blood sugar level. Glucagon consists of a straight peptide chain of *29* amino acids. Its structure, which is free of cystine and isoleucine, is given in Formula *9*. It has been synthesized; the synthetic product has the full biological activity of natural glucagon.

H.S.Q.G.T.F.T.S.D.Y.S.K.Y.L.D.S.R.R.A.Q.D.F.V.Q.W.L.M.N.T

Formula 9: Amino acid sequence of the pancreas hormone glucagon.

Pituitary hormones. The pituitary gland has an anterior lobe, a posterior lobe, and an intermediate portion; they differ in cellular structure and in the structure and action of the hormones they form. The posterior lobe produces two similar hormones, oxytocin and vasopressin. The former causes contraction of the pregnant uterus; the latter raises the blood pressure. Both are octapeptides formed by a ring of five amino acids (the two cystine halves count as one amino acid) and a side chain of three amino acids. The structure of oxytocin is given in Formula *10*. The two cystine halves are linked to each other by a disulfide bond, and the **C** terminal amino acid is glycinamide. The structure has been established and confirmed. Human vasopressin (Formula *10)* differs from oxytocin in that isoleucine is replaced by phenylalanine and leucine by arginine. Porcine vasopressin contains lysine instead of arginine.

A Cys.Tyr.Ile.GluN.Asn.Cys.Pro.Leu.Gly($CONH_2$)

B Cys.Tyr.Phe.GluN.Asn.Cys.Pro.Arg.Gly($CONH_2$)

Formula 10: Amino acid sequence of the two similar hormones of the posterior lobe of the pituitary gland. (A) Oxytocin. (B) Human vasopressin. The solid line represents the disulfide bond between the two halves of cystine.

The intermediate part of the pituitary gland produces the melanocyte-stimulating hormone (MSH), which causes expansion of the pigmented melanophores (cells) in the skin of frogs and other batrachians. Two hormones, called α-MSH and β-MSH, have been prepared from hog pituitary glands. a-MSHconsists of *13* amino acids (see Formula *11*); its N terminal serine is acetylated (*i.e.*, the acetyl group, CH_3CO—, of acetic acid is attached), and its C terminal valine residue is present as valinamide. β-MSH contains in its *18* amino acids many of those occurring in α-MSH (see Formula *11*).

The anterior pituitary lobe produces several protein hormones—a thvroid-stimulating hormone, molecular weight *28,000;* a lactogenic hormone, molecular weight *22,500;* a growth hormone, molecular weight *21.500;* a luteinizing hormone, molecular weight 30,000; and a follicle-stimulating hormone, molecular weight *29,000.* The thyroid-stimulating hormone (TSH, thyrotropin) consists of a and β subunits with a composition similar to the subunits of luteinizing hormone. When separated, neither of the two subunits has hormonal activity; when combined, however, they regain about *50* percent of the original activity. The lactogenic hormone (prolactin) from sheep pituitary glands contains *190* amino acids. Their sequence has been elucidated; a similar peptide chain of *188* amino acids that has been synthesized not only has *10* percent of the biological activity of the natural hormone but also some activity of the growth hormone. The amino acid sequence of the growth hormone (somatotropic hormone) is also known; it seems to stimulate the synthesis of RNA and in this way to accelerate growth, The luteinizing hormone (LH) consists of two subunits, each with a molecular weight of approximately *15,000;* when separated, the subunits recombine spontaneously. LH is a mucoprotein containing about *12* percent carbohydrate. The urine of pregnant women contains chorionic gonadotropin, the presence of which makes possible early diagnosis of pregnancy. The amino acid sequence is known. The sequence of *160* of its *190* amino acids is identical with those of the growth hormone; *100* of these also occur in the same sequence as in lactogenic hormone. The different pituitary hormones and the chorionic gonadotropin thus may have been derived from a common substance that, during evolution, underwent differentiation.

Derivation of pituitary hormones

Peptides with hormonelike activity. Small peptides have been discovered that, like hormones, act on certain target organs. One peptide, angiotensin (angiotonin or hypertensin), is formed in the blood from angiotensinogen by the action of renin, an enzyme of the kidney. It is an octapeptide and increases blood pressure. Similar peptides include bradykinin, which stimulates smooth muscles; gastrin, which stimulates secretion of hydrochloric acid and pepsin in the stomach; secretin, which stimulates the flow of pancreatic juice; and kallikreine, the activity of which is similar to bradykinin.

IMMUNOGLOBULINS AND ANTIBODIES

Antibodies, proteins that combat foreign substances in the body, are associated with the globulin fraction of the immune serum (see IMMUNITY). As stated previously, when the serum globulins are separated into α-, p-, and y-fractions, antibodies are associated with the y-globulins. Antibodies can be purified by precipitation with the antigen (*i.e.*, the foreign substance) that caused their formation, followed by separation of the antigen–antibody complex. Antibodies prepared in this way consist of a mixture of many similar antibody molecules, which differ in molecular weight, amino acid composition, and other properties. The same differences are found in the y-globulins of normal blood serums. It is believed that the y-globulin of normal blood serum is a mixture of thousands of different y-globulins, each of which occurs in

(CH_3CO)S.$\underline{\text{Y}}$.S. $\underline{\text{M.E.H.F.R.W.G}}$.K.P.V($CONH_2$)	porcine α-MSH, melanocyte-stimulating hormone
D.S.G.P.$\underline{\text{Y}}$.K.$\underline{\text{M.E.H.F.R.W.G}}$.S.P.P.K.D	porcine β-MSH
A.E.K.K.D.E.G.P.$\underline{\text{Y}}$.K.$\underline{\text{M.E.H.F.R.W.G}}$.S.P.P.K.D	human β-MSH
S.$\underline{\text{Y}}$.S. $\underline{\text{M.E.H.F.R.W.G}}$.K.P.V.G.K.K.R.R.P.V.K.V.Y.P.D.G.A.E.D.Q.L.A.E.A.F.P.L.E.F	porcine β-corticotropin

Formula 11: The amino acid seauence of hormones produced by the intermediate part of the pituitary gland. The amino acid sequence, M.E.H.F.R.W.G. occurs in all melanocyte-stimulating hormones and in adrenocorticotropic hormones (corticotropins).

amounts too small for isolation. Because the physical and chemical properties of normal y-globulins are the same as those of antibodies, the y-globulins are frequently called immunoglobulins. They may be considered to be antibodies against unknown antigens. If solutions of y-globulin are resolved by gel filtration through dextran, the first fraction has a molecular weight of 800,000. This fraction is called IgM or γM; Ig is an abbreviation for immunoglobulin and M for macroglobulin. The next two fractions are IgA (γA) and IgG (γG), with molecular weights of about 300,000 and 150,000, respectively. Two other immunoglobulins, known as IgD and IgE, have also been detected in much smaller amounts in some immune sera.

γ-globulins of blood serums

The bulk of the immunoglobulins is found in the IgG fraction, which also contains most of the antibodies. The IgM molecules are apparently pentamers—aggregates of five of the IgG molecules. Electron microscopy shows their five subunits to be linked to each other by disulfide bonds in the form of a pentagon. The IgA molecules are found principally in milk and in secretions of the intestinal mucosa. Some of them contain, in addition to a dimer of IgG, a "secretory piece," the structure of which is not yet known. The IgM and IgA immunoglobulins and antibodies contain 10 to 15 percent carbohydrate; the carbohydrate content of the IgG molecules is 2 to 3 percent.

IgG molecules treated with the enzyme papain split into three fragments of almost identical molecular weight of 50,000. Two of these, called Fab fragments, are identical; the third is abbreviated Fc. Reduction to sulfhydryl groups of some of the disulfide bonds of IgG results in the formation of two heavy, or H, chains (molecular weight 55,000) and two light, or *L*, chains (molecular weight 22,000). They are linked by disulfide bonds in the order *L*–*H*–*H*–*L*.Each *H* chain contains four intrachain disulfide bonds, each *L* chain contains two. The structure of antibodies and normal immunoglobulins of the IgG type is shown in Figure 6.

Antibody preparations of the IgG type, even after removal of IgM and IgA antibodies, are heterogeneous. The *H* and *L* chains consist of a large number of different *L* chains and a variety of *H* chains. Pure IgG, IgM, and IgA immunoglobulins, however, occur in the blood serum of patients suffering from myelomas, which are malignant tumours of the bone marrow. The tumours produce either an IgG, an IgM, or an IgA protein, but rarely more than one class. A protein called the Bence-Jones protein, which is found in the urine of patients suffering from myeloma tumours, is identical with the *L* chains of the myeloma protein. Each patient has a different Bence-Jones protein; no two of the more than 100 Bence-Jones proteins that have been analyzed thus far are identical. It is thought that one lymphoid cell amona hundreds of thousands becomes malignant and multiplies rapidly, forming the mass of a myeloma tumour that produces one y-globulin.

Bence-Jones proteins

Analyses of the Bence-Jones proteins have revealed that the L chains of man and other mammals are of two quite different types, kappa (κ) and lambda (A). Both consist of approximately 220.amino acids. The N–terminal halves of κ- and A-chains are variable, differing in each Bence-Jones protein. The C–terminal halves of these same L chains have a constant amino acid sequence of either the κ- or the A-type. The fact that one half of a peptide chain is variable and the other half invariant is contradictory to the view that the amino acid sequence of each peptide chain is determined by one gene (see GENE). Evidently, two genes, one of them variable, the other invariant, fuse to form the gene for the single peptide chain of the *L* chains. Whereas the normal human L chains are always mixtures of the κ- and A-types, the H chains of IgG, IgM, and IgA are different. They have been designated as gamma (y), mu (μ), and alpha (α) chains, respectively. The N terminal quarter of the H chains has a variable amino acid sequence; the C terminal three-quarters of the H chains have a constant amino acid sequence, as indicated in Figure 6.

Some of the amino acid sequences in the L and H chains are transmitted from generation to generation. Thus, the constant portion of the human *L* chains of the κ-type has in position 191 either valine or leucine. They correspond to two alleles (character-determining portions) of a gene; the two types are called allotypes. The valine-containing genetic type has been designated as InV(a$^+$), the leucine-containing type as InV(b$^+$). Many more allotypes, called Gm allotypes, have been found in the gamma chains of the human IgG immunoglobulins; more than 20 Gm allotypes are now known. Certain combinations of Gm types occur; the combination of Gm types 5, 6, and 11 has been found in Caucasians and Negroes but not in Chinese; the combination of 1, 2, and 17 has not been found in Negroes; and the combination of 1, 4, and 17 has not been found in Caucasians. Allotypes have also been discovered to occur in rabbits, mice, and other animals.

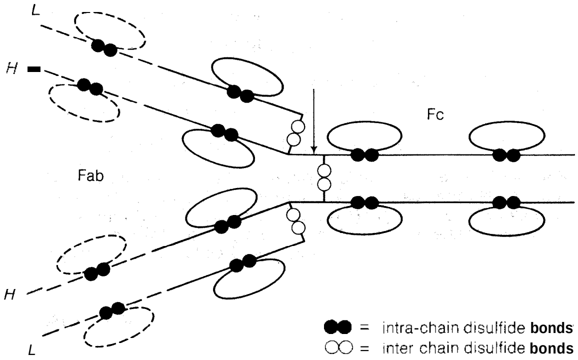

L
H
Fc
Fab
H
L
●● = intra-chain disulfide bonds
○○ = inter chain disulfide bonds

Figure 6: Diagram of an IgG immunoglobulin. Two heavy chains (*H*) and two light chains (*L*) are linked to each other by inter-chain disulfide bonds. Intra-chain disulfide bonds cause loops to form in the 12 peptide portions, each of which contains about 110 amino acid residues. The 12 peptide regions have cystine residues at similar positions and other similarities in their amino acid sequences. The broken lines represent variable portions and the solid lines represent constant portions of the chains. Specific sites that bind antigens are formed by the variable portions. The vertical arrow indicates cleavage of the IgG molecule into two Fab fragments and one Fc fragment by the action of the enzyme papain.

It is understandable from the occurrence of a large number of allotypes that antibodies, even if produced in response to a single antigen, are mixtures of different allotypes. The existence of several classes of antibodies, of different allotypes, and of adaptation of the variable portions of antibodies to different regions of an antigen molecule results in a multiplicity of antibody molecules even if only a single antigen is administered. For this reason it has not yet been possible to unravel the amino acid sequence in the variable portion of antibody molecules. Much of the amino acid sequence in the constant regions of the *L* and H chains of man and rabbit immunoglobulins, however, has been resolved.

BIBLIOGRAPHY

Books: HANS NEURATH (ed.), *The Proteins,* 2nd ed., 5 vol. (1963–70), a comprehensive discussion of the structure, function, and biology of proteins; FELIX HAUROWITZ, *The Chemistry and Function of Proteins,* 2nd ed. (1963), a textbook for graduate students; M.F. PERUTZ, *Proteins and Nucleic Acids* (1963), deals with hemoglobin and the role of nucleic acids in protein biosynthesis; M.O. DAYHOFF (ed.), *Atlas of Protein Sequence and Structure* (1969, published in intervals of 1 or 2 years), describes 3-dimensional protein structure; H.E. SCHULTZE and J.F. HEREMANS, *Molecular Biology of Human Proteins, 2* vol. (1966); L.P. CAWLEY, *Electrophoresis and Immunoelectrophoresis* (1969), a practical manual; E.M. SLAYTER, *Optical Methods in Biology* (1970); STANLEY BLACKBURN, *Amino Acid Determinations* (1968), *a* manual of method.

Progress reports: Advances in Protein Chemistry (annual); *Annual Review of Biochemistry* (annual); *Amino Acids, Peptides and Proteins* (annual); A. NIEDERWIESER and G. PATAKI (eds.), *New Techniques in Amino Acid, Peptide and Protein Analysis* (1971), *a* report on recent advances; B. WEINSTEIN (ed.), *Chemistry and Biochemistry of Amino Acids, Peptides and Proteins* (1971), articles on recent advances; IUB, *Enzyme Nomenclature* (1965).

(F.Ha.)

Protestantism

Protestantism, characterized by its belief in the doctrine of justification by grace through faith, the authority of the Bible, and the universal priesthood of believers, is (along with Roman Catholicism and Eastern Orthodoxy) one of the three major branches of Christianity.

This article is divided into the following sections:

NATURE AND SIGNIFICANCE

The origin and meaning of the term Protestant. The name Protestant was given to a collection of Christian parties at the Diet of Speyer, an assembly of the Holy Roman Empire, in Germany in 1529. It has since been applied to virtually all non-Roman Catholic Western Christianity. Though such a negative definition ("non-Roman Catholic") may be emotionally unsatisfying to Protestants, and though it may impart almost no clue as to the doctrinal substance of this branch of Christianity, it is perhaps the only terminology that is both inclusive and accurate.

Principle of protest

In theory, Protestantism has stood for a principle of protest that calls under judgment not only the beliefs and institutions of others but also one's own movements and causes. On those grounds, however, most students of Protestantism would recognize that the Protestant tradition has not been substantially more successful than have other faiths at remaining self-critical or at rising above institutional self-defensiveness.

Within the spectrum of non-Roman Catholic Western Christianity, a great variety of doctrinal views and polities have been expressed. Not all Western non-Roman Catholic Christians have been ready to be included in Protestantism. Some Anglicans and Lutherans, for instance, have been so eager to stress their continuity with the historic Roman Catholic Church and their distance from extreme Protestantism that they have asked for separate designations. Courtesy suggests that such appeals be taken seriously; however, ultimately habits of speech and sociological usage tend to predominate and, despite their protestations, these groups are usually included in the Protestant cluster.

Historical and cultural importance. However difficult it may be to define Protestantism, there are no problems in indicating its world-historical importance. The rise of the Reformers in the 16th century meant the beginning of the end of a unified religious force in the Western world. Never again could the civil and religious realms cohere on a transnational basis. The ancient spiritual centre of the culture had been criticized and questioned, and Western men could never again so easily unite on values as they had before the rise of Protestantism.

Hundreds of millions of people have found in Protestantism a means to express their faith. Whole populations (in Germany, Scandinavia, The Netherlands, the British Isles, and early America) have been termed "Protestant" by demographers. From these bases, Protestant culture has been transported, along with Western technology, capitalism, democratic government, and educational ideals, throughout the world.

The problem of an "essential" Protestantism

Those who argue that there is an "essential" Protestantism usually stress its witness to the gracious justification of man by God or to Protestant reverence for and submission to the Bible. Others point to its stress and enlargement of the role of the laity. Some parts of each of these teachings, however, are adhered to by many Roman Catholics and denied by many Protestants. It seems inadvisable, therefore, to speak of an essential Protestantism and instead more suitable to seek for unifying strands, positions of consensus, and varieties of historical manifestations.

HISTORY

The heritage of the magisterial Reformers (classical Protestantism). Protestantism is older than the Diet of Speyer. It has often been traced to some late medieval reformist elements in the Roman Catholic Church, among them the Hussites (followers of Jan Hus, 1371–1415) in Bohemia, the Wycliffites and Lollards (followers of John Wycliffe, died 1384) in England, and the followers of Girolamo Savonarola (1452–98) in Italy. These reformist beginnings were consolidated and advanced in the heart of the European continent in the second decade of the 16th century, when the Catholic monk Martin Luther (1483–1546) experienced a rediscovery of grace and began to propagate his teachings at the expense of his and others' loyalty to the existing Catholicism.

Almost simultaneously, Protestantism took rise in Switzerland, first under the efforts of Huldrych Zwingli (1484–1531) and his colleagues and later through John Calvin (1509–64). From there, Reformed Christianity spread to the Lowlands and to Scotland. In England, during the reign of Henry VIII (1509–47), submission to the pope was rejected, and the Church of England was born as still another manifestation of Protestantism, under the tutelage of Thomas Cranmer (1489–1556) and others. These men and hundreds of almost as well-known associates helped spread the Protestant movement into northern and central Europe and, a century later, to North America.

Rejection of the medieval Catholic system

At the heart of earliest Protestantism was a rejection of the medieval Catholic system, wherein men were regarded as having been brought back to God in part through their own merits and efforts and also through the mediation of a sacramental and clerical system. The Protestants wanted to see men free of that system, more immediately responsible to God, and freer to experience grace. The conservative Reformers at first trembled at what they were doing, fully aware of the personal risks involved as well as conscious of what they were doing to the fabric of social life and of inherited theology and piety. They believed that compensatory values were present, however: truth would be served even if old vows of obedience had to be broken.

These conservative Reformers, sometimes called "magisterial" (established), were often reluctant to push ahead. They engaged in prolonged efforts to make their early efforts intelligible to Rome and seemed genuinely reluctant to spawn separate movements in their own names. One by one, however, they were excommunicated or made uncomfortable in Roman Catholicism, or they saw that their doctrinal or ecclesiastical positions drove them increasingly from its sphere of influence. The Protestant movement thus evolved into groups of churches, often organized according to the outlines on the map of the modern nations that were beginning to be formed in that period.

Variations in forms of church government

Without recent precedent for their acts, they turned to the Bible for guidance about church governance. Some of them, the Anglicans in particular, retained bishops as part of the "apostolic succession" and thus shared some Roman Catholic views of church order. Lutherans ordinarily retained bishops, but on differing rationales, without always claiming the succession or arguing that episcopacy was essential to the life of the church. On Calvinist Reformed soil, the presbyterian order, also with biblical precedent, was established; in it, lay elders helped govern, and their "synods" allowed for common action between congregations. Still other Protestants, including most of those who belonged to the radical Reformation parties (*e.g.*, Anabaptists), considered that the only way to keep a pure church was to follow what they regarded to be the New Testament policy of having local control in congregations. The result was a crazy-quilt pattern of Protestant views of church authority, and non-Roman Catholic Christianity in the West was never able to resolve its differences or to find the kind of unity Roman Catholics had under the pope.

The heritage of the radical Reformers ("left-wing" Protestantism). Repudiation of the Roman Catholic sacramental system did not mean that Protestants would do without sacraments, but they never agreed on their meaning. This was typical of a fundamental problem that haunted the evangelical, or Protestant, parties: how sweeping should reform be? Some Reformers, psychologically attuned to people's love of tradition and habit, wanted to be moderate and to save everything they had known except that which manifestly stood in the path of the Gospel. The radical Reformers believed that the church of their day was grossly apostate and could not be rescued; instead, there was to be restoration or restitution of original and primitive (New Testament) Christianity.

The question of the extent of reform led to persecution within Protestantism, for the conservative parties tended to be established by the respective states and were embarrassed by their radical cousins as much as Catholic Christians were. They viewed the radicals as threats to both the civil order and their own ministries, and they made every kind of effort, including persecution and the making of martyrs, to bring the radical sectarians into line. Their efforts, however, were without consistent success.

Theological and social movements in Europe and the Americas that modified the Reformation heritage. Original Protestantism went through many developments. It is possible, however, to see considerable coherence from denomination to denomination on an international basis through more than four and a half centuries. To say that is to note that the Protestant witness is often shaped by "the spirit of the times" and that some common problems produced similar solutions that transcended confessional or creedal differences. A chronological survey of Lutheranism, then, would find many of the same kinds of forces present in differing epochs as would a comparable survey of Calvinism or Anglicanism. While it will be important through most of this article to concentrate on the original definitions and early transformations, a survey of these epochs or episodes will provide background for understanding Protestant development.

Influence of "the spirit of the times"

Reformation social movements. During the origins of Protestantism in the Reformation period there was an outbreak of reform and experiment. Sociologists of religion regularly note that the first generation of a movement often relies on the sometimes erratic, exciting words and actions of charismatic prophets who cannot give stability or provide settlement. In the case of Protestantism, this meant years of at first timid and then violent rejection of the church, of efforts at common work that were soon frustrated in competition of Lutherans versus Reformed, Continent versus England, and other rivalries. This was a period of polemics, of occasional vandalism and image burning, of called-for and aborted revolution (*e.g.*, the Peasants' Revolt of 1524–25 in Germany), and of a full range of options from a conservative tampering with liturgy to claims that new messiahs with special revelations were present.

By the middle of the 16th century there were attempts at political settlements of religious problems. Wherever in Europe the Reformation eventually prevailed, the classical Reformation churches (*i.e.*, Lutheran, Anglican, Reformed) early became well established, but some minor religious wars and some truces (*e.g.*, the Peace of Augsburg in 1555 or the Elizabethan settlement, beginning in 1559) were necessary before permanent forms could be recognized. The accent later fell on regularizing church order, developing hymns and liturgies, preserving the writings of the various Protestant founders, and learning responsibility for churchmanship.

Periods of Protestant orthodoxy, Puritanism, and Pietism. The end of the 16th century and the first half of the 17th was a period of another interterritorial and transconfessional phenomenon, a time of doctrinal definition. Even the radical Reformers, who were sometimes anti-intellectual in outlook, found themselves impelled to articulate their dogmatic positions. To a greater extent, men with a passion for orthodoxy and a love of intellectualization used the Protestant-dominated universities of northwestern Europe as a base for formulating creedal positions. As authors of huge, multivolume works of doctrine, they prided themselves on their ability to relate their Reformers' positions to the various philosophies then prevailing.

Intellectualization and piety

By the end of the 17th century in almost every Protestant region, reaction had set in. In England, Puritanism carried the Reformation further; it passed through a time of intellectualization toward a new kind of piety that often verged on mysticism. On Reformed and Lutheran soil the movement known as Pietism served as a counterpart to this phase of Puritanism. The Pietists, typified by the German Philipp Jakob Spener (1635–1705), were not ill at ease in intellectual circles. They even established universities of their own. But the Pietists undervalued intellect and made more of religious devotion and passion. Not eager to upset the whole established order, they did want to vivify it. Consequently they developed smaller, more intimate circles for prayer and worship. They learned to speak to each other of their hearts' concerns. They traded on the doctrine of the orthodox period but wanted to help produce new works of love. Significantly, they helped initiate the period of Protestant missionary and benevolent activities.

Rationalist and Evangelical Protestantism. Just as Pietism appeared almost wherever Protestantism did, so did its heir, its successor, and its competitor, Rationalist Protestantism. In Germany, men like the philosophers Gottfried Wilhelm Leibniz (1646–1716) and Gotthold Ephraim Lessing (1729–81) drew on Protestant resources. In The Netherlands, Hugo Grotius (1583–1645) had incorporated "natural law" philosophies into his Christian system. In England, the moderate Deists related positively to some forms of witness, and Anglicanism produced latitudinarian clerics---men of an Arminian persuasion (*i.e.*, adherents of the views of the liberal Dutch Calvinist Jacobus Arminius) who stressed the benevolence of the deity, the capability of man, the importance of morality and reasonableness, and the need for tolerance. Many of them occupied Anglican pulpits and even bishoprics, or taught in theological faculties. From the Pietist era they carried over a view that minimized doctrinal differences, but they were now much more interested in the mental and moral aspects of man's life than they were in his piety.

That late 18th-century form of Protestantism may have been more or less restricted to its clerical and theological leadership, but its viewpoint seeped into the broader streams of the laity. The subsequent Protestant phase, often called "evangelicalism," was a very widespread popular movement, attracting clergy and laymen of all Protestant communions. Evangelicalism reacted against the elitism of rationalism and the intellectualism of both orthodoxy and the Enlightenment.

Emphases of evangelicalism

Generally conservative in its theological position, evangelicalism emphasized doctrines concerning the person and work of Jesus Christ, particularly those dealing with his sacrificial death for man's experience of a relation in grace to God and his motive for Christian living. From Pietism it absorbed paradigms for humanitarian, benevolent, and missionary activity. This was the form of Protestantism that, beginning in Germany in the late 18th century, England in the 1790s, and the U.S. in the 1820s, went literally "into all the world" with its message. It was impelled by the new techniques of revivalism and gave birth to a new hymnody. The genius of the era was the British Anglican who came to lead Methodism, John Wesley (1703–91). Despite all the assaults of modernity, this movement lived on through the 19th century and has much influence even today.

Ecumenism ***and encounter with secularism.*** In the 20th century, however, the rise of still another international Protestant movement came into being, one which affected all the divided bodies. This was the tendency toward reunion, first of Protestant and later of all Christian churches, a tendency that was channelled into the ecumenical movement. In part the result of disaffection by Protestants with the undercutting of each other's mission, the ecumenical movement produced organizations such as the World Council of Churches and various national councils of churches and helped each confession (*e.g.*, Anglican, Baptist, Congregational, Lutheran) set its own house in order.

Goals of desacralization

At the heart of the Protestant world of thought throughout four and a half centuries, and especially in the recent past, has been the effort to come to terms with the complex process of secularization, which has sometimes been seen to be concurrent with Protestantism and, for many, to be in part the result of the Protestant impulse. The Protestant wanted to see the world disenchanted, or deprived of its magic and mystery, so that man under God could dominate creation and produce effects that would serve man.

It was the Protestant who often wanted to desacralize politics, which meant that he was ready to see the eventual end of the "divine right of kings" and was willing, as in the United States, to help write constitutions or charters for nations without employing positive references to deity. He did this not because he believed that God had nothing to do with the political order but because he had come to believe that coercive governments could not properly impose religious views or that religion was most healthy when it did not live off the support of the civil ruler.

These disenchanting and desacralizing movements evolved into a force that, much like Protestantism, claimed to stand in the biblical prophetic tradition. From there it learned to bring judgment on rulers, to counter ecclesiastical authorities, and to work critically for an improved world. The benefits of the secular movements have been balanced by liabilities and limitations, however. Secularization has led many persons to remove themselves from the sphere of Protestant interpretations of life, to drift from or to rebel against a witness to God, as Protestants have "traditionally" propagated this witness, or to lose themselves in worldly concerns that bring them beyond easy reach of Christian witness. The relationship of Protestantism to a secular world no doubt will remain a central issue in the future, and the terms for the relations will probably be set by the original witness of Protestantism in the 16th century.

TEACHING. WORSHIP, AND ORGANIZATION

Common principles, ideas, and practices of the magisterial Reformers and their successors. ***Justification by grace through faith.*** The original Protestant leaders united in their contention that what separated them from the Roman Catholicism of their day was their teaching that man is justified by grace through faith. Devotion to this teaching has been central to Protestantism throughout its history. Although there have been subtle variations in the differing Protestant church bodies, a central core of shared belief was at first easily discernible.

Concern for "justification"

Concern for "justification" was related to the obsession that the 16th-century person often expressed in terms of finding himself on good terms with God. He drew his metaphors from the courts of law. Aware of his shortcomings, his ignorance, his sin, and his guilt, he saw himself standing before a bar of justice presided over by God. Without help, he could expect nothing but God's wrath and condemnation. This meant that he would perish everlastingly, and his present life would be full of torment. Yet the Bible also presented him with a picture of a loving and gracious God, who may very well desire happiness for him. The question was: how could he be sure that God would reveal his gracious, and not his wrathful, side? How could he have the confidence that he was included in the positive loving action of God?

The teaching of the Reformers becomes most intelligible when seen over against the Western Catholic doctrines (*e.g.*, sin, grace, atonement), as they saw them. In the Protestant view, the late medieval Catholic teaching held that man was brought back to God only when so much grace had been infused into his soul that he merited the favour of God. God could not have been expected to accept someone who was unacceptable, but he could impart something that would make man acceptable. This something was grace, and its flow depended upon the merits of God's perfect Son, the man Christ Jesus. The church, according to medieval Catholicism, in a sense controlled the flow through the sacramental system and through its hierarchy. To the Reformers, the Roman Catholic sacramental system seemed to be part of a transaction that was always going on between man and God. In it, men made sacrifices designed to appease and please God. They would attend the mass, bring offerings, show sorrow, do penance—which might involve self-punishment or compensatory good works—until God would be gracious. The leaders of the church, from priests through bishops and popes, mediated the transaction. The Reformers believed that such an arrangement could easily be misused as a political instrument for forcing rulers to comply with the church's wishes and as a personal instrument for keeping people in uncertainty or terror. It was this vision of Catholicism that helped inspire the Protestant leadership to rebel and to define justification in other terms.

Influence of the writings of St. Paul

The terms for this Protestant teaching came from the Bible, especially from the New Testament, and even more so from the writings of St. Paul. In St. Paul they saw a religious hero and thinker who had endured a spiritual quest similar to their own. He could be described as having been brought up in a legalistic version of Judaism, a system in which he was constantly striving to please God by following his Law, particularly as set forth in the Old Testament through the Ten Commandments. Yet, Paul failed and was assailed by doubts about his worthiness and his salvation. His conversion meant a radical turning and a free acceptance of God's favour "in Christ." This meant that in faith man could be so identified with Jesus Christ that when God looked at him, he saw instead the merit that Christ had won through his self-sacrifice on the cross. God looked, in short, at the sinner; but he did not see the sinner. He saw his perfect Son. So he could declare the man righteous; he could justify him--even though the man was still a sinner.

When taken out of the historical context of St. Paul's teachings in the letters to the Romans or the Galatians and transferred to their own times, the Reformers' teaching of justification relied heavily on the work of the Holy Spirit. The Holy Spirit, in effect, made Christ's action contemporaneous with the sinner's quest. God was working now on behalf of the man in need. Through the preaching of the Word of God, man was exposed to the story of Jesus Christ's sacrifice and death. If he believed this historical narrative and, more importantly, if by the power of the Holy Spirit he believed that it was told and enacted for him, he stood before God in a new light. He did not have grace infused into him to the point that he became acceptable and pleasing to God. Instead, while man was still a sinner, God accepted him favourably and justified him. Christ's death on the cross was then the only "transaction" that mattered between God and man. The sacraments reinforced the relation and brought new grace, but no pretense was made that the human subject had achieved satisfaction before God or produced enough merit to inspire God to act.

Man's freedom and the problem of "good works"

In the Reformers' view, the new situation was one of freedom. Whereas Catholic man constantly stood in fear as to whether he had provided enough merits, had achieved enough good works, or had pleased the church as God's bargaining agent, the Reformers' version had man standing before God completely freed of these nagging questions. He was liberated both from the terrors of sin, death, and the devil, on one hand, and, on the other hand, from the enslaving pride that went with men's belief that they had achieved or at least had substantially cooperated in their own salvation.

This left the Reformers with a serious question, one to which their Roman Catholic opponents regularly referred. What had happened, in this teaching of justification and freedom, to the biblical accent on good works? Jesus himself, in the Synoptic Gospels (Matthew, Mark, and Luke), was constantly preoccupied with the effort of making men better, of having them bring forth "good fruit" and good works. Even Paul shared such concerns. Had the Protestant movement slighted these concerns in its desire to free man from the necessity of merits and good works?

The literature of Protestantism is rich in its expression of answers to such problems and questions. The Reformers were virtually unanimous: good works did not produce appeasement of God or salvation, yet they inevitably flowed from the forgiven heart and were always the consequence of the justified person's life. The Law of God could never be used as the saving path along which man walked, as a sort of obstacle path or road map to God. Instead, the Law of God measured man's shortcomings and judged him. A gracious God acting through his Gospel brought man back to him.

Man as simultaneously righteous and sinful

The Reformers' vision of man implied in such teachings was doublesided. They believed that from God's point of view, the justified man was so identified with Jesus Christ that he shared Christ's perfection. The same man, throughout all his life, left to his own devices or when seen by God apart from Christ's sacrificial work, remained a sinner of the worst kind. The difference came through God's gracious initiative; nothing that man did started the process of his justification. To the eyes of many in subsequent generations, the result was an apparently pessimistic and gloomy view of man's potential in Protestantism. His will was bound, apart from God's loving activity. He had no merits or good works that would satisfy God. Sometimes the phrase total depravity was used to describe his condition, though it must be said that the term had connotations in the 16th century that were different from those that it has today. It was used not so much to provide lurid connotations for descriptions of the depth of his sin but rather to describe its extent; man as a total being was in trouble. Even his good works, his piety, his religiousness, and his efforts, apart from justification by grace through faith, fell under God's curse. On the other hand, the justified sinner could be described in the most lavish terms, as one who could be "as Christ" or even sometimes "a Christ."

Those who have heard this Protestant teaching outlined through the centuries have regularly seen the difficulties it raises insofar as the portrait of God's character is concerned. Protestants never came up with logically satisfying answers to the resultant questions, though they were convinced that they were faithful witnesses to biblical teachings concerning the mystery of God's nature. The central question: if everything depended upon God's initiative and yet the majority of men are not saved, does this not mean that God is responsible for creating men only to have them suffer; is he not guilty of the worst kind of cruelty by being the sole agent of their damnation?

In facing the question, Protestant leaders differed slightly from each other. Some said that whenever men were saved, it was to God's credit; whenever they were lost, it was through their own fault. They were free to hear the Word; they were free to respond and accept the gift of grace in Christ; their own hardness of heart kept them from freedom and new life. Others ran the risk of presenting cruel pictures of God's nature and action, in their interest to witness to his sovereignty and initiative. The view that God predestined some men to be saved and others to be damned was called "double predestination." Some theologians argued that God did this predestining before men fell into sin; others saw it as a new act of God consequent upon man's fall. Those Protestant parties that were generally non-Calvinist in outlook were usually less systematic and less logical in their statements. The non-Calvinists taught a doctrine called "single predestination." They shared the Calvinists' affirmation of God's total responsibility for man's salvation; but they tended to be silent or to relegate to the area of mystery and unanswerable questions, the issue of how God could then be other than responsible for man's damnation. In general, the Protestants saw themselves to be more successful at preserving the teaching of God's sovereignty and the corollary of human helplessness than they were at making his character attractive to all. They saw themselves overcoming this problem in biblical terms by a stress on his loving relation to men in sending his own Son, Jesus Christ, to suffer for them.

Predestination

*The "priesthood of **all** believers."* If the teaching of justification had important consequences for the doctrines of God and of man in Protestantism, it had equal import for any statement of the meaning of the church and especially of clergy–laity relations. The medieval system, sacramental and hierarchical, in effect gave the priests a monopoly in monitoring the transaction between God and man. The Protestant teaching of justification broke this down and the Protestant leaders reverted to what they held to be the biblical view, that all believers have a share in spreading the word of grace and the acts of forgiveness. The result was an emphasis that stressed not the privileges of a priestly caste but rather "the priesthood of all believers."

The Reformers viewed this teaching as being based on the free-flowing sense of authority that existed between Christ and his Apostles, who had been pictured in the Gospels as being active apart from an elaborate clerical church order. At the same time, they believed that their doctrine would effectively displace the Roman Catholic hierarchical thought and action. Now all men were to be enjoined to take the responsibility for each other's salvation; any Christian man could represent the needs of all others before God. Originally the priesthood of all believers was basically an enlargement of the view that all Christians had intercessory powers, that they could all pray for one another. But it came to refer to the whole Protestant view of an equality of status between clergy and laymen and to the common calling of all Christians to be agents of God's Word and grace.

Priesthood and coresponsibility

The affirmation of the priesthood of all believers had widespread implications in society. In Protestant areas and nations the privileges of the clergy were limited and the scope of lay activity enlarged. All men shared a "vocation" (calling), and priestly vocations were not considered to be more meritorious or nobler than lay vocations. Monastic vocations were almost entirely swept away, and restorations of the monastic ideal have been rare and exceptional in Protestant history. Protestants kept, for the most part, a rite of ordination (though some Anabaptists dispensed with all acts that seemed to imply separation between a ministry of ordained persons and laymen) but did not regularly view it as a sacrament. That is to say, ordination conferred no special grace on men. In part, a ministry was kept on a pragmatic basis; the clergy were to tend to the business of studying and preaching the Word, properly administering the sacraments, and caring professionally for the health of the church. A set-aside ministry was also derived from biblical precedent in the Book of Acts and early Christian letters.

Protestants, while giving lip service to the equality of laymen and clerics in the priesthood of all believers, have not always seen themselves to be so successful in clarifying the laity's role. In most cases, laymen were not to be the preachers in public worship, and administration of the sacraments usually remained in clerical hands. By demanding of preachers expertise at expounding the Bible, Protestants often have made educational requirements a basis for ordained ministry, at the expense of a full lay involvement. Yet their views did greatly enhance both the theological and practical status of laymen, when contrasted to the situation in medieval Catholicism.

If all believers were priests, then no single church could monopolize the mediation of grace, since Protestants saw that there were believers in all churches, Roman Catholic and Protestant, Lutheran or Calvinist or Anglican. As a result, the teachings inherited from medieval Catholicism about the visible and the invisible church were called into question. To many Reformers, most notably Luther, the church was always visible because it was made up of

The visible and invisible church

people. But its limits and borders were invisible since one could not examine the heart of others to see exactly who were the true believers. This inability to define the boundaries of the church led other Reformers, among them Calvin, to continue to employ the distinction between a visible church and an invisible one, the latter referring to the people who were saved, even if they were in churches where full doctrinal purity had not been achieved. Men see the visible, humanly organized church of Christ, but they cannot simply identify this with the Bible's one, holy, catholic, and apostolic church, which is properly discerned only by God and hence invisible to men. The visible church, in the Reformers' view, almost certainly contained a mixture of members of the invisible church, on the one hand, and hypocrites, or false believers, on the other.

The Word of God and the Bible

Authority of the Word. Justification by grace through faith and the priesthood of all believers were affirmations that challenged the inherited Catholic views of authority because they undercut the monopoly of the hierarchy in the system of grace. Downgrading the medieval system of authority left a vacuum that Protestantism hastened to fill. Full of variety and pluralism as the movement was from the first, it was rarely characterized by a love of anarchy or indiscipline, and the Reformers set to work at once to establish the locale and extent of authority in the church and the believer's life. Almost unanimously, they saw final authority to reside in the Word of God, which tended in the minds of many to be simply equated with the Bible. The Protestant movement's need for authority enhanced its view of Scripture just as, one might argue, the rediscovery of scriptural teaching was seen to be the impetus behind the Protestant movement.

Later generations of Protestant thinkers sometimes resorted to scholastic philosophical definitions similar to those of the medieval Catholic theologians; in such definitions, justification became the material (or substantive) principle of the Reformation, while the matter of scriptural authority became the equally important formal (or structural) principle. In some epochs, debate about the nature of the Word of God or the Bible was even more preoccupying than was discussion of justification. Protestants often have portrayed medieval Catholicism as being a nonbiblical or even an antibiblical faith, one that denied the Bible to the laity. The expense of reproducing manuscripts led many libraries to chain books to the wall, and the Bible chained to the wall entered Protestant mythology as a symbol of the denial of lay access to the Bible in Roman Catholicism. In many circles, Protestantism has been celebrated as a religion of the "open Bible" in opposition to the closed book of Catholicism.

Interpretation of the Bible

Mythology aside, Protestants without exception concentrated on biblical teaching, and this led to a new passion for translating the Bible into the vernacular and disseminating it as widely as possible, an activity aided by the almost simultaneous invention of movable type and the resultant progress in printing technology. It was to be put into the hands of as many ministers and laymen as possible. Thus, they also had to be taught to read, and the Protestant movement claims some credit for hastening the modern impetus toward the ideal of universal literacy. While the Bible was ordinarily read in the churches and interpretation was shaped by the old and new traditions of these churches (Anglicans read the Bible's teachings on apostolic succession in a way different from that of Anabaptists, for example), what came to be called "the right of private judgment" was often exalted.

While Protestants could agree that the Word of God was authoritative in matters of faith and that the Bible as the book inspired by the Holy Spirit had unique status, they did not agree on all interpretations of the Scripture, nor did they unite in a single doctrine of scriptural authority. The Anabaptists, and later the Quakers, stressed an immediate experience of God and thus tended to qualify the impact of the Bible in shaping Christian life. But even among the Lutherans, Calvinists, and Anglicans there were differences of opinion about the Bible.

Latter-day Protestant Fundamentalists have argued that scriptural authority, if it is to be believable and if it is indeed the Word of an all-knowing, perfect God, dare not be described as including errors, even errors in historical or geographical detail. They have backed up this contention with citations from John Calvin and Martin Luther. At the same time, subsequent Protestant contenders for a more open and critical attitude toward biblical literature are equally capable of citing the Reformers — Luther, for example. Luther often spoke of the Bible as "the cradle in which Christ lies" and as a book that derives its authority from Christ and not from an a priori (presumed) doctrine of inerrant inspiration. So he could quite freely relegate some books of the Bible to secondary status or criticize the argument of others, even as he could complain of Paul's grammar or illogical argumentation, and point to errors of detail. Part of the confusion results from the fact that the Reformers in their era did not see questions regarding truth in the same terms as they have been recognized in the later scientific world, and, also, they could accept the full and governing authority of the Bible without elaborating theories concerning its perfection in detail.

De-emphasis of institutional bases of authority

The unquestionable elevation of the Bible as the authority in matters of faith led to a corollary downgrading of other authorities in the church. The hierarchy, and especially, the pope, were hardest hit, and papal authority was denied in almost every sector of Protestantism. More regard was paid, at least by conservative reformers, to the Fathers (doctrinal teachers and interpreters) of the early church, who were sometimes cited in the confessional writings of the various Protestant bodies. The Fathers were revered as guides rather than as final authorities. Similarly, a critical attitude toward councils of the church came to prevail. On the one hand, it was widely asserted that councils can and do err, and historical studies pointed to contradictions between various conciliar statements and to unsubstantiated assertions by past councils. On the other hand, many formulas and creedal statements of the ecumenical councils, particularly as these referred to the Trinity or to the Person and work of Jesus Christ, were highly regarded, and many Protestant churches took the chief creedal statements of past councils into their own official body of teaching.

Canon law, which was the inherited body of legal materials that regulated faith and morals, quite naturally also suffered because of the high regard for the Bible. In most Protestant circles, it was difficult to make legislation binding upon conscience unless it was based on clearly affirmed biblical legal teaching; more important, accent on the Gospel of grace led most Protestants to want to undervalue the whole role of law in the life of the church. At the same time, new church order soon developed, and it must be said that Protestants often acted as legalistically as did the Roman Catholics, whom they were repudiating. Most Protestant bodies, notably the Anglicans, developed their own versions of canon law or rules of church older and discipline.

Concern for further reformation

The ongoing reformation of the church. The church that is to be judged not by the pope but by a normative Bible, grounded in a priesthood of all believers, and critically affirmative of Church Fathers and councils and that rejects inherited codes of canon law differed vastly from medieval Catholicism. In few respects did it differ more than in its establishment of the principle of an ongoing reformation. While most of the Reformers, once established, tended practically to resist extensions of reformation that would jeopardize their status and definition, almost all Protestants, at least nominally, assented to the idea that "ecclesia reformata semper reformanda"—*i.e.*, that the church was always reformed and always in need of further reformation. The Protestant movement, then, was conceived as an unfinished product, constantly to be judged by a reading of the Bible, its polity continually subject to debate, its policy open to ongoing appraisal and change. It was in that climate that the sacramental teaching of Protestant churches was argued and developed.

Emphasis on *the* sacraments. On few points did Protestants disagree more than in their interpretation of the sacraments, but they did unite in their rejection of some aspects of Roman Catholic teaching. Since attack on the sacramental-hierarchical system of salvation was at the heart of their reform, almost nothing of it survived intact. In place of the churchly system, the new accent fell on the limitation of sacramental teaching to those acts clearly commanded by Christ and connected with his promise in the Scriptures. One can argue that, since "sacrament" was not a biblical term, the debate had to do simply with definitions. Most Protestants define sacraments, then, as acts that impart grace and the new life. They must combine the Word of God and some visible means (like bread, wine, and water); they must have been established by God and instituted by Christ. On these terms, five of the seven chief Roman Catholic sacraments failed to meet the definitional tests: marriage, ordination, confirmation, penance (now called repentance), and extreme unction (now called anointing of the sick). Not in every case did Protestants abolish these acts from their rites, but they ruled them out of consideration as sacraments. Thus the Protestant teaching on marriage was normally as "high" as Catholic doctrine, and it may be considered quasi-sacramental. But it was seen chiefly as a civil act blessed by the church, and it did not convey grace to the participants, nor was there a visible "means."

The sacraments commanded by Christ

Though Protestants—with a few exceptions, chiefly Anabaptist and Quaker—had little difficulty limiting the number of sacraments and perpetuating a high regard for those that survived the change in definition, they were far apart in their understandings of what went on in sacramental acts. Basically, three views were debated. To the "right," as one might call it, was the Lutheran view, which critics saw to be quite close to Roman Catholicism. Luther seemed to bring with him something of a medieval world-view, in which symbols of the material world were transparent to another invisible, divine order. This made it possible for him to make much of the material objects in the sacraments. When he connected them with biblical words, he was able to say of bread and wine that these are the body and blood of Christ, and of Baptism, that it effected a change in the believers' status before God.

At the "left" was the view of Huldrych Zwingli and other Swiss Reformers, who accented the spiritual side and downgraded the material. In some respects, they shared more of a modern view of matter and spirit, in which the symbols were opaque, disengaged from an invisible "other order." Such teaching meant that what mattered most in the sacraments was the following of Christ's commands, the reminiscence of his participation in the world of his disciples, and the spiritual intentions brought to the acts of believers. For Zwingli, the bread and the wine merely represented the body and blood of Christ, and Baptism was more a sign of a Covenant with God than a supernatural imparter of grace. Between the Lutheran and Zwinglian views were Calvinist and Anglican attitudes and definitions. All Reformers agreed in their criticism of the Catholic teaching called "transubstantiation," which held that the actual "substance" of the bread and wine in the Lord's Supper was turned into the body and blood of Christ. But they did not agree over the alternatives to that teaching, and debate over the sacrament did as much as any other theological factor to contribute to internal Protestant division.

Relationship between the community of the baptized and the political community. Equally varied were the attitudes toward civil authority among the various Protestant parties. Martin Luther expressed what in theory could have been a most radical theological view of the separation of civil and religious realms through his doctrine of "the two kingdoms." He could reduce his teaching virtually to an aphorism: God's Gospel ruled in the churchly realm and his Law ruled in the civil society. To rule the church by the Law or the civil realm by the Gospel would be to bring legalism to the sphere of grace and sentimentalism into the orbit of justice and thus dethrone God and enthrone Satan. In practice, however, the Lutheran Reformation worked to keep its ties to the civil order and was the established religion wherever it predominated in Germany and Scandinavia. In many territories, princes actually took on the superintending roles that bishops had known in Roman Catholicism.

Separation of church and state

John Calvin made less of a theoretical effort to separate civil and religious realms. For him, Geneva was to be a theocracy in which the saints would rule. God's covenanted community was to be based on his Law, as revealed in the Scripture. No detail of civil or community life was too remote, too secular, or too petty to escape inclusion by the Calvinists in the ecclesiastical sphere of supervision or regulation. Huldrych Zwingli taught a variation of this version, one that asked the Christian to be a zealot or patriot in the civil society—a teaching that he confirmed with his blood, for he was killed in battle in 1531. In the Anglican approach there was also no attempt to separate the civil and religious realms; in England the church was given the mandate to press conscientious matters upon the king and other civil authorities. These established Protestant views were to be subverted or countered by radical Reformers who did want a separation from civil spheres. These views were also constantly revised in later Protestant history, with the rise of the modern secular state.

Modes of expression of the ideas of the magisterial Reformers and their successors. Protestantism was forced to find means to propagate and sustain itself through time. Reformers had removed many of the inherited props or means and developed, within a century, parallel structures of most of those that had been repudiated along with Roman Catholicism. Lacking papal authority, canon law, and "international" connection with civil authority (as there had been in the old Holy Roman Empire), along with the binding power of church councils or a single philosophy on the basis of which to argue their case, they came up with alternatives or surrogates for most of these, though the new systems were more varied than the at least nominally homogeneous Catholic skein.

Most notable among the structural necessities was the formulation of "confessions," or creedal statements—and Protestants met frequently and regularly to write them—by which they could define their positions for the benefit of their adherents and their opponents. The Lutheran Augsburg Confession (1530), Reformed documents such as the Second Helvetic Confession (1566) and the Westminster Confession (1646), Anglican affirmations such as the Thirty-nine Articles (1563), and Anabaptist confessions such as that of Dordrecht (1632), all gave evidence of the Protestant impulse to define their positions. The Protestant leaders recognized that their movement could not long exist or continue with the fervour and ferment of first-generational impulses.

Protestant confessions

Confessions of the church appealed to the minds of theologians, the administrative passions of leaders, and the legalistic spirit of those who would impose them as doctrinal standards, but they did not warm believers' hearts. Thus Protestant leaders had to concern themselves with the affective side of church life in order to hold the attention of masses of people and to give them opportunity to express their faith and life in God. The chief instruments to achieve these aims were liturgies and hymns. The inherited liturgies included much of the Roman Catholic sacramental teaching. As such, they were given over too much to an accent on the sacrificial character of the mass and thus had to be purged. Conservative Reformers retained the shell, or outline, at least, of these formulas for worship, though they took great pains to bring both these outlines and the nuances of expression into line with what they considered to be a more evangelical teaching. Since worship is perhaps the chief public expression of gathered Christians, all Reformers had to give attention to its detail.

Liturgies and hymns

Martin Luther initiated the process with his ***Formula Missae*** ("Formula of the Mass") of 1523, a service which retained the Latin language; but he soon devised (in 1526) a ***Deutsche Messe*** ("German Mass"), ***a*** vernacular and folk expression of greater informality. At about the same time, Huldrych Zwingli was producing a Reformed order with two liturgies for the Word and the Lord's Supper in 1525, soon to be followed by Martin Bucer's (1491–1551) work on Psalms and church practice in 1539 and John Calvin's ***Form of Church Prayers in*** 1542 and 1545. Across the English Channel, the Anglicans were preserving stately forms of worship used in subsequent centuries, chiefly in ***The Book of Common Prayer*** of 1549 and 1552, and in Scotland John Knox helped formulate Presbyterian worship in ***The Forms of Prayers*** in 1556.

Emphasis on preaching

While Protestant orders were somewhat less ceremonial than the Roman Catholic liturgies they replaced, the human impulse to routinize ceremonies prevailed, and almost everywhere these forms for worship took on a more or less formal character. They differed from Catholicism chiefly in their elevation of the act of preaching the Word of God. Preaching was viewed as the means of grace whereby men were encouraged to repent and accept the grace of God through faith in Christ, just as the sermon was used to shape the community and give guidance for interpretation and action in life. For some, this accent on preaching meant a downgrading of the Lord's Supper; for others, there was to be a parity, with the sacrament providing a necessary parallel means of conveying grace. Communion "in both kinds," with reception of both bread and wine, prevailed (whereas in the Catholicism of the era of the Reformers, the cup was withheld from the laity), and, except in Anabaptist circles, the Catholic practice of infant baptism by means other than total immersion was retained. The Protestants, for the most part, took over existing Roman Catholic church buildings for worship, or they met in academic or civil halls or homes; but as time passed, they also took responsibility for erecting church buildings.

Hymnody played a major role in giving voice to Reformation sentiment, never more successfully than in the activity of Martin Luther, whose "A Mighty Fortress Is Our God" came to be called "the battle hymn of the Reformation." The Genevan Reformation and the Presbyterian churches in general tended to prefer simple and sometimes stolid hymnody in the form of rephrased and parsed psalms, such as ***The Genevan Psalter*** of 1563. The rejection of hymns and attention to sung versions of Scripture also prevailed in early Anglicanism, not so much because of principle but because of the failure of Anglican Reformers to devote themselves to the propagation of their movement through song. The great tradition of Protestant hymn writing developed later, in the 18th and early 19th centuries.

Systematic theologies and dogmatics

Liturgies and hymns appealed to the heart and soul, but Protestants also addressed the mind through an impressive outpouring of works in systematic theology and dogmatics. John Calvin was the supreme systematizer of first and second generation Protestantism, and his ***Institutes of the Christian Religion*** (1536) is a classic on even the shortest shelves of Christian doctrinal literature. Luther was, of course, a first-rate theologian, but he made less effort to be systematic, and his scores of volumes of theology usually grew out of comments on issues that agitated him or inspired or disturbed his movement at any moment. His disciple and colleague Philipp Melanchthon, in the ***Loci Communes*** of 1521, showed more systematic discipline.

In the 17th century the Protestant movement tended toward more rigid doctrinal expressions, as individuals interpreted the confessional statements of the earlier century with an almost fanatic attention to detail. Huge works of Lutheran and Reformed dogmatics poured forth from presses, most of them based on a kind of Protestant reversion to the type of scholastic philosophy that had prevailed in the late medieval period. Leaders in the period of Lutheran orthodoxy were Martin Chemnitz (1522–86) and Johann Gerhard (1582–1637); Reformed Orthodoxy was marked by the scholarship of Theodore Beza (1519–1605) or, in England, men like William Perkins (1558–1602). The ponderous and often lifeless writings of lesser orthodoxists than these were often expressions of internecine Protestant warfare, and the Reformation parties frequently used theology as an instrument of verbal warfare in that period. Debates raged over the sacraments, over the two natures of Christ, over the relations of ecclesiastical and civil realms, and over the part man played in salvation. Almost never did these debates lead to concord, and despite occasional irenic figures, such as Georg Calixtus (1586–1656) or Hugo Grotius, Protestantism was fated to remain divided, at least until the ecumenical movement in the 20th century began to produce new amity and common purpose or assent.

Common principles, ideas, and practices of the radical Reformers and their successors. The interpretation of Protestantism up to this point has been, with only a few noted exceptions, based on the majority view among the 16th-century Protestant movements. No single term adequately covers the Lutheran-Calvinist-Anglican complex, though "magisterial," "establishment," mainline," "conservative," and "classical" have frequently been applied to these movements. Of considerable parallel significance was the Protestant activity of another, and even more complicated, cluster of movements, for which also no single term can be agreed upon. Some historians speak of "the radical" Reformation or "the left-wing of the Reformation"; others have concentrated on components, such as the Anabaptist-sectarian or the spiritual-mystic or the rationalist-unitarian versions. In almost every case, these were the expressions of the economically and socially deprived classes in the 16th-century societies, though their latter-day heirs have sometimes known or sought the favour of civil authority and social arbiters.

The "radical" Reformation was radical in that it deliberately chose to repudiate as much as possible of traditional Roman Catholicism, in various acts of "restitution" of what it held to be the obscured and eclipsed but true original apostolic church. The "conservative" or "magisterial" Reformation, on the other hand, tended to keep whatever it could of the medieval ecclesiastical tradition and to affirm continuities in the life of the church wherever possible.

Problems of generalizations about the radical reform movements

The varieties of radical expressions are rich and bewildering. They grew in virtually every Protestant land, sometimes as an extension of the logic of the conservative Reformation but more often as original movements bearing a logic of restitution all their own. The radical Reformation also occurred in Catholic territories, such as Italy, where the mainline Protestant movement never knew much success. In Lutheran circles, men like Karlstadt (c. 1477/1481–1541) and Thomas Müntzer (c. 1490–1525) set out, in Luther's prime years, to shatter much of what he wanted to retain and to carry reform in new directions. Debates over the Lord's Supper and Baptism led to new radical movements in Switzerland, southern Germany, and Bohemia-Moravia. In Strassburg, a significant group of radicals, including Kaspar Schwenckfeld (1489–1561), Melchior Hofmann (died 1543/1544), and Sebastian Franck (c. 1499–c. 1542), gathered around 1529. The north of Germany and the Netherlands were havens of early Anabaptism (re-Baptism), and in the southern Netherlands, Menno Simons (c. 1496–1561) spread the movement that has come to be called Mennonitism. In Poland and eastern Europe, the radical Reformation often took spiritualist and unitarian (anti-Trinitarian) turns, as it did in Italy. "Radical reform" was also behind some of the Puritan and separatist movements in England. Because they were by nature competitive, free-formed, and varied, it is difficult to generalize about the radical Reformation movements, but some assertions common to major segments are possible, and the study of these movements is important because of the role they were to play in shaping modern Protestantism, especially as it developed in North America.

The gathered church. The radical Reformers were united in their opposition to established Protestantism's view of ecclesiastical continuity with the church of Christ in every age. The mainstream Reformers were radical in their rejection of what they regarded to be false teaching in the medieval church and almost never had kind words to say about any of its forms. But they did believe that God had kept a body of faithful teachers and respondents through the millennium or so after what they considered to be the "fall" of the church during the closing years of the Roman Empire, and this view of the succession of believers was integral to their doctrine of the church. Just as emphatic was rejection of this view in radical circles. Some radicals were willing and eager to trace a kind of continuity from John the Baptist down to the 16th century, but it was significant that they found virtually every evidence of true faith only in the sectarian movements that had separated themselves from official Roman Catholicism or that were condemned, harassed, and persecuted by Catholics. Among these were the Waldenses (a medieval religious movement espousing voluntary poverty and lay preaching); the Albigenses, also called Cathari (medieval sect espousing dualism and asceticism); some forms of Spiritual Franciscanism (branch of the Franciscan order espousing poverty); and other reform movements of the Middle Ages. Just as often, however, radicals taught that the true church had died not long after Christ and had to be restored as if from the foundation itself.

The "believer's church"

The repudiation of continuity was paralleled by rejection of a tie between the civil and ecclesiastical realms. The bond between these two, in the era of the Roman emperor Constantine (died 337), was viewed by the radical Reformers as the root of the church's fall and later vicissitudes or death. From that experience, it was argued, the church ought to have learned not to let the spiritual infection of political authority prevail nor to permit any one to be regarded as a member of the church without an explicit personal affirmation of faith. In a phrase that is often widely used, the church was to be "the believer's church," made up of assenting and consenting people of decision who chose to respond to God's Covenant. This view appeared in contrast to the view held by those who argued that Baptism of infants, who of course could not make personal decisions, conferred church membership and that, thus, virtually entire populations of territories could be members.

The keystone of the concept of the believer's church is that men voluntarily choose to be members. No one can be coerced into it nor can he become a member automatically, as it were, through a sacramental act. It was on this ground that infant Baptism was condemned by almost all radical Reformers. A result of this accent on voluntaryism has been a strong stress on the will of the believer and the giving of a voice to all believers in the questions of the governance and destiny of the church.

The theological counterpart to the teachings that disengaged radical Protestantism from Catholic continuity or established life was the view that no human authority determined modes of church life. The church is Christ's, and not man's. As such, it seeks to transcend territorial, racial, and ethnic bounds in theory, even if it is rarely consistent or successful in practice. As Christ's church, it is capable of representing him fully in each place, and thus local governance or authority, and even autonomy, was universally stressed.

The radical Reformation almost always restored the sense of an apostolate (missionary outreach); whereas the conservative Reformers had often neglected a sense of witness and missionary activity, and some even ruled it out from the church's present-day mandate. Anabaptists and spiritualists and "free" church (nonstate) advocates tended to be missionary, even if this meant a kind of subversion of established Protestant churches, filled as these were—in radicals' eyes—with unbelievers or inadequate believers.

Concern for the "last times"

Relationships between church and state. Churches as disengaged as these were from established structures were in principle devoted to, and in practice successful in adhering to, ideas that called for sharp distinctions between Christian and non-Christian, sacred and secular, religious and worldly life. This is not to say that radicals took no interest in the civil or social realm; they often did, indeed. But they brought a special viewpoint. They were "eschatological"; that is, they almost always were moved by dramatic views of the future, in which Christ would come again or the Kingdom or Day of the Lord would be announced to change everything. Worldly conditions were temporary and were judged by the saints as ephemeral and corrupting, even if they found it necessary to live with or to employ earthly instruments in the meantime. At the same time, for the sake of the freedom and purity of the believer's church, its members advocated separation from the civil realm, permitting no intrusion by civil authorities in church affairs and seeking no direct involvement in administration of the state by ecclesiastical figures.

Because many of the radicals believed that Christ's new order was imminent, they took a negative view of most human means of facing problems. Many of them advocated a rejection of warfare and saw in the Gospels a support of pacifist positions. The modern "peace church" witness of Mennonites, Brethren, and Quakers was born of this impulse. Paradoxically, other radicals (such as Thomas Müntzer) on occasion saw violence and warfare as legitimate means for them to help hasten Christ's new order.

Church discipline. Separation between the church and the world, and membership based on clear commitment, made it possible for radicals to insist on higher standards of church membership and stricter means of church discipline than could their magisterial counterparts. Social control was more feasible in these smaller and well-defined groups than in the established churches, and "the ban," as a form of excommunication, was the instrument which supported discipline. The use of the ban meant expulsion from the congregation of believers and, with it, social exclusion. The ban was not conceived as merely a punitive measure; brotherly admonition and discipline were to continue, with the hope that the wayward could be rescued.

Rejection of infant Baptism

Believers' Baptism. A special word must be said concerning Baptism since it gave its name Anabaptism to part of the movement and was one of the radicals' most dramatic points differentiating them from the rest of Protestantism. Infant Baptism, from the radicals' viewpoint, cheapened the standard of church membership and was not clearly designated or foreseen in the New Testament documents that chartered the church. Michael Sattler (*c.* 1500–1527), Menno Simons, and Balthasar Hubmaier (1485–1528) led the opposition to infant Baptism. Radicals would follow Jesus, who underwent Baptism as an adult, and they also would be "buried" (in water) with him, as St. Paul said baptized people would be. "New birth" would come from this act, and the reborn believers would restore the church.

Doctrine of the ministry. The concept of ministry was also changed more drastically in radical groups than in the more established Protestant circles. When priests became Lutherans, Calvinists, or Anglicans, there could be a rather subtle transition in their calling. The Anabaptists and spiritual Reformers, however, wanted a clean break with the past. The minister was viewed chiefly as a prophet, not as priest. As an agent of a new order, anticipating Christ's fulfilled Kingdom, he was not to care about earthly prerogatives or routines. Some men, such as Menno Simons, believed that the only way to take on the new ministerial vocation was to repudiate their Roman Catholic ordination. But such conversions from Catholic to radical clergy were rare, and the radical wing of the Reformation more frequently expressed its views on the ministry by simply placing a low valuation on ordination. The classical Reformers wanted university-trained, theologically expert ministers. The radicals, on the other hand, permitted laymen to be ministers: leaders such as Kaspar Schwenckfeld and Konrad Grebel (*c.* 1498–1526) were probably never ordained.

The suffering of persecution. The radical Reformation and the believer's church were made up of people who were prepared to suffer for their faith at the hands of both civil authorities and Catholic and other Protestant ecclesiastical leaders. The story of the rise of Anabaptism is one of persecution, of exiles and fugitives, and of a pilgrim church. The story of the rationalist form of Reformation, as in the case of Michael Servetus (anti-Trinitarian; *c.* 1511–53), often ended in something that can be called a Protestant Inquisition, in which men died for their ideas. Though some erratic personalities may have revealed a desire for martyrdom, more characteristic were those who upheld the idea of patterning one's life after Jesus, the great example. He had not known status or security and was eventually condemned to death; how could his true followers evade a similar path?

Doctrinal variations. Doctrinal varieties among the radicals were many, and it is hardly fair to cluster the various emphases. Certain features stand out, however. First of all, the role of Christ, central to Protestant Christianity, shifted subtly but significantly. The emphasis on Christ's priestly work, in which he brought sacrifice for men before the altar of God, was displaced by a new regard for his prophetic role. He had thundered against the powers of religion and civil society, against established forces, and against the rich; so would his followers. He was seen less as an agent in a divine-human transaction culminating in death on the cross as a sacrifice, and more as the supreme exemplar and leader.

Christ as the prophet and example

The radicals spoke critically of scholastic philosophy and the intellectualized theology built upon it, so they displayed a distaste for the more arcane expressions of classical theology. Faustus Socinus (1539–1604) in Poland and Michael Servetus in Strassburg became shapers of modern Unitarianism. They believed that the doctrine of the Trinity was an unscriptural abstraction and that simple monotheism could best be protected if Christ were not defined as a full expression of the Godhead. Unitarianism remained a distinctly minority emphasis in the radical Reformation. The Bible was usually highly regarded, but whereas the magisterial Reformers tended to see it in the context of tradition, the radicals stressed contemporary personal experience and often allowed for or claimed new special revelation.

PROTESTANTISM'S INFLUENCE IN THE MODERN WORLD

Influence on nationalism. Protestantism eventually became the majority faith throughout northwestern Europe and in England and English-speaking America. From there, in the great 19th-century Protestant missionary movement, it was carried into all parts of the world, joining Roman Catholicism as a minority presence in Asia and Africa and at the same time also establishing beachheads in largely Catholic Latin America. It is impossible to separate Protestantism from the general history of the North Atlantic nations, where it was firmly established for centuries and where its "free" churches or, after "separation of church and state," its voluntary churches, still predominated.

Thus it is possible to speak of Protestantism's contribution to modern nationalism, one of the major historic forces of recent centuries. It shared in shaping this force initially by helping bring to an end the Holy Roman Empire, which was disintegrating already at the time of the Reformation but which finally collapsed in 1806. The old *corpus Christianum* (body of Christ; *i.e.*, Christian society) did not survive; the Christendom initiated by the Emperor Constantine had not envisaged formal Christian division, and the presence of Protestantism spelled the doom of an international, transterritorial, unified Christianity under one head. Protestantism's desire to cultivate literacy and to spread regard for the vernacular served to remove the Latin linguistic bond of older Christendom and to encourage the rise of national boundaries based on languages. All but the radicals tended to make much of loyalty to the existing state, and Protestants often provided an ideological base for each new state as it rose to self-consciousness—as was the case in Prussia or in the United States.

Influence on social ethics. Similarly, many social thinkers, economists, and historians have discerned a much-debated "Protestant ethic," which purportedly had a decisive significance in forming modern attitudes toward raw materials, waste, production, and capital expansion. Max Weber (1864–1920) in *The Protestant Ethic and the Spirit of Capitalism* traced the roots of capitalism to those cities where Calvinism was strong. Today not many historians see such a simple connection: there has been capitalism before and apart from Calvinist influence, and Calvinism has not always produced capitalism. Still, in the Protestant doctrine of the "calling," wherein men at machines or desks could be fulfilling God's purpose as profoundly as did any monk or nun in a convent, men have seen the rise of a new seriousness toward work. In Protestant doctrines of the use of the earth and of stewardship, scholars have noticed a major contribution to the industrial ethos.

Positive concept of work

Influence on the arts. Protestant attitudes toward the arts have been ambivalent and therefore have produced mixed results. For the most part, Reformed and spiritualist Protestants have been uneasy about the arts, fearing lest the symbol be confused with the reality—and lest, therefore, the symbol be idolized and the reality forgotten. Thus Calvin and Zwingli found little place for the visual arts, though Luther showed interest and was a friend of some artists of his time, including Lucas Cranach (1472–1553). Luther also revealed a more affirmative attitude toward music than did the Swiss Reformers, though through the centuries most of Protestantism encouraged the use of music. When Protestant historians want to point to past glories in the aesthetic realm, they cite men like John Milton (1608–74) in literature, Rembrandt (1606–69) in painting, and Johann Sebastian Bach (1685–1750) in music, though such a group has few heirs in more recent centuries.

Ecumenical concerns. While it is clear that Protestantism by nature had to allow for great variety, not all Protestants have rested content with division and separation. They were caught between two biblical mandates. One commanded them to seek the truth and not to express full fellowship with those they considered to be in error. The other stressed the values of Christian unity as a witness in the mission of the church and as a foretaste of the eschatological, or fulfilled, life of Christians when, all agreed, they will all be one. The ferment of the 16th century and the doctrinal formulations of the 17th century led to ever-increasing divisions and hardening of lines or positions. The 18th-century Enlightenment, which in its British and German forms lived off and fought against Protestantism just as the French forms similarly related to Roman Catholicism, tended to breed a spirit of consensus. The Enlightenment placed an exceedingly high value on toleration of differences even as its spokesmen worked for agreement on doctrines based on a search for what they viewed as natural in reason and law. Such a tendency inevitably served to minimize doctrinal differences among Protestants.

Protestant discontent with division and separation

The 20th century, however, has seen more effort toward producing consensus than did the previous three and a half centuries. The modern ecumenical movement, today thoroughly Protestant-Catholic-Orthodox in its outlook, was first born and institutionalized on Protestant soil by men who saw the mission of the church frustrated by competition and division. Beleaguered, huddled together like sheep in a storm, to use a familiar picture, they sought each other's company.

At the same time, modern transportation and communication techniques effectively reduced their world and made uniting symbols accessible. A theological recovery was fused with a new vision of common tasks to produce a Protestantism eager for common statement and often for common action in an ecumenical era. The ecumenical movement has led to denominational mergers and to conciliar organizations, on both confessional and transconfessional lines.

In the meantime, the openness of Roman Catholicism, particularly exemplified in the career of Pope John XXIII (1881–1963), has led to new amity and concord between Protestants and Roman Catholics. In the last third of the 20th century, both of the old warring parties, without formally repudiating their polemical positions of the 16th century, have tended to move beyond its terms and to find new bases for meeting. Modern Catholic biblical commentators speak in what sounds much like Protestant terms of grace and faith. Protestants have new appreciation for a Roman Catholic view of the interconnectedness of the components of the church. More and more, Protestants view the Scriptures as rooted in a tradition and tradition as rooted in the Scriptures. Thus they have a new sympathy for Catholic views of tradition—even as some Catholics criticize unreflective responses to ecclesiastical authority on coercive lines in their own communion. Protestants and Eastern Orthodox Christians, generally quite spatially separated, have begun to understand each other through agencies and organizations such as the World Council of Churches.

CONCLUSION

"The end of the Protestant era"

In the latter half of the 20th century, many heirs of Protestantism, among them the philosophical theologian Paul Tillich (1886–1965), have begun to speak of "the end of the Protestant era," or of the times as being "post-Protestant." This does not mean that they all waver in their faith in Protestantism's general witness. Tillich, for one, argued that "the Protestant principle" of prophetic criticism had to be included in any authentic expression of church life and that it was a genuine value in the secular world. But these thinkers believed that the cultural dominance of Protestantism on its own historic soil was waning.

Emerging at the time of the Renaissance and developing during the Enlightenment, the adherents of Protestantism saw their thought-world repeatedly challenged from without. During the 19th century, with the rise of industrialism and urbanization, a changing world presented new problems to "sociological Protestantism." Meanwhile, ideologues, some of them avowed "god-killers," rose up on Protestantism's territory to challenge its deepest beliefs: the economic theorist Karl Marx (1818–83), the evolutionary theorist Charles Darwin (1809–82), and the philosophical nihilist Friedrich Nietzsche (1844–1900), to take three examples, were thoroughly at home with the Protestant experience and used it as a foil to develop many of their own views.

In the 20th century, Protestantism has become uncertain about its "foreign mission" of expansion in a postcolonialist, anti-imperialist world. The modem appreciation for values in non-Christian religions has led many Protestants to adopt positive attitudes toward these at the expense of the desire to extirpate or displace them with an expanding Protestantism. Totalitarian forces, particularly in Nazi Germany, absorbed some Protestant emphases and changed them beyond recognition, or they persecuted those Protestants who radically opposed suppression.

The attractions of modern life, secularization, and a crisis of faith, all have contributed to a general Protestant decline, beginning with a measurable decrease in church membership, first on the Continent in the 19th century and then in England around the turn of the century. Therefore, while huge majorities of the population (as in Scandinavia and England) are baptized members of established Protestant churches, only a small percentage of these are attendants at worship services or responsive to the disciplines and mandates of the church. Those who use church attendance and support of ecclesiastical appeals as indicators of Protestant fortunes unite with those who see that Protestant dogma no longer defines belief, nor do its divisions any longer excite Western man—and then note the end of the Protestant era.

On the other hand, Protestantism is deeply integrated into so many elements of Western culture that it can be expected to continue to assert subtle influence. It has experienced ebb and flow or revival and decline periodically and now may be going through an extended period of decline. At the same time, it lacks little so far as devoted members in many parts of the world are concerned, and they may rally as they have in the past to spread their version of belief in Christ and their loyalty to a Christian way of life.

BIBLIOGRAPHY. E. MOLLAND, ***Christendom: The Christian Churches, Their Doctrines, Constitutiorzal Forms and Ways of Worship*** (1959); J. DILLENBERGER and C. wELCH, ***Protestant Christianity Interpreted Through Its Development*** (1954); J.S. WHALE, ***The Protestant Tradition*** (1955), a summary of the creedal positions of Protestant bodies; W. PAUCK, ***The Heritage of the Reformation*** (1950), reflective essays on the theological and practical impact of Protestantism; R.M. BROWN, ***The Spirit of Protestantism*** (1961), the main themes of Protestant life summarized by a gifted theologian; J.H. NICHOLS, ***Primer for Protestants*** (1947), a brief survey of Protestant history and theology for the layman; J.B. COBB, ***Varieties of Protestantism*** (1960), a theological analysis of living alternatives in Protestantism, ***Living Options in Protestant Theology: A Survey of Methods*** (1962); R. MEHL, ***The Sociology of Protestantism*** (1970), the best survey of Protestant sociology; G.H. WILLIAMS, ***The Radical Reformation*** (1962), the most comprehensive and authoritative work in English on this subject; F.H. LITTELL, ***The Origins of Sectarian Protestantism: A Study of the Anabaptist View of the Church*** (1964), a historical analysis of the main themes in the radical Reformation; LOUIS BOUYER, ***The Spirit and Forms of Protestantism,*** trans. by A.V. LITTLEDATE (1956); J.L. DUNSTAN (ed.), ***Protestantism*** (1961), a useful book combining sources with narrative and interpretation; PAUL TILLICH, ***The Protestant Era,*** trans. by J.L. ADAMS (1948), a collection of essays, one of which discusses the "end of the Protestant era"; C.W. KEGLEY, ***Protestantism in Transition*** (1965), a theologian's survey of Protestant tendencies after the middle of the 20th century; ERNST TROELTSCH, ***Protestantism and Progress,*** trans. by W. MONTGOMERY (1958), a classic interpretation of Protestant contributions to modernity; MAX WEBER, ***The Protestant Ethic and the Spirit of Capitalism,*** trans. by TALCOTT PARSONS (1930), much debated seminal study of Protestantism and economic life; J.A. HARDON, ***The Protestant Churches of America*** (1956), a Catholic's summary of Protestant doctrinal positions, based on a reading of their official writings; F.E. MAYER, ***The Religious Bodies of America,*** 4th rev. ed. (1961), denomination-by-denomination study of doctrinal positions in American religious groups; W.S. HUDSON, ***American Protestantism*** (1961), an excellent brief survey of Protestant history in America; J.C. BRAUER, ***Protestantism in America: A Narrative History*** (1953), a simple statement of the main themes of American Protestant history; A.L. DRUMMOND, ***Story of American Protestantism*** (1949), a British view of Protestant history in the United States; M. MARTY, ***Righteous Empire: The Protestant Experience in America*** (1970), written for the national bicentennial.

(M.E.M.)

Protestantism, History of

Protestantism, beginning in northern Europe in the early 16th century in reaction to medieval Roman Catholic doctrines and practices, became, along with Roman Catholicism and Eastern Orthodoxy, one of three major forces in Christianity. After a series of European religious wars, and especially in the 19th century, it spread rapidly in various forms throughout the world. Wherever Protestantism gained a foothold, it influenced, to a greater or lesser extent, the social, economic, political, and cultural life of the area.

This article is divided into the following sections:

I. "Protestantism"

MEANING OF THE TERM

Origin of the term Protestant

Protestantism, one of the three major branches of Christianity, was given its name at the Diet of Speyer in 1529. At that imperial assembly the Roman Catholic princes of Germany, along with the Holy Roman emperor Charles V, rescinded most of what toleration had been granted to the followers of Martin Luther three years earlier. On April 19, 1529, a protest was read against this decision, on behalf of 14 free cities of Germany and six Lutheran princes, who declared that the decision did not bind them because they were not a party to it, and that if forced to choose between obedience to God and obedience to Caesar they must choose obedience to God. They appealed from the diet to a general council of all Christendom or to a congress of the whole German nation. Those who made this protest became known as Protestants. The name was adopted not by the protesters but by their opponents, and gradually it was applied as a general description to those who adhered to the tenets of the Reformation, especially of those outside of Germany. In Germany the adherents of the Reformation preferred the name evangelicals and in France Huguenots.

The name Protestant was also attached not only to the disciples of Luther (*c.* 1483–1546) but to the Swiss disciples of Huldrych Zwingli (1484–1531) and later of John Calvin (1509–64). The Swiss Reformers and their followers in Holland, England, and Scotland, especially after the 17th century, preferred the name Reformed.

VARIOUS USES OF THE TERM

In the 16th century the name Protestant was used primarily in connection with the two great schools of thought that arose in the Reformation, the Lutheran and the Reformed. In England in the early 17th century the word Protestant was used in the sense of "orthodox Protestant," as opposed to those who were regarded by Anglicans as unorthodox, such as the Baptists or the Quakers. Roman Catholics, however, used it for all who claimed to be Christian but opposed Catholicism (except the Eastern churches). They therefore included under the term Baptists, Quakers, and Catholic-minded Anglicans. Before the year 1700 this broad usage was accepted, though' the word was not yet applied to Unitarians. The English Toleration Act of 1689 was entitled "an Act for exempting their Majesties' Protestant subjects dissenting from the Church of England." But the act provided only for the toleration of the opinions known in England as "orthodox dissent" and conceded nothing to Unitarians. Throughout the 18th century the word Protestant was still defined in relation to the historical reference of the 16th century Reformation. Samuel Johnson's dictionary (1755), which is characteristic of other dictionaries in that age, defines the word thus: "one of those who adhere to them, who, at the beginning of the reformation, protested against the errours of the church of Rome."

II. The Reformation period

THE IMPULSE OF THE REFORMATION

The role of Luther. In 1517 Martin Luther published at Wittenberg his Ninety-five Theses on the subject of indulgences (the remission of temporal penalties for sins) and other matters, arguing that an external act, such as buying an indulgence, could never be a substitute for inward penitence. Since the theory of indulgences rested upon the authority of the pope, an attack upon them soon broadened into an attack upon the pope's particular claims to authority. Germany was full of discontent against papal interference, and the German people rose in support of Luther.

The person of Luther was a key to understanding the way in which Protestantism developed, for his experience conditioned Protestant theology. Before 1517 he had attempted faithfully to practice the life of a monk but had found that his efforts brought him no peace of soul. By the study of St. Paul and St. Augustine he was led to an ever-growing trust in God's help and an ever-growing distrust of his own efforts. Between 1512 and 1518 (probably in 1512) he had an experience in the tower of the Austin Friars at Wittenberg in which he apprehended the full force of St. Paul's text (Rom. 1:17): "The just shall live by faith." Henceforth justification by grace through faith was to be the doctrine central to Luther and thereafter to the Protestant churches. Almost all forms of Protestantism have repudiated doctrines such as merit or works of supererogation, which implied that a man could by good conduct set up claims upon God or win his own way toward heaven.

Importance of doctrines of justification and the authority of the Bible

The role of the Bible. The Protestant churches appealed to the Bible and the primitive church and therefore aimed in a measure to imitate, whether in ministry, or in organization, or in simplicity of worship (Lutherans and Anglicans, however, made little effort to imitate the primitive church in worship), what they took to be the conditions of primitive Christendom. They aimed to purify and simplify the church; but for all their use of the word reformation, which suggested a looking back, they looked forward toward a time of what they believed would be purer sacraments, devout preaching, godly ministry, and moral conscience among Christian people. Though Luther and Calvin aimed primarily at a doctrinal reformation, a drive toward what they thought to be a better morality, both private and public, was the most marked feature of other leaders of the Protestant movement of the 16th and early 17th centuries. Protestant emphasis was on the reconciliation of men not by their own endeavours but by God's help, which was granted to them through what Christ had done and suffered and received through the willing assent of the heart. Protestants had a deep sense of the sinfulness of the human race, of its fearful predicament before the judgment of God, and of its powerlessness to save itself.

Protestantism made its basic appeal to the Bible as the ground of its belief. In the early 17th century the English theologian William Chillingworth defined it thus: "By the religion of Protestants I do not understand the doctrine of Luther, or Calvin, or Melanchthon . . . nor the articles of the Church of England; no, nor the harmony of Protestant Confessions; but that wherein they all agree . . . that is, the Bible." It is true that the varieties of Protestants have included thinkers who would not quite accept Chillingworth's definition. But the common factor in Protestantism has been the acceptance of the supremacy of the Bible over the churches; the belief that ecclesiastical ministries or hierarchies were to be tested against the Bible as the word of God; the doctrine that all things necessary to salvation were to be found in Holy Scripture; the belief, summarized in the phrase "priesthood of all believers," that truth may not be with constituted authority but with the ordinary soul quietly following the leading of God.

This led to a flowering of biblical commentaries and exegetical, or interpretive, works. Luther and Calvin were famous for their commentaries upon books of the Bible. Luther finished a translation of the Bible into German in 1534, and that version was important in the further development of the German Reformation and in influencing Protestant translations into other languages (Danish, Swedish, Dutch, Lithuanian, and English).

Emphasis on the sermon and education

Protestants laid much emphasis on the delivery of the Word of God to the people, and the sermon became a most important part, sometimes the most important part, of every Protestant act of public worship. The major Reformers placed much stress upon education and from the time of Luther onward emphasized catechisms—*i.e.*, classes before confirmation—and the common devotional study of the Bible. Martin Bucer (1491–1551), the Reformer of Strassburg, introduced confirmation as a formal act, after a period of preparation, in which the vows taken on behalf of an infant at Baptism were renewed. The Reformers also sought to take the Bible into the household and, with their belief in the priesthood of the laity, created family prayers in middle-class homes.

THE ESTABLISHMENT OF THE REFORMATION CHURCHES, 1521–80

The development of territorial churches. In the early days of the Reformation period, the Protestant states were saved from suppression because they could have been suppressed only if the Holy Roman emperor in Germany had been given power to suppress them, and many rulers who were not Protestants---other German princes, the king of France above all, and even at times the pope himself—were afraid to give any German ruler such vast powers.

Between 1521 and 1580 most of northern Europe became Protestant. The strength of the movement lay in the educated middle classes and the tradesmen of the towns, for many inhabitants of which the Renaissance and northern humanistic studies had made some of the traditional practices of the medieval church appear to be superstition or magic. The cities of Germany and Switzerland—Nürnberg, Bern, Geneva, Zurich, Strassburg—were free cities and among the earliest to go Protestant. By 1539 a league of Protestant States included Saxony, Brandenburg, Prussia, Württemberg, Hesse, Brunswick, Anhalt, and half of Switzerland. By 1560 Protestantism carried England and Scotland, Sweden, Norway, and Denmark and looked as though it might soon include France. From 1565 on there was a possibility of a Protestant Netherlands, and Protestantism was spreading rapidly in Austria, Bohemia, Poland, and Transylvania. Between 1521 and 1580 the Protestant map of Europe was drawn, almost as it has re-

Advances and retreats of Protestantism

mained. It suffered its severest setbacks in France (where long civil wars prevented the Huguenots from becoming more than a minority but where they were able to achieve toleration by the Edict of Nantes (1598), in the southern Nethlerlands (Belgium) and the nearby prince-bishoprics of the Rhineland (which Spanish troops helped to preserve for Catholicism), in Austria and Bohemia (where the Thirty Years' War—1618–48—forcibly re-Catholicized those regions), and in Ireland (where the Irish population identified resistance to Protestantism with resistance to the English). In Spain and Italy, though some distinguished converts were made, Protestants never constituted more than a few small groups. In the Eastern churches Protestantism affected only a few theologians who had been educated in the West, especially Cyril Lucaris, the patriarch of Constantinople who was murdered in 1638.

From the earliest years of the Reformation, Protestants were divided primarily into Lutheran and Reformed groups. Lutheran states included Sca'ndinavia, north Germany, and Wiirttemberg. Reformed states included Holland, Scotland, and French cities of the Huguenots. England was Reformed for the most part but suspected by other Reformed churches as leaning too much towards the Lutherans.

Ecclesiastical reforms of the Reformation churches (16th to the early 17th century). ***Worship.*** Luther and his colleague Philipp Melanchthon (died 1560) were conservative in their ideas of reform. They had little sense of a breach with the church of the past and thought of themselves as purifying it from abuses. The Lutheran liturgies were merely simpler forms of the Latin mass, and for a time some of them retained the Latin language in parts of the service. In their worship, Lutherans emphasized the doctrine of the Real Presence (of Christ) in the sacrament of the Lord's Supper.

Zwingli, the Reformer of Ziirich, however, held that what was not in the Scripture was not permitted. The forms of prayer in Swiss Reformed churches were simple—readings, prayers, sermons, psalms. The psalms were at first said, not sung, but metrical versions of the psalms with tunes were soon introduced. A French version of the psalms by Clément Marot (c. 1497–1544) and the English version by Thomas Sternhold (died 1549) and John Hopkins (died 1570) became the hymnbooks of the Reformed congregations. A simple and austere congregational worship, suspicious of ritual as a way toward formalism, became a major characteristic of Reformed Protestant worship. Lutherans, however, had a hymnody that did not confine itself to the psalms. Luther wrote 37 hymns and hymn singing became a moving feature of worship in Lutheran churches. The congregational hymn or psalm was a distinctive innovation made by Protestants in Christian worship. More than any other development, it transformed the character of lay worship.

Organization. Besides their forms of worship, another difference between Lutheran and Reformed churches lay in their views on church organization. Churches were reformed by governments (princes or city councils), though they often acted in response to a popular movement. The necessary legal changes required that governments intervene. All Protestant states dissolved monasteries and nunneries and discouraged or suppressed the monastic life; and since the monasteries possessed land and endowments, government had to deal with the legal and personal consequences.

One of the fundamental consequences of the Reformation was the diminution of the legal force of Canon Law when Protestant states declared that no outside authority (church, bishop, or pope) had the power to make laws binding upon some of its subjects without the consent of the government of that state. Thus the Protestant movement not only depended for its success upon the development of the modern sovereign state but by its success assisted that development.

The development of the consistory and presbytery

Where the prince, as in Luther's Saxony, actually took charge of the Reformation, the device for reorganizing the church was called the consistory (an ecclesiastical court of law), whose members were appointed by the prince, sometimes with the prince as its chairman, and which usually consisted of lawyers and clergy in equal numbers. In the German states the consistory became the supreme court, appointing clergy, visiting parishes, and reorganizing church life. In a free city the city council, as in Zwingli's Ziirich, took charge of the Reformation. The city council itself acted as the government of the church as well as of the state.

In Geneva, however, Calvin argued that this mode of church government was unscriptural, that although the government of the state was indeed an instrument of God, the government of the church was divinely ordered in the Scriptures and was not the same thing. God, he asserted, had established the offices of pastors and elders as authorities in the congregation; and these pastors and elders had rights in the government of the congregation that no civil authority could overrule. Thus the Calvinist consistory, though called by the same name as the Lutheran, was an entirely different body; it was created by the pastors and the congregation as a Christian and spiritual, but not as a civil, body, with supreme authority in all spiritual matters--especially moral rebuke and excommunication. This was the basis of the Presbyterian system of church government as it came to exist among all the disciples of Calvin—in Scotland, the Netherlands, some of the German cities in the northwest, the Palatinate after 1570, and the French Huguenots. It became the characteristic organization of Reformed Protestantism, based upon a scriptural authority, strong in its own authority, and yet relying upon an element of Christian democracy.

In Sweden, as in England, the monarchy retained the supreme power in the church and it also retained the bishops. There was, therefore, no normal consistory in Lutheran Sweden. The same, or similar, condition existed in Denmark and Norway.

The development of English church government

The English Reformation was different. There, King Henry VIII, between 1533 and 1547, tried to maintain Catholicism without the pope, dissolved the monasteries or forced them to dissolve themselves, introduced Bibles into the churches, and elevated a scholarly and hesitant Protestant, Thomas Cranmer (1489–1556), as archbishop of Canterbury. Under Henry's successor, Edward VI (1547–53), the Council reformed the church, inclining more toward the Swiss pattern of reformation than to the Lutheran, with a simplified liturgy, a transformed appearance in churches, and a belief in the symbolic presence of Christ in the bread and wine in the Lord's Supper. Protestantism, however, was not yet a popular movement in England and went little deeper than the city of London, the universities, and the merchant class. Under Edward's successor, Mary (1553–58), England submitted without enthusiasm to the Pope, and Cranmer, Hugh Latimer (*c.* 1485–1555), Nicholas Ridley (c. 1503–55), and several other leading Protestants were burnt at the stake. When Elizabeth (1558–1603) ascended the throne, the country was finally Protestant. But the previous swings in public opinion, the Queen's personal position, and the desire not to offend Spain and the more conservative reformers in Germany meant that the Reformation under Elizabeth allowed room for both Catholic and reformed elements. A growing body of English Calvinists, drawing strength from communications with Geneva, the Palatinate, Holland, and Scotland, became discontented with the Catholic inheritances preserved in the worship services and the office of bishops and led directly to the Puritan and separatist movement in England.

The development of the congregation

Though not recognized by the established Protestant bodies, another organizational form appeared—the congregation. In the congregation a group of men and women came together, elected their own pastor or dismissed him, organized their own worship, studied the Scriptures themselves, and remained sovereign, as a democratic group, over their own church life.

In Germany and Switzerland in the 16th century these congregational groups were loosely called Anabaptists (rebaptizers). Organized in small groups, they were usually agreed upon denying the validity of infant Baptism, upholding the sovereignty of the congregation, and declaring that the congregation must have nothing to do

"The Massacre of St. Bartholomew's Day," oil painting by François Dubois (1529–84). The Huguenot leader, Gaspard de Coligny, is shown twice, hanging from a window of his house and lying beheaded in front of the house with the Duc de Guise standing over him. In the Musée Cantonal des Beaux-Arts, Lausanne, Switzerland.

BY courtesy of the Musee Cantonal des Beaux-Arts, Lausanne: photograph. Andre Held

with the civil magistrate. They were world denying, and usually pacifist, in their way of life. In the 16th century, however, many groups of dissenters were called Anabaptists, whatever their doctrines or practices might be, and under this name were heaped together persons holding mystical doctrines, such as Kaspar Schwenckfeld (1489–1561) and Hans Denck (1495–1527); those who would introduce the Kingdom of God by force, such as at Miinster in 1533–34; and those groups advocating a form of religious communism, such as the Hutterites of Bohemia. They suffered severe persecution from Protestants and Catholics alike. For a time they found refuges in Bohemia, Poland, and Hungary, until they were suppressed by force in the 17th century. But their concept and form of church organization later became a prevailing pattern in much of Protestantism. An "Anabaptist" leader in the northern Rhineland, Menno Simons (1496–1561), gave his name to a moderate group of such congregations, the Mennonites, and Holland reluctantly granted asylum to them.

In Elizabethan England many Puritans believed that the state church had not been fully reformed and pressed for further reformation—in discipline, polity, and simplicity of ritual. Since further reformation was not conceded, some of them began to argue for separate congregations. In 1582 Robert Browne (c. 1550-1633) published at Middelburg ***A Treatise of Reformation** without **Tarrying for Any*** and argued for independent congregations that would bind themselves by a covenant, would not attempt to embrace all the world but should gather men and women out of the world, and would have no relations with the state. These features of early Separatist or Independent church views resembled those of some of the groups descended from the Anabaptists and of the Mennonites of Holland, who had an influence on some of these early anti-establishment English groups. The Pilgrims who sailed on the "Mayflower" transplanted these ideas to New England in 1620, and after John Winthrop and his company had settled Massachusetts, the colony worked, in effect, under a state policy of Congregationalism. The experience of the early American colonies was influential in the development of English Congregationalism. During the English Civil War (1642–51) the Parliament became identified with Presbyterianism, and Oliver Cromwell (1599–1658) and his officers with Separatism. English Congregationalists have often looked back with respect toward Cromwell and his regime of comparative toleration of various forms of religious organizations. In 1658 deputies from 120 congregations met at the Savoy Chapel in London and issued the Savoy Declaration, containing a confession of faith and a platform of discipline.

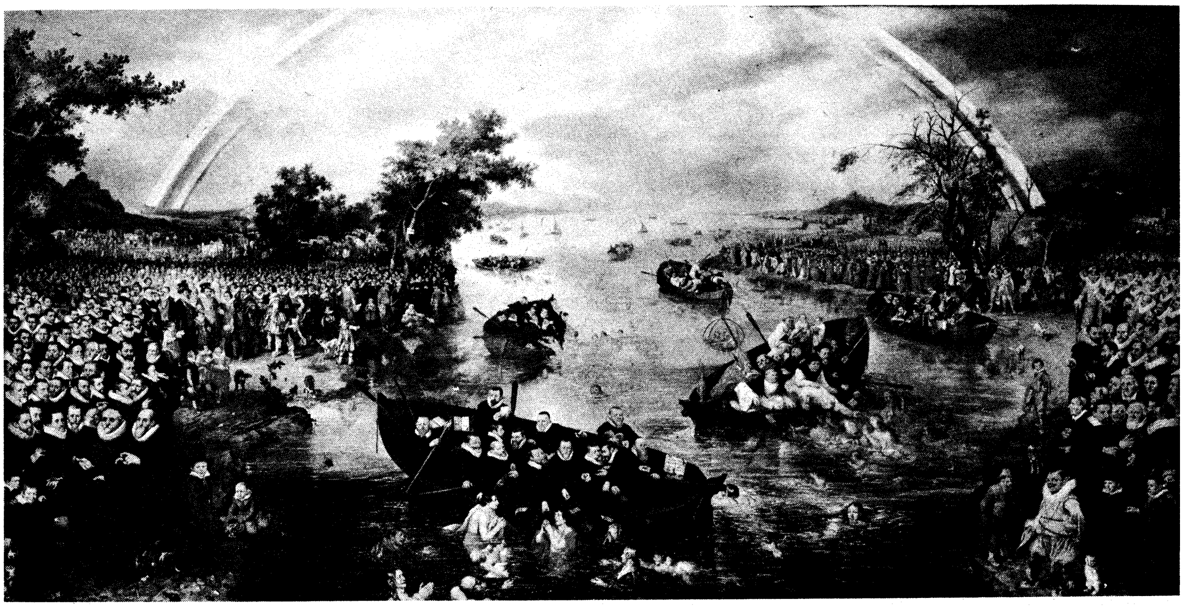
Church organizations and doctrines of Baptists and Friends

Closely associated with the Congregationalists were Baptists, who in origin differed from Separatist congregations only in believing infant Baptism to be wrong. Through the Mennonites of Holland, they were descended from the Anabaptists of the Reformation, but most of them, like the Congregationalists, were Calvinist in their thought and were thus known as Particular Baptists, to distinguish them from a smaller group of General Baptists who were not Calvinist. The English Particular Baptists dated their origins from 1633, and the General Baptists from the forming of a church by Roger Williams in 1639 at Providence in Rhode Island. They both suffered from the English authorities up to 1640 but profited from the toleration practiced during Cromwell's regime.

Of the many little groups that flowered under Cromwell's rule, one became permanent and influential—the Friends, or Quakers, who were organized by George Fox (1624–91) in 1668, when he issued a "Rule for the Management of Meetings." The Quakers became especially known for their emphasis on pacifism and their doctrine of the authority of the "inner light," which was a kind of mystical revelation to the awaiting Friend.

Religious liberty and moral reform. In several Protestant states the Reformation became a symbol of national freedom and independence as well as of ecclesiastical and moral reform, which helps to explain the great influence of Protestant ideas in the later history of Europe and America.

Holland secured its independence from Spain by the wars of 1568–1648. In Scotland the Reformation, under the guidance of John Knox (c. 1514–72) between 1559 and 1564, was successful against the will of two successive Catholic sovereigns, Mary of Guise (1516–58) and Mary, Queen of Scots (1542–87). Under the Stuart monarchs of the 17th century the Presbyterian policy established by Knox became the centre of Scottish resistance to English and royal domination. Despite the existence of a Roman Catholic minority in the Highlands, Protestantism became identified with Scottish nationalism. Protestantism also became identified with an English nationalism, when, after the reign of the Catholic Queen Mary, the English people under Queen Elizabeth I fought against the threat of Catholic conquest by King Philip II of Spain. Though the papacy had long since ceased to be a political force that could dominate all of Europe, the English regarded the pope as a potential threat to their



"Fishing for Souls," by Adriaen Pieterszoon van de Venne, 1814, depicting Catholics and Protestants competing for converts in The Netherlands. Protestant James I of England and Prince Maurice of Nassau are shown standing on the left bank and the Catholic Archduke Albert of Austria and his wife, Isabella, on the right bank. In the Rijksmuseum, Amsterdam.

By courtesy of the Rijksmuseum, Amsterdam

liberties. In France, the Huguenots achieved the right to exist as a result of the most terrible wars of the 16th century, including the infamous Massacre of St. Bartholomew's Day, August 24, 1572. When the edict of toleration, the Edict of Nantes (1598), was revoked by King Louis XIV in 1685, they fled from France. The French Huguenot exiles in England, Germany, and America carried with them the conviction that Catholicism and freedom were incompatible.

Calvinist efforts to reform church and society

The Protestants set out to reform the church, involving not only its doctrine but, as they saw it, the morals of Europe as well. They promoted education, simplicity, and hard work and discouraged luxury, fornication, extortion, and drunkenness. Within the Calvinist churches especially, the drive to reform was both effective and controversial because of the nature of the Calvinist consistory, with its independent government and its special task of supervising public and private morality. The tradition of the Calvinist churches, whether in Switzerland, Holland, Scotland, or New England, was distinguished by austere simplicity, an immediate sense of vocation (of an individual life calling), a deep attention to biblical texts, and a refusal to compromise. The Calvinists' belief that they were of the "elect" gave them an assurance of salvation and a sense of total dependence upon God.

The name Puritan was first applied to these Calvinists as a term of abuse in England, and there the movement became controversial because it was caught up in the constitutional conflict between crown and Parliament. The alliance between church and king tended to create an alliance between Puritans and the parliamentarians who were against the king, though some parliamentarians were far from Puritan and some Puritans were royalists. Under the rule of Parliament during the Civil War of 1642–51, the Puritan Westminster Assembly attempted to reform the Church of England, by making it a Presbyterian Church in association with the Church of Scotland. The fall of Parliament before Cromwell's army in 1647–48 meant that the Presbyterian system was never fully established in England; but the tone of Cromwell's England was, in general, Puritan. With the restoration of King Charles II (1660) the royalists reacted against everything Puritan, and the movement was driven underground, to be revived later among the Methodists.

III. The late 17th through the 19th century

THE LATE 17TH AND 18TH CENTURIES

Catholic recovery of Protestant territories. After the Peace of Westphalia in 1648 that ended the Thirty Years' War, Catholicism regained some territories from Lutheran Protestantism: first, because the rise of toleration was somewhat more rapid in Protestant countries than in Catholic lands and, secondly, because Louis XIV identified French power with universal French acceptance of the Roman Catholic faith and in 1685 revoked the Edict of Nantes and expelled thousands of Huguenots, who thus fled to England, Holland, or Germany, much to the advantage of those countries. Several of the French refugees became prominent in English religious life, and in Prussia groups of them founded flourishing congregations known as the French Reformed. In 1702 a determined group of Huguenots in the mountains of the Cévennes in France, known as the Camisards, rose in rebellion but were suppressed by military power two years later. There was a further small outbreak of war in 1709. For a time the few surviving Huguenot congregations met only in secret. They were led by Antoine Court (1695–1760), who secured ordination from Ziirich and founded (1730) a college at Lausanne to train pastors. French Protestants barely held out until the French Revolution, after which they had a revival.

Survival of French Protestantism

The gaining of Alsace in 1648 for France had enabled Catholics to increase rapidly, and Protestants decreased in strength. Strassburg, once one of the leading cities of the Protestant Reformation, returned its cathedral to the Catholics (1681) and became a town with a large Catholic population. Louis XIV ruled the Palatinate for nine years and allowed the French Catholics to share the churches with the Protestants; and though he was compelled to surrender the country at the Treaty of Rijswijk (1697) to the Holy Roman Empire, a clause (the *Simultaneum*) of the treaty (added at the last moment and not recognized by the Protestants) preserved certain legal rights and endowments of Catholics in Protestant churches. The consequence of the increased power of France was to diminish Protestant authority in the Rhineland between Switzerland and the Netherlands.

Another shock to Protestantism was the conversion of Augustus II, Elector of Saxony, to Roman Catholicism in 1697. It appeared as though Protestantism was not even safe in its original home. The conversion involved political motives; Augustus was a candidate for the throne of Poland and was loyal to his new allegiance, assisting the Catholic Church in Poland and also, somewhat, in Saxony; but such assistance had no effect on the Lutheranism of Saxony.

Protestant scholasticism. The second half of the 17th century was at once the high age of Protestant systematic orthodoxy and the age when the first signs of its dissolution appeared. The axioms of the Reformation were worked out in a great and systematic body of doctrine.

The theologians defended and the pastors taught Luther's or Calvin's dogmatic systems—relying also upon authoritative sources such as the Formula of Concord (1577) in Lutheranism or the conclusions of the Synod of Dort (1618) in Calvinism—which were extended and made into a tradition. Even when the system was not of the ordinary Protestant tradition, it was generally worked out in many volumes, based upon coherent axioms, defended against all assailants, appealing always to reason and to biblical authority and seldom to feeling or conscience. This age has sometimes been known as the age of Protestant scholasticism. But that pejorative term came from a posterity that would no longer accept the axioms on which the systems were founded. These were the last scriptural theologians before the period of the Enlightenment, when the understanding of Scripture was altered. The old axioms were changed by Pietism, science, and philosophy.

The influence of P.J. Spener and his *Pia Desideria*

Pietism. The Pietists, groups of earnest students of the Bible and seekers after faith and moral experience, were founded by P.J. Spener (1635–1705), an Alsatian Lutheran of Strassburg. At Frankfurt am Main in 1667 he initiated Sunday devotional meetings for laymen, and in 1670 the *collegia pietatis* ("assemblies of piety"), which were held twice a week in his house for the discussion of the previous Sunday's sermon or of some religious book. In 1675 he published *Pia Desideria* ("Pious Desires"), a plea for a deepening of church devotion, charging the clergy with lacking self-denial, theology with being little but controversial, and the universities with teaching knowledge useless for pastoral concerns. The *Pia Desideria* exercised a great influence throughout Germany and gave birth to many small groups of pious laymen meeting to study the Bible.

In 1694 the Elector of Brandenburg (Frederick 111), under Spener's influence, founded the University of Halle, brought A.H. Francke (1663–1727) there, and filled other chairs with men of his views. Halle became the centre of German Pietism with rapidly rising numbers of students. Francke administered an orphanage with printing house and schools. The Halle faculty, which ceased to be Pietist before 1750, showed the strength and weakness of the movement. It was deeply concerned with saving souls and with the practical and pastoral work of the church; it laid much stress upon the need for conversion and held up before young men high standards of moral responsibility. Pietism attempted to show men that religion was true only if it was personal and individual. It was inclined to be censorious of weaker folk. Also, it was little concerned about the formal accuracy of doctrine and thus was able to cross the barriers that divided denominations. The Pietists raised the general standards of church life. Yet because they shattered the conventional and formal Protestant theological systems, the Pietists opened the way toward more questioning attitudes. They demolished old theological structures and served as a bridge between orthodox and 19th-century theological thought. From Spener onward the movement cared for education and gave impetus to the creation of schools for the poor and also for the rich. Under Pietistic influence confirmation became almost universal among the German churches. The Pietists also first directed the German churches toward foreign missions. One of the Pietistic groups, the Moravians, led by Graf von Zinzendorf (1700–60), established missions in the West Indies (1732), Greenland (1773), and Georgia (1735). The Pietistic movement reached Sweden (where it met much opposition), Denmark, and Norway by the beginning of the 18th century. Under Danish sponsorship, two Pietist missionaries (B. Ziegenbalg and H. Plütschau) established the first Protestant mission in India.

Rationalism. The first signs of a Rationalist movement, which was to have as powerful an influence on Protestantism as the Pietists had had, may be traced back to those few who at the end of the 16th century attacked Calvinism on grounds of reason. In Leyden, the Netherlands, Jacob Arminius (1560–1609) reacted against Calvinist doctrines of predestination (God's foreordaining men to heaven or hell). Though anyone not a Calvinist after a time came to be called Arminian, there were groups so designated in Holland and England that contained men who were more marked by their use of reason in theology than by their opposition to Calvin. In England, the enemies of such liberal theologians gave them the name Latitudinarians. The so-called Latitudinarians sought to maintain church unity based upon a few fundamental articles of faith and otherwise to allow for a wide diversity of doctrine, polity, and ways of worship. Their best representatives were the Cambridge Platonists—philosophical theologians at Cambridge (*c.* 1640–80)—who claimed that reason is the reflection of the divine mind in the soul.

During the 17th century, philosophy, hitherto considered a handmaid to theology, was expanded beyond the limits of Aristotle and the Bible and—partly due to natural science and partly due to the reflections of thinkers from Francis Bacon (1561–1626) and René Descartes (1596–1650) onwards—developed its independence. The successes of science, especially to be noted in the work of Sir Isaac Newton (1642–1727), persuaded many men of the power of reason and, by 1680, of the necessity that all things be tested by reason, including even those realms of the conscience or spirit that hitherto had been thought to be above reason or inaccessible to reason. The symbols of the age of Rationalism were: the rapid decline of belief in witchcraft; the slow and painful rise of a belief in toleration; a more widespread spiritualizing of conceptions like heaven and hell; and the recognition of the small size of the planet Earth within the universe. On the Continent Spinoza (1632–77) and G.W. Leibniz (1646–1716), in England John Locke (1632–1704), were regarded as the philosophers of the age. Among the German theologians Christian Wolff (1679–1754) of Halle approached theology almost as if it were a form of mathematics, seeking for a truth that would be incontrovertible among all reasonable men. He believed that he could demonstrate the truths of God and immortality and allowed the possibility of other languages of revelation as probable in the light of those proofs. Under prompting from Pietists of Halle, he was expelled from Prussia in 1723. But before Wolff's death, Rationalist theologians had displaced the Pietists in control of Halle University and had made it the centre of Rationalist theology among Protestants.

Effects of Rationalism

In England the same trend among the disciples of John Locke issued in the Deists (especially John Toland, 1670–1722) for whom Christianity was never mysterious and was understood only as a republication of the natural religion of the human race. Like Wolff and his disciples, the English Deists had no permanent influence on the history of Protestantism, except by forcing the theologians to answer them and thereby to treat the philosophy of religion with seriousness. The most important of all the answers to the Deists lay in the work of Bishop Joseph Butler (1692–1752), whose sermons and *Analogy of Religion* formed the most cogent defense of the basis of Christian philosophy known in that age.

Rationalist theology, working at the same time as, though certainly not in harmony with, Pietism and evangelicalism, began to modify or even destroy the traditional orthodoxies—*i.e.*, Lutheran or Calvinist—of the later Reformation. The Rationalist theologians insisted that goodness in God could not be different in kind from goodness in men, and therefore that God cannot do what in a man would be immoral. Though for the most part they accepted the miracles of the New Testament—until toward the end of the 18th century—the Rationalists were critical of miracles outside the New Testament, since they suspected everything that did not fit their mechanistic view of the universe.

The role of John Wesley

Methodism. Closely parallel to the Pietists in Germany was the evangelical, or Methodist (named from the use of methodical study and devotion), movement in England led by John Wesley (1703–91). While a fellow of Lincoln College, Oxford, Wesley gathered a group of earnest students of the Bible about him, made a missionary expedition to Georgia, and became a friend of the Moravians. Henceforth, like the Pietists, he laid much emphasis upon the necessity of conversion and devoted the remainder of his life to evangelistic preaching in England. He did not intend any separation, but the parish system of the Church of England as then organized was incapable of adjustment to his plan of free evangelism and lay preachers. In 1744 Wesley held the first conference of his preachers; soon this became an annual conference, the governing body of the Methodist societies, and was given a legal constitution in 1784. The Methodist movement had remarkable success, especially where the Church of England was failing—in the industrial parishes, in the deep countryside, in little hamlets, and in hilly country, such as Wales, Cumberland, Yorkshire, and Cornwall. In 1768 Methodist emigrants in the American colonies opened a chapel in New York, and thereafter the movement spread rapidly in the United States. It also succeeded in French-speaking cantons of Switzerland.

The evangelical, or Methodist, movement seized upon the elements of feeling and conscience that Protestant orthodoxy had tended to neglect. It gave a renewed and devotional impetus to the doctrines of grace and justification and to the tradition of moral earnestness, which had once appeared in Puritanism but which had temporarily faded during the reaction against Puritanism in the middle and late 17th century. In England, it slowly began to strengthen the tradition of free churchmanship over against the tradition of the established church, though for a century or more many English Methodists believed themselves to be much nearer the Anglican Church from which they had issued than any other body of English Protestants. It enabled hymns—hitherto confined (except for metrical psalms) to the Lutheran churches—slowly to be accepted in other Protestants bodies, such as the Church of England, the Congregationalists, and the Baptists. The evangelical movement of the 18th century produced several of the most eminent of Christian hymn writers, especially Philip Doddridge (1702–51) and Charles Wesley (1707–88).

Though John Wesley himself had not been Calvinist, in Wales the Methodists retained both the name and the theology of Calvinistic Methodists. In the United States Methodism made even more rapid progress.

The development of Protestantism in the English colonies. Churches in the 13 colonies of the American states practiced the Congregational or Baptist church polity on a scale not known in Europe. The small Anabaptist groups had required evidence of faith, and this sometimes meant public testimony to the experience of conversion. In the larger congregations of America a similar testimony—because it was given to a wider circle—became more evident, more solemn, and at times more emotional. The pastors of the Calvinistic tradition of New England, trying to escape from the religion of forms and to seek the religion of the heart, gave unusual stress to the necessity for an immediate experience of salvation. Pastors found that under certain conditions a wave of emotion could sweep through an entire congregation and believed that they could here observe conversion and its subsequent issue in a better life. The movement owed something to the German Pietist T.J. Frelinghuysen (1691–c. 1748) and something to John Wesley's colleague George Whitefield (1714–70). The chief mind at the beginning of the Great Awakening, however, was that of an intellectual mystic rather than of a conventional Calvinist preacher. Jonathan Edwards (1707–58) was the Congregational pastor at Northampton in Massachusetts, where the conversions began in 1734–35. In the middle years of the 18th century waves of revivals and conversions spread through the colonies. Though the revivals were led by Congregationalists and Presbyterians, many small, independent Bible-centred groups, which often professed allegiance to Baptist teaching, came into being because of the revivals. As Wesley in England and Zinzendorf in Germany had been forced to carry their new methods outside the established churches of their lands, so also was the experience of American revivalistic leaders. Emotional phenomena and disorders were not welcomed by conservative churchmen who cared for decorous reverence in church.

The Great Awakening

The movement was not native to America. But the conditions of the American frontier gave this kind of evangelicalism a new vigour, and from America it permanently influenced the future development of Protestantism. In the towns and new cities with moving populations, Protestantism found methods that became a feature of evangelical endeavours to reach the unregenerate or the unchurched crowds of the coming industrial cities.

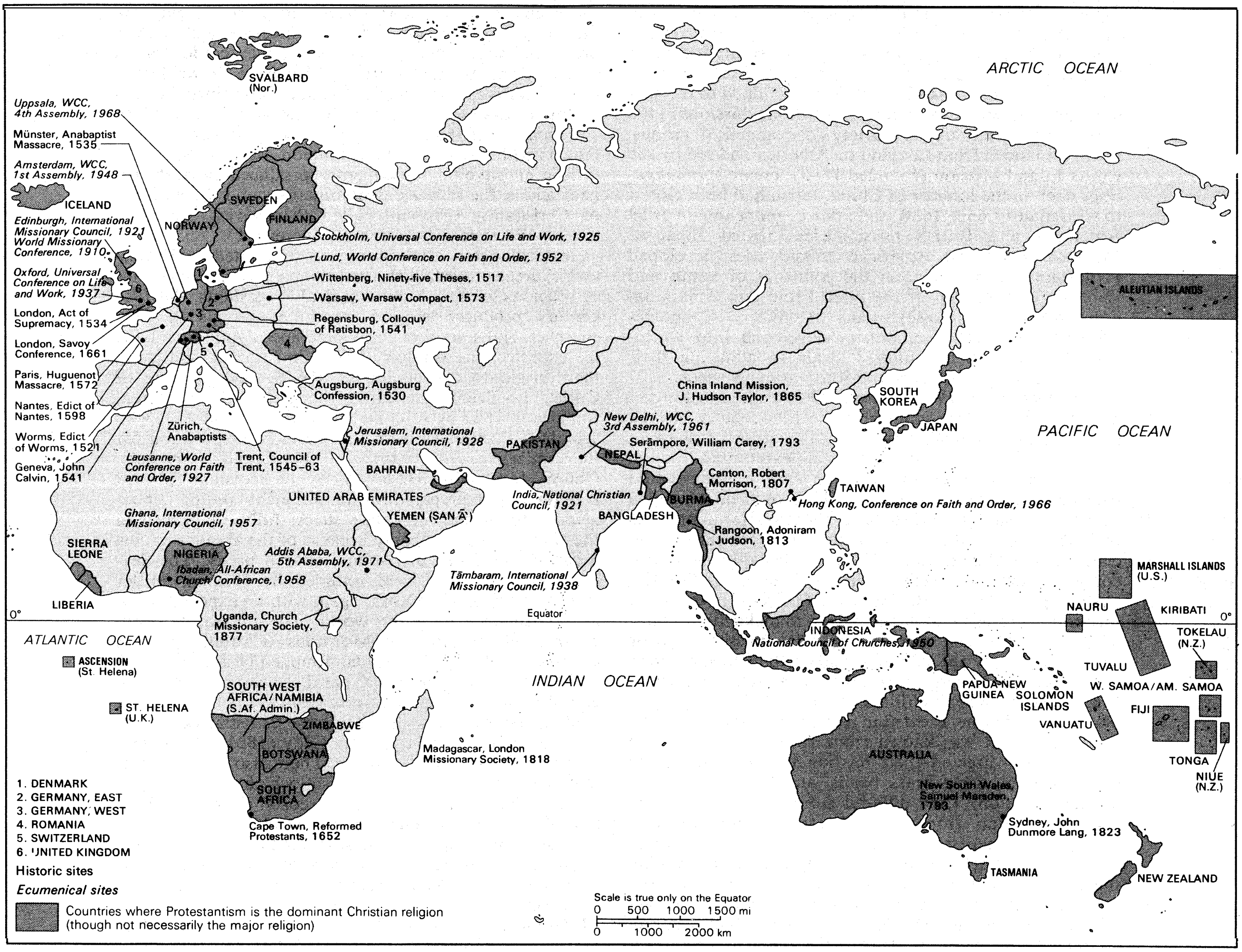
THE 19TH CENTURY

Effects of the American and French Revolutions. The American Revolution and the French Revolution changed history and within it the history of the Protestants. The American Constitution, with its inferred separation of state and churches, owed something to the spirit of free churchmanship inherited from colonial days, something to the religious mixture of immigrants from Europe, something to the reaction against the "Church and King" alliance that prevailed in Britain, and something to the secular spirit of the Enlightenment. With the French Revolution and Napoleon, the idea of the secular state became an ideal for many European liberals, especially among the anticlericals in Roman Catholic countries. The American pattern was probably more influential than the Napoleonic in Protestant Europe. The Protestant states of Germany, Scandinavia, Holland, Switzerland, England, and Scotland, all accustomed to established Protestant churches. for a time met no strong demand anywhere for disestablishment. In all those countries the members of the free, or dissenting, churches won complete toleration and civil rights during the 19th century, but in no Protestant country was the formal link between state and an established church totally broken during the 19th century, except in Ireland (1871) and in Wales (1914–19), where the Church of England was a minority. But, at least as an outward and historical form, established churches remained in England, Scotland, and all the Scandinavian countries.

Developments in Europe. Early in the 19th century the greatest acts leading to reunion since the Reformation were initiated.

During the later 17th century the states of Europe—especially as they allowed for citizens of more than one denomination—moved toward toleration for all men as long as they were good citizens, though this advance toward tolerance was very slow. The Christian leaders, especially of the new Rational, or Latitudinarian, school, sought to show that the doctrines that divided Protestants from each other (if not Protestants from Catholics) mattered less than the truths upon which they agreed. Among the Lutheran and Reformed, the German theologian George Calixtus (1586–1656) already had sought to prove their essential unity, by showing that the doctrines that divided them were not essential to faith. A Scotsman, John Durie (1596–1680), travelled from England to eastern Germany and from Sweden to Switzerland on practical endeavours to persuade churchmen to unite. In 1631 the Huguenot Synod of Charenton (France) agreed to accept Lutherans who married Reformed or were godparents, without compelling them to abandon their special beliefs, on the ground that there was a sufficient agreement in the essential gospel between the Lutheran and Reformed. Lutherans (except for Calixtus and his school) could not take this view. Neither Calixtus nor Durie had much influence. Leibniz (1646–1716) and the French Roman Catholic bishop Jacques Bénigne Bossuet (1627–1704) corresponded about the possibility of union between Catholics and Protestants, but in vain. In Prussia, with a mainly Lutheran population and a dynasty of Reformed princes, the policy of reconciliation became more effective. In 1708 King Frederick I built a "union-church" in Berlin, with the Lutheran Catechism and the Heidelberg Catechism side by side on the altar. In 1817, in a Prussia stimulated by the national revival that followed the fall of Napoleon (1815), King Frederick William III (1770–1840) used the third centenary of the Reformation to unite the Lutheran and Reformed of

Pre-19th-century efforts at reunion


ARCTIC OCEAN
SVALBARD (Nor.)
Uppsala, WCC, 4th Assembly, 1968
Münster, Anabaptist Massacre, 1535
Amsterdam, WCC, 1st Assembly, 1948
ICELAND
Edinburgh, International Missionary Council, 1921 World Missionary Conference, 1910
Oxford, Universal Conference on Life and Work, 1937
London, Act of Supremacy, 1534
London, Savoy Conference, 1661
Paris, Huguenot Massacre, 1572
Nantes, Edict of Nantes, 1598
Worms, Edict of Worms, 1521
Geneva, John Calvin, 1541
Zürich, Anabaptists
Lausanne, World Conference on Faith and Order, 1927
Trent, Council of Trent, 1545–63
NORWAY
SWEDEN
FINLAND
Stockholm, Universal Conference on Life and Work, 1925
Lund, World Conference on Faith and Order, 1952
Wittenberg, Ninety-five Theses, 1517
Warsaw, Warsaw Compact, 1573
Regensburg, Colloquy of Ratisbon, 1541
Augsburg, Augsburg Confession, 1530
1
2
3
4
5
6
ALEUTIAN ISLANDS
China Inland Mission, J. Hudson Taylor, 1865
SOUTH KOREA
JAPAN
PACIFIC OCEAN
Jerusalem, International Missionary Council, 1928
PAKISTAN
New Delhi, WCC, 3rd Assembly, 1961
Serāmpore, William Carey, 1793
NEPAL
BAHRAIN
UNITED ARAB EMIRATES
YEMEN (SAN'Ā')
India, National Christian Council, 1921
BANGLADESH
BURMA
Canton, Robert Morrison, 1807
TAIWAN
Hong Kong, Conference on Faith and Order, 1966
Rangoon, Adoniram Judson, 1813
Ghana, International Missionary Council, 1957
SIERRA LEONE
NIGERIA
Ibadan, All-African Church Conference, 1958
Addis Ababa, WCC, 5th Assembly, 1971
Tāmbaram, International Missionary Council, 1938
LIBERIA
0°
Equator
Uganda, Church Missionary Society, 1877
ATLANTIC OCEAN
ASCENSION (St. Helena)
ST. HELENA (U.K.)
INDIAN OCEAN
INDONESIA
National Council of Churches, 1950
MARSHALL ISLANDS (U.S.)
NAURU
KIRIBATI
0°
TOKELAU (N.Z.)
TUVALU
W. SAMOA/AM. SAMOA
PAPUA NEW GUINEA
SOLOMON ISLANDS
VANUATU
FIJI
TONGA
NIUE (N.Z.)
SOUTH WEST AFRICA/NAMIBIA (S.Af. Admin.)
ZIMBABWE
BOTSWANA
SOUTH AFRICA
Cape Town, Reformed Protestants, 1652
Madagascar, London Missionary Society, 1818
AUSTRALIA
New South Wales, Samuel Marsden, 1793
Sydney, John Dunmore Lang, 1823
TASMANIA
NEW ZEALAND
1. DENMARK
2. GERMANY, EAST
3. GERMANY, WEST
4. ROMANIA
5. SWITZERLAND
6. UNITED KINGDOM
Historic sites
Ecumenical sites
▨ Countries where Protestantism is the dominant Christian religion (though not necessarily the major religion)
Scale is true only on the Equator
0 500 1000 1500 mi
0 1000 2000 km

Significant sites, missionary expansion, important events (with associated persons), and relative numerical strength of Protestantism in the Eastern Hemisphere.
Data from K. Latourette. Christianity in a Revolutionary *Age* (1961); Harper & Brothers

The Prussian Union

Prussia by royal decree (the Prussian Union), and despite resistance, the union was slowly accepted by the majority of Prussian congregations. Other, though not all, German states succeeded in uniting their Protestant communities about the same time. Many of the more conservative Lutherans, rejecting the Prussian Union, emigrated to the United States.

The rise of American Protestant influence in the world. Since the 16th century, the two centres of Protestant political power had been Germany and England. With German unity effected under Prussia and the rise to world power of Britain, the political force of Protestantism was stronger during the 19th century than at any time since the Reformation. But about 1860 it began to be clear that a third force was emerging in the United States. After 1820 there was an extraordinary development of the American nation. American frontier conditions helped to extend the variety of Protestant forces, and denominations such as the Disciples of Christ, formed in 1832 from revivalist groups, arose. These Protestant denominations in time extended their influence beyond America. Many of the immigrants to America were Catholic, and in time the largest single denomination in the United States was to be the Roman Catholic. But the tone of American leadership and culture remained Anglo-Saxon, liberal, and Protestant, and this was one of the great moments in the history of Protestantism. Many Germans and Scandinavians, usually of the Lutheran persuasion, emigrated to America, and American Lutheranism expanded until it became a centre of Lutheran life and thought of a weight equal to the original homes of Lutheranism in Germany and Scandinavia. Because the Lutheran leadership came largely from European Pietistic groups, the American Lutheran churches tended to be more conservative in theology and discipline than the churches in Germany. The element of revivalism in American Christianity continued throughout the 19th century and helped the concept of a personal Christian faith to penetrate deeply into the American way of life. To an observer like the Frenchman Alexis de Tocqueville (1805–59), the culture seemed to be evangelical: usually individualistic, free, emotional, with far more interest in practical consequences than in theology, and manifold in its variety.

The era of Protestant expansion. With this background of European strength in Germany and Britain, with the rising strength of the United States, and with the longest period of peace that Europe had ever known, the Protestant churches entered their greatest period of expansion. At home, they were confronted by the new cities of the century and developed social services on a scale hitherto unknown, such as in hospitals, nursing, orphanages, temperance work, care of the old, extension of education to the young and to working adults, Sunday schools, boys' and men's clubs in city slums, and the countless organizations demanded by the new city life of the 19th century. Abroad, they carried Protestantism effectively into all those parts of Africa that were not under French or **Por-**

Social service and **missionary activities** tuguese influence, so that in southern Africa the Bantu became largely a federation of Protestant peoples. In India, British and American missionaries steadily increased the strength of the newer Indian Christian churches. In China, Christianity had been hitherto confined to the seaports and the survivors of Roman Catholic missions in the 17th century; but now a variety of evangelical groups, mostly financed from England or America and led by the China Inland Mission (founded 1865), created congregations deep in the interior of China. Japan had been closed to Christianity since 1630, and after its reopening in 1859 American and British missionaries created Japanese Christian churches. American missionaries developed Protestant congregations in the countries of South and Central America. All of the main Protestant denominations—Lutherans, Presbyterians, Anglicans, Congregationalists. Baptists, Methodists-developed into worldwide bodies, and all suffered strain in adjusting their organizations to meet these extraordinary new needs.

Revivalism in England and Scandinavia

Revivalism in the 19th century. One of the most prominent features of Protestantism in the 19th century was the development of revivalist methods to meet the needs of an industrial and urban society. It appeared that though many urban poor seldom went to church, they would listen to evangelical preachers in halls or theatres, or on street corners. Methodists and Baptists, familiar with revivalistic methods, made many strides forward, especially in the United States. These endeavours were not confined simply to reaching the working class. The English Baptist Charles H. Spurgeon (1834–92) secured a large audience in London and helped to make the ministry of Protestant dissent very powerful. His mission was for the most part to the educated rather than to the urban poor. For the lowest end of the social scale, a former Methodist preacher, William Booth (1829–1912), and his wife, Catherine, created in east London the agency of evangelism that was known from 1878 as the Salvation Army. They directed their mission to the men on the street corners, using brass bands, and even dancing, to attract attention. They differed from the Methodist revivalist tradition, from which they had sprung, by their belief in the necessity of a strong central government under a "general" appointed for life, and by abandoning the use of sacraments. Their noise in the street at first met much hostility and even persecution, but by the end of the 19th century the Salvation Army had securely established its place in British life and had become a worldwide organization.

In Sweden a Methodist preacher influenced Karl Olof Rosenius (1816–68), who introduced revivalism into Swedish Lutheranism. He and some disciples also were influenced by the movement that stemmed from Zinzendorf. Though there were links with Pietism, the new movement was quite unlike the little groups of Pietism. The Pietists wanted to gather men to salvation out of the world, whereas the Bornholmers (as they later came to be called in Denmark because of a famous episode in evangelism on the island of Bornholm) wanted to declare salvation to the world. The movement had effects in Norway and Denmark and in the Lutheran Church—Missouri Syqod in the United States but never became as separate as the Salvation Army.

Revivalism in the United States

In the United States the development of revivalism was particularly marked in the expansion of the moving frontier. The memory of the Great Awakening (c. 1725–50) was always powerful, and in halls of cities as well as in the camps of the west, revivalistic preaching methods were effective. Protestantism was exceptionally strong because, in many cases, immigrant groups found in religion that link with their historic past that secular society could not for the time give them. Famous evangelists appeared to meet the need of the cities, especially Charles Grandison Finney (1792–1875) and Dwight Lyman Moody (1837–99).

Thus, some of the evangelistic power in Protestantism of the 19th century was drawn away from the traditional churches of the Reformation—Lutheran, Calvinist, and Anglican—and tended to create new forms of church life and new organizations. These almost always used lay preachers, were far more concerned with bringing the individual to conversion and little concerned with church order, and were sometimes content if they could draw a soul to Christ without worrying if it were drawn into a historic Christian community as understood since the Reformation. Consequently, they developed a tendency, not common before the Pietist movement, to identify Protestantism with individualism in religion. Because the evangelistic endeavours subsequently produced separate organizations, the separate denominations and the varieties of Christianity that still called themselves—and with justice—Protestant were rapidly increased.

The secular state allowed or even stimulated the Protestant churches to establish further and powerful varieties of religious groups. Among radical Protestants, the 19th century produced several important groups or new churches, and several of them were apocalyptic (involving the expected intervention of God in history) and owed their origin to expectation of the Second Coming of Christ. In Britain appeared the Plymouth Brethren, founded in 1827 by John Nelson Darby (1800–82), who separated themselves from the world in preparation for the imminent coming of the Lord. The Catholic Apostolic Church, formed in 1832 largely by the Scotsman Edward Irving, likewise prepared for an imminent coming. Apocalyptic groups and sects were successfully established in the United States, probably because of the absence in new areas of any settled or habitual church polity. The Seventh-day Adventists were founded by William Miller (1782–1849) of New York, again with an expectation of an immediate end of the world. Though not self-proclaimed Protestants, the Mormons (Church of Jesus Christ of the Latter-day Saints), founded by Joseph Smith (1805–44), came out of a parallel waiting upon the end. Another set of groups arose from the revival of faith healing, the most important being the Christian Scientists, founded in 1879 by Mary Baker Eddy (1821–1910), who set up her first church in Boston. There also were holiness churches influenced by Methodism, which taught the gift of perfection in this life, and Pentecostal churches, which looked for an immediate outpouring of the Holy Spirit, such as in speaking in tongues. In the 20th century these Pentecostal churches were to develop into the most expansive of all the many groups that appeared during the 19th century.

Toleration. The great Protestant advance depended in part on the existence of the secular state and toleration. As late as 1715 the Austrian government had denied all protection of the law to the numerous Hungarian Protestants. But after the French Revolution the few survivals of this old church-state unity were rapidly whittled away. Even in countries in which one church was established, all churches were given some form of protection; Protestant groups could spread, though slowly and under difficulty, in Spain or Italy. Even in tsarist Russia, which did not recognize toleration, Baptists obtained a foothold from which they were to build the second largest Christian denomination of Soviet Russia. Wherever western European and American ideas were influential, Protestant evangelists could work fairly freely, especially in the colonial territories of Africa and India.

The issue of church and state relationships

Though the secular state thus helped Protestant (and Roman Catholic) expansion and variety, it also confronted all churches with urgent new problems. The American pattern, in which the state must have no constitutional connection with religion, stemmed as much from the old Congregational tradition as from the ideas of the Enlightenment and was never antireligious in intention. It was influential among the older churches of Europe. In Protestant countries where state and church had been in alliance since the Reformation, the effect was twofold: the state became more neutral in its attitude toward the leading denominations of its territory; and the state church pressed harder toward independence from all forms of state control. Lutheran Germany produced a strong movement toward independence in the mid-19th century. In Scotland the evangelical movement demanded independence from the state in the appointment of ministers to parishes, and when this was refused by the courts and by the government, nearly half the Church of Scotland (1843) under the leadership of Thomas Chalmers (1780–1847) left the established church to found the Free Church of Scotland. The two churches continued side by side (until their eventual reunion in 1929). In Switzerland a Reformed theologian, Alexandre-Rodolphe Vinet (1797–1847), pressed for the separation of church and state and (1845) founded the Free Church.

The Oxford Movement and the revival of Protestant communal orders

In England the move toward independence in a state church was a feature of the Oxford Movement, founded by John Henry Newman (1801–90) in 1833. Here the movement took a course unique in Protestantism. It asserted independence by emphasizing all the Catholic elements within the traditional heritage of Protestantism and so created a school of thought that, though remaining within a Protestant Church, came close to repudiating the Protestant tradition as it was then commonly understood in Europe and America. Newman himself became a Roman Catholic in 1845 and was made a cardinal in 1879. Under the leadership of the survivors, the Oxford Movement brought about a transformation in the worship, organization, and teaching of the Church of England, within the traditional polity of an established and Protestant church. The remarkable sign of this change was the revival from 1840 on of nunneries and from 1860 on of monasteries. In German Lutheranism, under the influence of Pietism, Theodor Fliedner (1800–64) established in 1836 a "mother-house" for deaconesses that became a model for the many successor diaconate orders in Germany, Scandinavia, and the United States. These were the first such to appear in Protestant communities since the dissolution of monastic communities during the Reformation. In the mid-20th century France produced one celebrated community at Taizé, devoted to ecumenical prayer and study.

On the whole, the trend was always, though slowly, toward a free church in a free state. A few powerful conservative theorists, especially Friedrich Julius Stahl (1802–61) among German Lutherans, strenuously defended one version or another of the old link between throne and altar, and the necessity for a single privileged church if revolution or rationalism were to be avoided. These theorists were usually viewed, however, as survivals from a past age. Much more powerful and contemporary were the theorists who, in resisting the trend toward denominationalism and pluralism, saw the church as the religious side of the nation and therefore wanted to broaden its doctrines and liberalize its polity. In England Frederick Denison Maurice defended the established church upon these liberal lines; and in Denmark, more easily because the population was so largely Lutheran, Nikolai Fredrik Severin Grundtvig (1783–1872) shrank from every form of denomination or confessionalism and wanted to make Christianity the spiritual aspect of Danish national life. Grundtvig's movement had extraordinary success; but Denmark, and to a lesser extent Sweden and Norway, were exceptions to the trend. The older Protestant churches steadily moved fuither away from the state and unsteadily but gradually secured more autonomy in their organization.

Churches and social change. Attacks on the churches during the 19th century (and after) were twofold: intellectual and social. On the one hand were the thinkers who declared that the advance of science and of history proved the Bible, and therefore Christianity, untrue. On the other hand, the city and industry created a proletariat estranged from religious life. Many of the political leaders, especially in Europe, claimed that the churches were bulwarks of that order of society which must be overthrown if justice was to be secured for the working man. Some of the earlier forms of socialism were atheistic or at least deistic and suspected free churches as fiercely as they suspected an alliance between altar and throne. Social and economic thinkers, such as Karl Marx (1818–83), told the working man that religion was the opium of the people, that it bade men be content with their lot when they ought to be discontented.

Christian Socialists and the Social Gospel

In response to such views, in nearly every European country, Catholic or Protestant, there came into existence groups of "Christian Socialists," men who believed (at least) in the doctrine that the working man had a right to social and economic justice, and that a Christian ought in conscience to work toward those political conditions that would achieve more social justice for the working man. Except for these basic views, the Christian Socialists varied greatly in their outlook and ideas, whether political or theological. Adolf Stocker (1835–1909), a court preacher in Berlin, was an anti-Semitic radical politician, and Charles Kingsley (1819–75), a clergyman novelist in England, was a warmhearted conservative who deeply sympathized with and understood the working man. The most profound of all the Christian Socialists was Frederick Denison Maurice (1805–72), a theologian of King's College in London till he was ejected in 1853, then a London pastor, and finally a professor of moral philosophy at Cambridge.

But in England and America, the radical wing of Protestants—especially Baptists and primitive Methodists—did as much for the workingman's religion as the intellectual leadership of a few Anglican theologians. In some cases the endeavours made Socialist parties possible for the Christian voter, in others they persuaded Christian voters or politicians—without actually voting for a Socialist party—to adopt policies that led toward a welfare state. Nevertheless, they made Christians more conscious of a social responsibility. In America the Social Gospel excited much influence in the churches at the end of the 19th century, and its most influential leader was a Baptist, Walter Rauschenbusch (1861–1918). Whereas in Catholic countries political parties arose that especially appealed to Christian voters and often used the word Christian in their name, in all the Protestant countries all political parties needed to appeal to Christian voters, and few avowedly secular parties have as yet had political success (except the Social Democrats in Germany after 1918 and the Communists in East Germany after 1945). Even the Nazis in Germany at first needed to make a show of defending the Christiantity of the Germans against Bolshevism.

Biblical criticism. Besides political, social, and economic criticism, Protestantism was encountering an intellectual onslaught on Christianity. The question of biblical criticism was first posed in the German universities; *i.e.*, whether a man might be a Christian and even a good Christian though he held some parts of the Bible to be not true. This became the great question for Protestantism, if not for all Christendom, in the 19th century. On the one hand, Protestantism stood by the Bible and declared that the truth of God came from the Bible. On the other, it rested in part on a fundamental conviction of the liberty of the human spirit as it encountered the Bible. Protestantism was thus seldom friendly to the tactic of meeting sane and reverent argument merely by excommunication or by the blunt exercise of church authority. The theological faculties of German universities, being state faculties and not church institutions, suffered much internal stress, but they arrived at last at the conviction that reasoned criticism-even when it produced conclusions opposed to traditional Christian thinking—should be met rather by refutation than by the way of authority. Thus German Protestantism showed at length an elasticity, or openmindedness, in the face of new knowledge, which was as influential in the development of the Christian churches as the original insights of the Reformation. Owing in part to this German example, the Protestant churches of the main tradition—Lutheran, Reformed, Anglican, Congregational, Methodist, and many Baptist communities—adjusted themselves relatively easily (from the intellectual point of view) to the advances of science, to the idea of evolution, and to progress in anthropology or comparative religion.

Influence of philosophical thought on biblical criticism

In such a flux of ideas, with the Protestant tradition seemingly under attack from Protestants, there was naturally a wide variety of approaches, both in philosophy and history. There was an opinion, represented by the German philosopher G.W.F. Hegel (1770–1831), that Christianity should be restated as a form of Idealistic philosophy. This view was influential for a time in Germany and afterward among Oxford philosophers of later Victorian England. Such restatements were subjected to destructive attacks, of which the most powerful were published by the Danish philosopher Søren Kierkegaard (1813–55), chiefly because such reasoned philosophy failed altogether to account for the depths and tragedies of human existence. An earlier opinion sought to base the justification of Christian faith in the religious feelings commonly found in humanity. A German philosopher, F.D.E. Schleiermacher (1768–1834), sought to infer the Christian and biblical system of thought from an examination of human religious experience. Schleiermacher's attempt had much influence on Protestant thought. Throughout the 19th century the appeal to religious experience was fundamental to liberal Protestant thinking, especially in the attempt to meet the views of modern science. Probably the most important of the successors of Schleiermacher was Albrecht Ritschl (1822–89), who wholly rejected the ideas of Hegel and the philosophers; he distinguished himself sharply from Schleiermacher by repudiating general religious experience and by resting all his thought upon the special moral impact made by the New Testament on the Christian community. Between 1870 and 1918 the Ritschlian school was one of the leading theological schools of thought within the Protestant churches.

Historical criticism of the Bible

Meanwhile, scholars made long strides in the study and exposition of the Bible. Freed from the necessity of defending every one of its details as historical truth, professors at Protestant universities were able to put the books of the Bible into a historical setting. This made an important difference in the study of the New Testament but was a revolution so far as the Old Testament was concerned, where the entire earlier accepted chronology was changed. German Rationalist or Hegelian historians were the first to study the problems with freedom. Ferdinand Christian Baur (1792–1860), of the University of Tübingen, applied the methods of Hegelian philosophy to the documents of the New Testament, which he conceived to be products of the clash between the Jewish Christians led by Peter and the Gentile Christians led by Paul. This theory, known as the Tiibingen theory, soon receded in influence; but in aid of this theory, Baur expounded the texts with such ability as to make his study a landmark in the study of the Bible. Among a large number of excellent biblical students, Joseph Barber Lightfoot (1828–89) of Cambridge finally demolished the Tiibingen theory by showing the 1st century origin of most of the New Testament texts; and Adolf von Harnack (1851–1930) of Berlin by the end of the century summarized the results of a century that was revolutionary in the area of biblical study.

IV. The 20th century

NEW CHALLENGES

The effects of wars. The war of 1914–18 broke Europe's waning self-confidence in the merits of its own civilization. Since it was fought between Christian nations, it weakened worldwide Christianity. The seizure of power by a formally atheist government in Russia in 1917 brought a new negative pressure into the world of Christendom and sharpened the social and working class conflicts of western Europe and America. During the following 40 years the Protestant churches suffered inestimable losses.

The effect of Nazi control in Germany

Germany under Adolf Hitler (in power 1933–45) professed to save Europe from the threat of Bolshevism; and the Nazi rule was at first welcomed by many German churchmen. Disillusionment was not slow to follow. From September 1933 there already existed a partial schism between churchmen willing to cooperate with the government in church matters—especially over the Aryan clause that demanded that no Jew should hold office in the church—and those, led by Martin Niemoller (1892–   ), who were not willing to cooperate in church matters. With the support of the state-aided Lutheran churches in the south (Bavaria and Wiirttemberg), Niemoller's group was able to form the Confessing (or Confessional) Church, and the schism was made manifest when the Confessing Church held the Synod of Barmen in May–June 1934. For a time the Confessing Church was strong throughout Germany; but when the German government provided a less doctrinaire government under the minister of church affairs Hanns Kerrl (1887–1941), the Confessing Church was itself divided—into those who were willing to cooperate and Niemoller's men, who were not willing to cooperate because it was a church government imposed by the Nazi government. At the Synod of Bad Oeynhausen (February 1936) the Confessing Church broke up and was never again so strong. In the later stages, especially during the war of 1939–45 (World War II) when the extreme Nazis secured complete control of Hitler's government, the churches came under increasing pressure and toward the end were struggling in some areas to survive. Bishop Theophil Wurm of Wiirttemberg (1868–1953) was a leader in protesting to the government against its inhumane activities, and Pastor Heinrich Griiber (1891–   ), until his arrest, ran the Buro Griiber, which sought to evacuate and protect Jews. Some church leaders, notably the theologian Dietrich Bonhoeffer (1906–45), paid with their lives for their associations with resistance to the Nazi government.

The results of World War II in Protestant Europe

The end of the war saw Russian armies in control 'of eastern Europe and Germany divided. All the churches in the area came under pressure. Most Germans were evacuated or deported from the three Baltic states of Lithuania, Estonia, and Latvia, and though Lutheran communities remained there, they were subjected to persecution, especially under the rule of Stalin. The Lutherans in Transylvania (Romania) and the Reformed in Hungary came under less severe pressure but were much diminished in numbers. The Protestants of Czechoslovakia, led by the theologian Joseph Hromadka (1899–1969), succeeded in maintaining more dialogue with Marxist thinkers than elsewhere in Europe. From the viewpoint of Protestant strength the greatest losses were suffered through the division of Germany. The settlement between the victorious powers gave large areas of former German-speaking (and largely Lutheran) areas to Poland, and many (approximately 8,000,000) East Germans were expelled; most went to western Germany. The East German state, as constructed in 1945, included Wittenberg and most of the original Protestant homeland. East Germany (the German Democratic Republic) was the sole country in which a Marxist government ruled a largely (70 percent) Protestant population. For a time the Lutheran churches were the chief link between East and West Germany and the annual meeting, or *Kirchentag*, the single expression of a lost German unity. But the building of the Berlin Wall in 1961 stopped this communication and isolated the East German churches. Despite governmental pressure, especially in relation to money, education, and church building, and in the national (and anti-Christian) form of youth dedication, the East German Protestants worked courageously and flourished. The 450th anniversary of the Reformation on October 31, 1967, showed how strong a hold the Protestant churches still possessed over the affections of a large number of people.

In Russia, before the Revolution of 1917 a deeply Orthodox state, the 40 years after the Revolution witnessed a growth in the Baptist community. The flexibility and simplicity of Baptist organization made it in some respects more suitable to activity under difficult legal conditions. In the years after Stalin's death in 1953 there was evidence of rapid advance; but after 1960 the Baptist communities, like the Orthodox, again came under pressure, which at times was severe.

The effect of the decline of the British Empire

The material losses that Great Britain suffered in World War II, and the end of the British Empire in the years after 1947, had serious effects on the Protestant churches in former British territories. The home country could no longer provide resources in money and men to the overseas churches on the same scale, and in a few areas church government was handed over to leaders who were not ready to take over church leadership. But in other areas the change of status for Britain hastened the process of change in leadership that had been proceeding slowly; and some of the failing resources were supplemented from elsewhere, especially from the United States, Canada, and Australia. Thus the so-called younger churches came to be a new fact of world Christianity, led by men who no longer saw the history of Christianity solely through European eyes and had an impatience partly derived from a different attitude to the Christian past. This was to be of primary importance in the ecumenical movement. Meanwhile, the secularizing trend of a technological age, as it developed in all the advanced countries, assailed the old European churches and had an even greater effect upon the areas where the younger churches ministered.

In 1948–49 the Communist seizure of power in China effectively ended Protestant missions there. By 1951 there were hardly any European missionaries in the country, and the Chinese churches had to stand without outside aid in men or money. They came under severe pressure, especially during the so-called cultural revolution in the 1960s. They could no longer evangelize and sought barely to survive.

Theological movements. Meanwhile, a certain reaction could be observed in the Protestant tradition of theology. This was partly due to a general doubt about European liberalism after World War I and particularly due, in its further development, to a reaction against attempts by the Nazis to use liberal theology for some of their views of society.

In both the 19th and 20th centuries, liberal theology met much criticism on the ground that it narrowed Christianity to the limits of what men believed themselves to be experiencing or turned what was objective truth into subjective feeling. Though himself no conservative, Kierkegaard was the most extreme of these critics. All the conservative theologians—including the earliest members of the Oxford Movement in England, the evangelical tradition generally, and those many who stood by the inerrant word of the Bible and in the 20th century came to be called by the name Fundamentalist—opposed the liberals on the same grounds. But in the 20th century there was a reaction even within the liberal camp. Beginning in 1918, a reaction against all theologies emphasizing religious experience was led by Karl Barth (1886–1968) of Basel and Emil Brunner (1889–1966) of Ziirich. This theological movement, called Neo-orthodoxy, widely influenced Protestant thinking in Europe and America. Barth and his disciples regarded their work as a reassertion of the true sovereignty of Scripture and as a return to the authentic principles of the Reformation. In America Reinhold Niebuhr (1892–1971) was almost as influential in reacting against liberal Christian philosophies as they applied to society and to man. Yet, that the questions the older theologians had sought to meet still remained was shown by the influence exerted by the German theologian Rudolf Bultmann (1884–) of Marburg, who sought to "demythologize" the New Testament by discovering its core truths and thus allowing its significance for faith to be more fully disclosed. Refugees from Nazi Germany, such as Paul Tillich (1886–1965), interpreted European developments to Americans.

The rise of Neo-orthodoxy

The continuity of the liberal tradition was strengthened by the rise of the United States to the position of leadership of Protestantism after 1945. Germany was divided, Britain was troubled by the loss of Empire, and the United States was stronger than ever. It could provide men and money on a scale that no other country could equal. Without American help it is possible that some of the younger African churches could barely have survived. American Protestantism, however, was more radical in character, and more varied in its diversity, than Protestantism in Europe. Therefore, the rise of the United States to leadership meant a greater influence by the radical Protestant groups in the general counsels of Protestantism and a keener concern to understand, and if possible remedy, the differences in religion through the ecumenical movement. Without American leadership and money the World Council of Churches, formed in 1948, could not have done nearly so much—*e.g.*, social and relief work—and would have been a weaker body.

The general missionary expansion of Protestantism thus took on a more radical appearance. The traditional Protestant churches still evangelized in Africa, South America, and India. But the most rapidly spreading groups were of the Protestant left wing, such as Pentecostal, Holiness, and other churches. An emphasis on the inspiration and the immediate guiding of the Holy Spirit marked all such movements, which first attracted public notice in America 1905–06 and in Britain from 1907 on. The movement won many converts in Brazil, Scandinavia, East Africa, and North America and in the 1950s and 1960s was numerically probably the most rapidly growing Protestant movement. Often it was associated with services for healing.

Expansion of left-wing Protestant churches

THE ECUMENICAL MOVEMENT

The ecumenical movement was in origin exclusively Protestant (though Eastern Orthodox leaders soon took part), and was at first largely dominated by Protestant thinking. Its origins lay principally in (1) the new speed of transport across the world and the movement of populations that mixed the denominations as never before; (2) the world reach of traditional denominations; (3) the variety of religion within America and the problems that such a variety created; and (4) the younger churches of Africa and Asia and their contempt for barriers raised by events of European history for which they felt no special concern. There was always a strong link with the missions, and an American missionary leader, J.R. Mott (1865–1955), whose travels did as much as anything to raise the various ecumenical endeavours into a single organization, represented in his own person the harmony of missionary zeal with desire for Christian unity. A conference of Edinburgh in 1910, which marks the beginning of the movement proper, was a World Missionary Conference. From it sprang conferences on Life and Work (led by the Swedish Lutheran archbishop Nathan Söderblom, 1866–1931), dealing with practical problems, and on Faith and Order, at which theologians sought to examine their theological differences with sympathy. At first Roman Catholics refused to participate; the Eastern Orthodox participated only through exiles in the Western dispersion; and the Nazi government refused to allow Germans to go far in participating. The end of World War II in 1945 created a new atmosphere, and the World Council of Churches was formally constituted at the Amsterdam conference in 1948. The entire movement depended for most of its money and for part of its drive on the Americans; but its headquarters was in Geneva, and, led by the Dutch ecumenical administrator W.A. Visser t'Hooft, it never lost sight of the fact that the traditional problems of divided Christian Europe had to be met if it was to succeed.

Formation of the World Council of Churches

Many problems remained. Fundamentalist and other conservative Protestants suspected the liberal theological views prominent in the ecumenical movement and refused to take part. Most of the British Methodists were successfully united; most of the Presbyterians and Congregationalists created the United Church in Canada. Crossing the border between episcopal and nonepiscopal church governments, the Church of South India in 1947 was created out of Anglicans, Presbyterians, Congregationalists, Methodists, and Baptists; a number of other unions were successfully brought about. But others failed altogether or made slow headway. An especial failure occurred in Britain in 1969, when a scheme for uniting Methodists to the Church of England failed to secure sufficient Anglican support. The ecumenical movement did more to change the understanding between the churches than to unite their separate organizations. In the United States, the Consultation on Church Union, involving the merger of some major Protestant churches, progressed slowly after its start in 1960, but suffered a setback in the early 1970s.

In the years after 1948 the ecumenical movement brought Protestants into an ever-growing dialogue with the Eastern Orthodox and the Roman Catholics. After John XXIII became pope in 1958, the Roman Catholics at last began to participate in the ecumenical movement. Although the definitions of the second Vatican Council (1962–65) were unacceptable to most Protestants, they had a breadth quite unlike the definitions of the first Vatican Council in 1870 and gave hope to those (usually liberal) Protestants who hoped in time to lower this greatest of barriers raised by the 16th century.

BIBLIOGRAPHY

General: K.S. LATOURETTE, ***A History of Christianity*** (1953), with useful bibliographies; E.G. LEONARD, ***Histoire générale du protestantisme,*** 3 vol. (1961–64; Eng. trans., ***A History of Protestantism,*** 1966–68). Many of the best treatments are given in histories of individual countries or in the histories of the churches in each country. Additional references may be found in OWEN CHADWICK, ***The History of the Church: A Select Bibliography,*** 2nd ed. (1966).

For Puritanism, see W. HALLER, ***The Rise of Puritanism*** (1938); CHRISTOPHER HILL, ***Society and Puritanism in Pre-Revolutionary England*** (1964); P. COLLINSON, ***The Elizabethan Puritan Movement*** (1967); and S.E. MORISON, ***The Intellectual Life of Colonial New England,*** 2nd ed. (1956). For Arminianism, see A.W. HARRISON, ***The Beginnings of Arminianism*** (1926). For Pietism, see K.S. PINSON, ***Pietism as a Factor in the Rise of German Nationalism*** (1934). For Protestant missionary expansion, see K.S. LATOURETTE, ***A History of the Expansion of Christianity,*** vol. 3–7 (1937–45); and S.C. NEILL, ***A History of Christian Missions*** (1964). For the 19th and 20th centuries, see K.S. LATOURETTE, ***Christianity in a Revolutionary Age,*** 5 vol. (1958–62); and S.C. NEILL (ed.), ***Twentieth Century Christianity*** (1961).

For American Protestantism, see H. SHELTON SMITH, R.T. HANDY, and L.A. LOETSCHER (eds.), ***American Christianity,*** 2 vol. (1960–63), a general guide with documents; W.W. SWEET, ***The Story of Religion in America,*** 2nd ed. (1950); and WINTHROP HUDSON, ***American Protestantism*** (1961). For the Social Gospel, see C.H. HOPKINS, ***The Rise of the Social Gospel in American Protestantism,*** 1865–1915 (1940). For the churches under the Nazis, see J.S. CONWAY, ***The Nazi Persecution of the Churches, 1933–1945*** (1968). For the Ecumenical Movement, see R. ROUSE and S.C. NEILL (eds.), ***A History of the Ecumenical Movement,*** 1517–1948, 2nd ed. (1967); and H.E. FEY (ed.), ***The Ecumenical Advance*** (1970).

(W.O.C.)

Protozoa

Protozoa is a phylum of mostly microscopic, one-celled organisms. In number of individuals, protozoans rival the bacteria and are almost as ubiquitous, being found wherever there is enough moisture for active life. Some species are worldwide, while others are restricted to special habitats. Probably the most widely scattered are the free-living marine protozoans that are part of the plankton and the parasites of man, including the amoeba that causes a form of dysentery. Living species of protozoans approach 30,000; probably even greater numbers are extinct.

Earlier, attention was focussed on the parasites that caused malaria and sleeping sickness, stimulating research in tropical medicine. More recently, with improvement in laboratory techniques for cultivating protozoans, these organisms are becoming biochemical tools, replacing conventional laboratory animals in many investigations, such as bioassays of vitamins and drugs, and in ecological research.

GENERAL FEATURES

The variety of form of protozoans is endless, from the flowing, protean blobs of amoebas to the exquisitely fashioned "sunbursts" of radiolarians and the delicate shells of foraminifers. The range in size is considerable, from the small blood parasite ***Babesia,*** only two micrometres (pm) long (a small red blood cell can easily contain a dozen such organisms), to the largest foraminiferan shell, almost five centimetres (cm) in diameter.

Disputed status

Although yrotozoans are often described as one-celled animals, some protozoologists prefer to think of them as being noncellular or acellular protistans, neither animal nor plant. The simplest protozoans, comparable in structure to a cell of higher animals, contain only a nucleus and the usual cytoplasmic inclusions (mitochondria, stored food, etc.). More complex species may be equipped with many nuclei, a variety of locomotor and feeding organelles, and sometimes with fibrils and contractile elements.

Many Protozoa seem to be as closely related to plants as to animals. Certain phytoflagellates, for example, form an apparently continuous series with typical algae. The slime molds are claimed by botanists as well as by zoologists. Accordingly, the protozoologist's concept of the phylum Protozoa is based in part upon arbitrary decisions that may or may not reflect the true relationships of these organisms. To this extent, the phylum represents a somewhat artificial assemblage.

IMPORTANCE

Significance in nature. As primary producers (phototrophs or autotrophs), photosynthetic flagellates share with typical algae a position near the base of the aquatic food chain. These photosynthetic organisms occupy much the same ecological position in fresh and salt waters as that filled by the grasses of terrestrial environments. There is some hope that the plankton of the oceans can be cultivated as "forage" to increase the yields of aquatic animals eaten by man.

The occurrence of population "blooms" and "red tides" reveals a harmful influence of dinoflagellates on other organisms. Certain species of ***Gonyaulax*** and ***Gymnodinium*** produce potent neurotoxins. ***Gonyaulax*** toxin qualifies as one of the most deadly poisons known. Similar neurotoxins from ***Gymnodinium brevis*** and ***G. veneficum*** also have been related to fish kill during blooms. In cultures of ***Gorryaulax monilata*** the toxin seems to be liberated when the flagellates disintegrate at death. Red tides kill marine fish in large numbers; occasionally other marine animals are affected. Mussels may retain dangerous concentrations of dinoflagellate toxin for several months after a bloom. Animals eating such shellfish may suffer severe or fatal effects. Seabirds are sometimes victims, as in the 1968 outbreak along the Northumberland coast in England. Man also has suffered from mussel poisoning. The first well-publicized outbreak occurred in 1927 in the San Francisco area (over 100 cases, 6 fatal); an outbreak occurred in England in 1968 (almost 100 cases, but none fatal),

The hemolytic (blood-destroying) toxin of ***Prymnesium parvum*** (a chrysomonad) has caused much loss of fish during blooms in brakish commercial fish ponds in Israel.

Soil protozoans

The effects of Protozoa on other micro-organisms in the soil have been disputed. The usual soil Protozoa are consumers (heterotrophs) rather than producers. One theory is that soil amoebas and ciliates are detrimental to maintenance of soil fertility because they reduce the number of beneficial bacteria. There is, however, no convincing evidence to confirm this theory. Furthermore, some observations suggest that ammonification and nitrogen fixation proceed more rapidly with Protozoa in the soil. The resolution of such a contradiction requires more carefully controlled observations than those reported so far.

Agents of disease. The importance of Protozoa in causing disease is well known. Endemic malaria, due to sporozoan parasites of the genus ***Plasmodium,*** remains a public health problem of major importance throughout much of Africa, the Pacific Islands, and southern and southeastern Asia. Estimates place the annual number of cases at about 100,000,000, with perhaps 1,000,000 deaths. Among the other sporozoan parasites, Coccidia ***(Eimeria, Isospora)*** are important primarily for their effects on domestic and certain game animals. They cause serious problems for poultry raisers. ***Eirneria bovis*** can be an important pest, particularly in young calves.

Visceral leishmaniasis (kala-azar), which attacks deep tissue, is still rated an important cause of death in southeastern and southern Asia and, to a lesser extent, in parts of Africa where the disease is still endemic. Dermal and mucocutaneous leishmaniasis, Old World and New World diseases, are less important overall but are serious enough to the victims.

Infections with trypanosomes, spread by tsetse flies, occur in an area of about 4,000,000 square miles in tropical Africa. Parasites of man cause African sleeping sickness, milder forms of trypanosomiasis, and acute infections in which the patients do not live long enough to develop sleeping sickness. As a secondary effect, even the milder cases may reduce general resistance and thus increase susceptibility of infected individuals to other diseases. Indirectly, man is affected also by those trypanosomes that make it difficult or impossible for him to maintain cattle and certain other domestic animals in some parts of tropical Africa. South American trypanosomiasis, or Chagas' disease, is estimated to affect 7,000,000 persons. It is sometimes fatal to children and is a frequent cause of heart failure in adults. The disease is spread by tropical and subtropical blood-sucking bugs.

Intestinal amebiasis, although seldom lethal, can progress to amebic dysentery; and even a mild infection may lead to liver abscess. The number of cases in the United States has been estimated at about 10 percent of the population. The general effects of even the mild cases result in an appreciable loss of human efficiency that cannot be estimated accurately. Toxoplasmosis has been rated as one of the most prevalent infections of man and domestic animals. The causative organisms (*Toxoplasma* species) are intracellular parasites probably spread by cats. Reported estimates of incidence are: man, 30–60 percent of urban populations; cattle, 1–2 percent; hogs, 24 percent; and sheep, 9–10 percent. Human infections are commonly asymptomatic, but an acute infection often occurs in infants or fetuses (congenital infections); and Toxoplasma seems to be responsible for many abortions, cases of hydrocephalus, and often a mild to serious mental deficiency in surviving infants.

Toxoplasmosis

Laboratory tools. The discovery of the malarial parasites and then trypanosomes and other important parasites of man furnished the initial practical stimulus for interest in the Protozoa. Several European nations with colonial interests in the tropics established institutes for research on tropical diseases, thereby stimulating research in protozoology. More recently, with increased knowledge of protozoan physiology and marked improvements in laboratory techniques, a number of species have been grown in pure culture (axenically) in chemically defined media. Determination of specific growth requirements has been possible in such cases. The result is an array of biochemical tools that can replace conventional laboratory animals in many investigations, with cost and time savings. Protozoa have been used extensively in bioassay of vitamins in biological materials and in estimating the vitamin content of fresh and salt waters. A few species have helped in tracing the action of drugs (antihistamines, herbicides, anticonvulsants, thalidomide, tumour-inducers, etc.), often with the aim of identifying metabolic and morphological effects on cells. Tissue cultures of parasitic Protozoa have been tried as substitutes for the natural hosts, the aim being closer and more precise observation of such parasites, which may lead to better methods of control.

NATURAL HISTORY

Ecology. Distribution. Free-living protozoans live in the soil and in fresh and salt water. Parasites occur in body cavities and tissues, some inside tissue cells. Differences among free-living species depend on the nature and abundance of food and other ecological factors. Other members of the community may significantly influence survival of any protozoan species. Laboratory analysis of such relationships is difficult. With direct competition, one species may replace another, or the competitors may both survive in reduced numbers. A species also may benefit from metabolic products of another, or antibiotic products may harm susceptible associates. Distribution of parasites depends upon distribution of their hosts.

Certain organisms form colonies characteristic of particular groups. Discoid and spheroid colonies—Gonium, Pandorina, *Volvox*, and certain ciliates—contain organisms embedded in a jellylike matrix. In branching colonies the framework is often a stalk, as in Epistylis and *Zoothamnium*, or plates (loricae) attached in a specific pattern, as in Dinobryon and Hyalobryon. Temporary aggregations include chains of dinoflagellates and phytomonads.

Environmental factors. In unravelling ecological relationships, conclusions based on field studies are not as precise as might be desired, while more specific conclusions from laboratory experiments often can be applied to natural conditions only with reservations. It is clear, nevertheless, that survival is influenced by factors characteristic of a particular environment. These factors include food supply, light (for photosynthetic species), temperature, oxygen and carbon dioxide (CO_2) tension, aci ty of the medium, and, for active life, the necessary water. Food requirements are critical. For photosynthetic flagellates, phosphorus and nitrogen sources are most likely to limit their distribution. Certain vitamins (thiamine, B_{12}, biotin), although their fluctuations influence the abundance of organisms requiring them, are seldom scarce enough to prevent bare survival of some protozoans. Except for an organic energy source, chlorophyll-free phytoflagellates need much the same foods as green flagellates. Other heterotrophic organisms, requiring more organic foods, may be more restricted in ecological relationships. In nature, however, they seldom suffer seriously from vitamin deficiency in tolerable environments because organic foods usually contain sufficient vitamins. Consumers may show certain dietary "preferences" that might influence distribution.

Food requirements and temperature range

Increases in food supply may cause striking increases in populations, as in blooms and red tides. Dinoflagellate red tides often follow heavy rains that carry food-laden water into the coastal waters. The increased food stimulates aquatic bacteria to produce vitamin B_{12}, which in turn promotes dinoflagellate growth. Data on distribution show that a few Protozoa can tolerate high temperatures. Among strains of Tetrahymena *pyriformis*, one from a cold mountain stream grows well at 35° C, while a "tropical" strain has failed to survive at that temperature. Parasitism may add complications. Young stages (mosquito phase) of Plasmodium *relictum* are injured quickly at 35" C; older stages last three to four days, and sporozoites (infective for pigeons) survive about two weeks. Unfavourable effects of high temperatures may depend partly upon increased food requirements. In cultures of *Ochromonas* shifted from 34" C to 38° C, the thiamine requirement is increased about one thousandfold, the B_{12} requirement even more.

There is little evidence that low temperatures are a serious natural hazard to Protozoa. In laboratory tests, free-living and parasitic species have survived temperatures well below freezing. At —95" C about 11 percent of a *Trichomonas* foetus population survived more than 5½ years. At the same temperature, Tetrahymena pyriformis can live at least four months. In these experiments, however, the medium contained a chemical protectant; survival time may be much shorter under more nearly natural conditions.

In relation to oxygen requirements, data on natural distribution may need careful interpretation. Many protozoan parasites of deep tissues are strict anaerobes (require an oxygen-free environment); others can adjust to a range of oxygen levels but do best at the lower end of the range. Even among oxygen-requiring protozoans (aerobes), different species are favoured by different oxygen levels. Some protozoans live under conditions apparently insuring anaerobiosis—ciliates in the rumina of cows, flagellates of the termite intestine, and ciliates in the bottom ooze of deep lakes or in sewage tanks. Dissolved oxygen in ponds and lakes increases in winter and spring, favouring aerobes, and decreases in summer and fall, favouring anaerobes. Less important daily fluctuations in shallow waters may be correlated with photosynthesis. Carbon dioxide, dissolved gas, or bicarbonates show fluctuations inversely proportional to those of dissolved oxygen. An adequate supply of CO_2 is essential for photosynthetic species, but little or nothing is known about quantitative needs of consumers. Bottom, or benthic, zones of ponds and lakes with dense bacterial populations show a high CO_2 concentration, which could possibly rise to levels unfavourable to certain Protozoa. The acidity–alkalinity range (pH) of the medium can limit the distribution of certain species. There is some evidence that protozoan population growth is most rapid at a particular pH level, which may vary with the medium. The general range for growth is approximately 2.2–9.2, although few species (*e.g.*, Polytomella caeca) can grow throughout most of this range. In general, the pH of the medium affects uptake of foods, so that pH relationships vary with the

pH optima

substrate. The rate of ingestion and activity also are influenced by pH. In films of moisture on soil particles, active life goes on in the upper few inches of the soil, where protozoan populations probably average 40,000 to 50,000 per gram. In rich soils populations may be several times as great. Small flagellates (Cercomonas, *Heteromita*, Oicomonas) and amoebas (Hartmannella, Naegleria) are usually predominant; ciliates (Colpoda species) form a small percentage of the population. Indigenous species represent perhaps 10 percent of the 250 or so reported; the rest are occasional invaders. Sand, as distinct from ordinary soil, also has a characteristic protozoan fauna including over 100 species of ciliates and a unique euglenoid (Pentamonas).

Species interactions. Symbiosis involving Protozoa includes associations in which the host is larger than its protozoan symbiotes and associations in which the protozoan is the host of symbiotic micro-organisms. Representing the first type, termite flagellates ingest wood chips swallowed by the insect and produce materials useful both to themselves and to the host. The hayeating ciliates of the cow's rumen perform a similar service. Dinoflagellates symbiotic in corals stimulate calcium metabolism of their hosts, possibly by contributing growth factors. In relationships of the second type, a strain of *Crithidia* oncopelti carries a bacterial symbiote that synthesizes useful amino acids. In Paramecium *bursaria* the symbiotic alga Chlorella contributes useful material; during starvation, a decrease in number of Chlorella suggests that the ciliates consume their algae. Association involving infection of a protozoan parasite is often termed hyperparasitism; intestinal flagellates, amoebas, and ciliates may serve as hosts of small amoebas, microsporidians, or bacteria.

Survival of a protozoan parasite may depend upon particular substances available in its natural host. *Trypanosoma* lewisi, a rat parasite, can live only a few hours in laboratory mice; however, as long as the mice are injected daily with normal rat serum, T. lewisi can survive. In some cases, protozoans can adapt to an unnatural host. Plasmodium malariae, which normally causes malaria in man, can produce infections in chimpanzees; likewise, P. cynomolgi of monkeys can be transferred to man by a mosquito and also from man to man and back to the monkey.

Pathogens and disease

Virulent strains of a pathogen can invade susceptible hosts and cause appreciable damage. In two-host cycles the vertebrate host is usually the one damaged by the parasite; Trypanosoma rangeli, in contrast, is pathogenic in the insect vector but relatively nonpathogenic in vertebrates. Pathogenicity may appear occasionally in species usually considered harmless—Naegleria gruberi was reported in fatal cases of amoebic meningoencephalitis. Virulence is inherited in strains of a species and may be controlled by deoxyribonucleic acid (DNA). Treatment of a nonvirulent strain of Trichomonas *gallinae* with homogenates of a virulent strain transformed the former into the latter type. More recently, a 1:10 mixture of DNA and ribonucleic acid (RNA) from a virulent strain produced the expected effect.

Life cycle and reproduction. Protozoan life cycles include reproductive stages and a stage for dispersal; the latter may or may not be encysted in a protective envelope. The period of active living (trophic phase) may, in some species, include more than one stage and may also involve a sexual stage. The survival value of sexual phenomena varies. In some species, sexual activity is sporadic. In others, syngamy, coupling to form a single cell or zygote, may be essential; in malarial parasites, for example, syngamy must occur within the mosquito for completion of the parasite's cycle (see Figure 1).

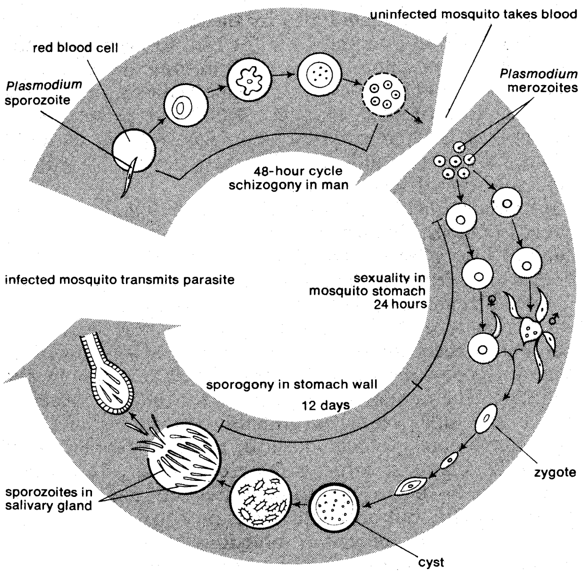

uninfected mosquito takes blood
red blood cell
Plasmodium sporozoite
Plasmodium merozoites
48-hour cycle schizogony in man
infected mosquito transmits parasite
sexuality in mosquito stomach 24 hours
♀
♂
sporogony in stomach wall 12 days
zygote
sporozoites in salivary gland
cyst

Figure 1: Life cycle of *Plasmodium vivax*.
From The *Science* of Biology by P.B. Weisz, Copyright 1966. Used with permission of McGraw-Hill Book Co.

Range of life cycles. The simplest life cycles include a trophic phase and a cyst; ability to encyst has been lost in some species. Modifications include more than one trophic stage, different kinds of cysts, and a sexual phase. The simplest modifications involve the trophic phase—amoeba/flagellate alternations in certain mastigophorans and sarcodines; amoeba/plasmodium in certain sarcodines; and flagellated/nonflagellated stages in certain mastigophorans. With the addition of sexual phenomena, modifications become more complicated (as in coccidians and malarial parasites). Metamorphosis is involved in cycles with two or more trophic stages. In Leishmania, a leptomonad in an insect host or a culture alternates with a nonflagellated leishmania1form in vertebrate cells. In certain trypanosomes vector stages contain a functional cytochrome system; stages from vertebrate blood lack this system (replaced by a glyceraldehyde dehydrogenase system for electron transport). Temperature may affect such cycles. Trypanosoma conorhini remains flagellated in cultures at 37" C, but at 28° C it assumes the characteristically flagellated vector form.

Encystment

Protective cysts typically enclose a dedifferentiated stage, in phytomonads a zygote. Coccidian macrogametes (*e.g.*, Eimeria) may produce such a cyst (oocyst) just before syngamy, leaving a pore for entrance of a microgamete. Also in coccidians, division products (sporoblasts) of the zygote may produce spores inside the oocyst. In the usually compound wall of the protective cyst, the outer or middle layer is typically thicker than the others. Protective efficiency varies. Cysts of Didinium remain viable for years if kept moist; cysts of Colpoda resist drying for several years. Soil amoebas and flagellates have survived for 49 years in dry soil samples. Such cysts are important in dispersal. Reproductive cysts (enclosing stages of fission or sometimes gametogenesis) have thin walls. Reproduction is not limited strictly to reproductive cysts, however; in Eimeria and related genera, the encysted zygote divides.

Production of a protective cyst in ciliates usually involves resorption of cilia, and the organism often rounds up and undergoes partial dehydration. Respiration may drop to about one-tenth'the trophic rate. Prior to secretion of the cyst wall, some species accumulate reserves such as glycogen. Chromatoid bodies also are reserves to be used during encystment. Induction of encystment has been attributed to starvation, vitamin deficiency, evaporation of the medium, critical pH changes, etc. Excystment and resumption of the trophic phase occurs when environmental conditions are again favourable.

Reproduction. Protozoan reproduction is not necessarily associated with sexual processes. Syngamy, the fusion of two protozoans, for example, produces one organism (the zygote); after syngamy, reproduction may begin. This process must sometimes initiate a reproductive phase.

Reproduction involves nuclear division and the separation of newly formed units. Separation may occur through simple cell division (binary fission), budding, plasmotomy, or schizogony (see Figure 2).

Binary fission

Binary fission, much like mitotic cell division in higher organisms, produces two new daughter organisms of equal size (see CELL AND CELL DIVISION). Budding pro-

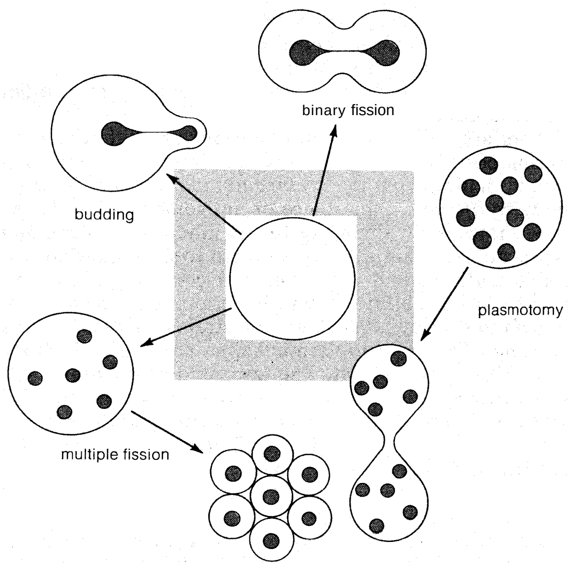


Figure 2: Asexual reproductive patterns in protozoans.

From The Science of Zoology by P.B. Weisz, Copyright 1966. Used with permission of McGraw-Hill Book Co.

duces one large and one small organism. Buds usually develop at the surface, but in certain suctorians they develop within a brood pouch, from which they leave as larvae. Schizogony, common in sporozoans (see Figure 3), is synchronized production of a number of young from a multinucleate stage. In small schizonts of Eimeria, bodies called merozoites are separated from one another by a sort of cleavage without the usual peripheral budding typical of the larger schizonts. Schizogony probably favours transfer of sporozoan parasites to a new host and also may facilitate completion of a reproductive phase before the host develops an immunity. Plasmotomy involves the division of multinucleate types into a number of organisms.

Reproduction may occur in an active or an inactive stage. Most ciliates reproduce in motile stages, but Colpodidae, like certain phytoflagellates, divide within a reproductive cyst. Fission may involve extensive reorganization. Hypotrichs (*e.g.*, Euplotes) may resorb ventral tufts (cirri) and produce new ones from new groups of special organelles called kinetosomes. In a holotrich, new kinetosomes arise next to old ones in alternate fashion; hairlike cilia then sprout from the kinetosomes during elongation of the ciliary rows during fission. Formation of a new mouth may involve extensive reorganization. In dinoflagellates that divide the old cell covering, or theca, the two daughters must build different missing portions. Shelled rhizopods may secrete new plates or collect extraneous materials for building a shell in the next fission. Behaviour of organelles in fission varies—the kinetoplast apparently divides, while the parabasal body is typically resorbed and replaced.

From The Science of Zoology by P.B. Weisz, Copyright 1966. Used with permission of McGraw-Hill Book Co.

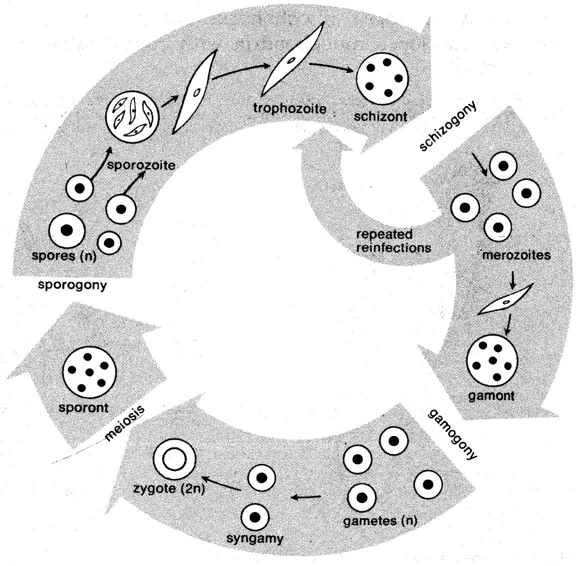


Figure 3: Basic sporozoan life cycle.

Regeneration after amputation has been investigated especially in *Stentor*, in which a fragment less than 1/100 the adult volume may regenerate a complete organism. In order for regeneration to occur, a micronucleus is essential in certain ciliates but not in others, although a portion of the macronucleus is always needed. Regeneration of flagella and cilia is a common phenomenon. Chlamydomonas loses its flagella during fission, and each daughter cell soon regenerates new ones. Hypotrichs also resorb and then regenerate locomotor organelles during fission.

Sexual phenomena. Sexual stages added to the life cycle include production of special gametic nuclei and formation of a fertilization nucleus (synkaryon) by fusion of such nuclei. Chromosome reduction (meiosis) occurs in early divisions of the synkaryon giving rise to a haploid phase (with a single set of chromosomes), from which the sex cells (gametes) are produced. This type of meiosis occurs in phytomonads, sporozoans, and certain parasitic flagellates. Gametic meiosis occurs in the production of sex cells (gametogenesis) of diploid organisms (with double sets of chromosomes) such as heliozoians, which undergo nuclear reorganization without benefit of a sexual partner (autogamy). Intermediate meiosis occurs in foraminiferans, with an alternation of generations, a haploid phase alternating with a diploid phase in which meiosis occurs. Syngamy may be between similar gametes (isogamous) or between obviously different gametes (anisogamous). Apparent isogamy may involve physiologically differentiated gametes (+ and − types), as in Chlamydomonas. Colonial phytomonads may show heterothallic and homothallic strains as well as monoecious and dioecious strains. Homothallic strains produce both types of gametes within a single colony. Heterothallic strains contain two kinds of flagellates; both must be present to insure gametogenesis and syngamy. In Gonium pectorale (strains often heterothallic) each flagellate becomes a gamete and leaves the colony. Gametes pair, and each resulting zygote encysts. Germination with meiosis produces a four-cell stage, which soon divides into four 16-cell colonies. In Pandorina morum (usually heterothallic) germination involves meiosis and degeneration of several nuclei, leaving one haploid flagellate to produce a new colony. In Volvox, a few reproductive cells are scattered in the colony; in Eudorina, reproductive cells are limited to the posterior hemisphere. In the life cycle of V. aureus a reproductive cell may produce microgametes (sperm), daughter colonies, or a macrogamete (egg). Released sperm packets stick to a presumptive female colony whose matrix promptly develops an opening; the microgametes enter and adhere to a reproductive cell, stimulating it to become a macrogamete. The complementary cells then fuse, becoming a zygote. Differentiation of an immature colony into a male is induced by a heat-sensitive factor liberated from mature male colonies. A comparable factor in V. carteri stimulates gonidia to become female colonies. In Chlamydomonas eugametos, a factor from female gametes induces differentiation of male gametes, and vice versa. Sexual activities are complicated by mating types and "varieties" (syngens) analogous to those of ciliates. Strains belonging to different syngens may be unable to mate. Compatible + and − strains produce gametes when mixed, and fusion results; with incompatible strains, no gametes appear.

Syngamy: fusion of gametes

Foraminiferans produce flagellated or amoeboid gametes that fuse in syngamy. Certain life cycles are complicated by alternation of generations comparable to that of plants. The diploid forms reproduce asexually and sooner or later undergo meiosis to produce haploid forms; these reproduce asexually also but eventually produce gametes. Syngamy forms a zygote, which develops into a new diploid form, and the cycle begins anew.

Conjugation involves ciliates belonging to complemen-

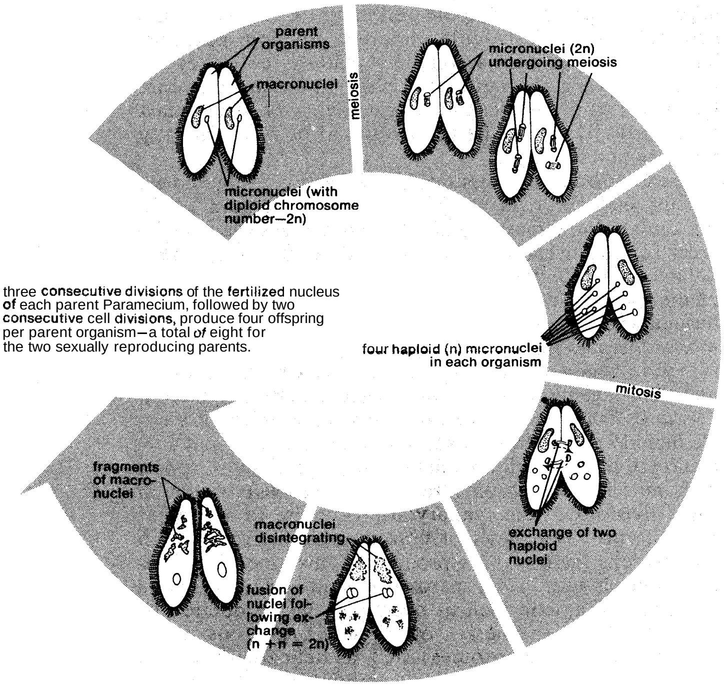

parent
organisms
macronuclei
meiosis
micronuclei (2n)
undergoing meiosis
micronuclei (with
diploid chromosome
number—2n)
three consecutive divisions of the fertilized nucleus of each parent Paramecium, followed by two consecutive cell divisions, produce four offspring per parent organism—a total of eight for the two sexually reproducing parents.
four haploid (n) micronuclei
in each organism
mitosis
exchange of two
haploid
nuclei
macronuclei
disintegrating
fusion of
nuclei following exchange
(n + n = 2n)
fragments
of macronuclei

Figure 4: Conjugation in Paramecium.
From Biolooical Sciences Curriculum Study, *Biological* Science: An *Inquiry* into Life, 2nd ed. (1968); Harcourt Brace Jovanovich

tary mating types within one syngen. Pairing of ciliates from different syngens typically is unsuccessful and results in abortive conjugation. Pairing is induced by specific proteinaceous substances in the ciliate; even free cilia, as in ***Paramecium bursaria,*** agglutinate with intact ciliates of a complementary mating type. After pairing, each conjugant undergoes a complicated nuclear reorganization in which three micronuclear divisions (including meiosis) produce gametic nuclei. Migratory nuclei are exchanged; fusion of a migratory with a stationary nucleus produces a synkaryon in each conjugant (see Figure 4). Each old macronucleus subsequently disappears, while the synkaryon divides into several nuclei, among which survivors differentiate into new macronuclei and micronuclei. Macronuclear differentiation includes chromosomal replication. The developing nucleus increases in size and, in many species, also changes form. The normal nuclear number is restored in reorganization fissions. In ***Paramecium,*** old macronuclear fragments are distributed to each daughter organism. A small amount of cytoplasm is transferred in ***Tetrahymena.*** In ***Vorticella,*** one conjugant is so small that it fails to survive. In autogamy (diploid organisms mostly) a single organism produces haploid nuclei, which then fuse into a synkaryon. In ***Paramecium*** such nuclear activities parallel those of conjugation. Both autogamy and syngamy occur in certain flagellates of the wood roach; in others (Rhynchonympha), only autogamy occurs.

Variation and heredity. In investigations on the haploid ***Chlamydomonas*** and ***Polytoma,*** Mendelian inheritance has been reported for such traits as size, shape, rate of swimming (normal and "lazy"), photosynthesis (normal and slow), storage of volutin (normal and excessive), etc. Linkage has been established for certain pairs of traits, and crossing-over has been demonstrated in several cases.

Observations on traits inherited through conjugation range from size and fission rate to specific enzyme systems, some of the latter observations involving pure cultures of ***Tetrahymena pyriformis.*** For example, the wild strain of ***T. pyriformis*** is unable to synthesize serine and must have this amino acid in defined media. The ability to synthesize serine, known in certain mutant strains, seems to be inherited in Mendelian fashion as a recessive trait. The status of the pyridoxine requirement in T. ***pyriformis*** is similar. Strains that grow without a source of pyridoxine and synthesize it appear to be homozygous recessives. When these are mated with the normal type (pyridoxine-requiring), all the progeny require pyridoxine. Mating types in ***Paramecium,*** *Tetrahymena*, and ***Euplotes*** also seem to be inherited in Mendelian fashion, although the genetic pattern may vary from one variety to another within a species.

There are indications that macronuclear differentiation during postconjugant divisions may influence determination of mating types in ***Paramecium aurelia*** and ***Tetrahymena pyriformis.*** Inheritance of antigenic types in ***P. aurelia*** shows some similarity to the inheritance of mating types, although antigenic types seem to be somewhat less stable under certain environmental conditions. Also inherited is the ability of certain strains of ***Paramecium aurelia*** to destroy other ***P. aurelia*** strains. "Killer" strains of ***P. aurelia*** bear cytoplasmic kappa particles, which secrete killer particles into the surrounding medium. All sensitive strains of paramecia die when they contact killer particles. Kappa particles seem to be self-reproducing, but their persistence in individual ciliates is ultimately dependent upon a gene (or genes).

FORM AND FUNCTION

Form varies in amoeboid types but is fairly constant in species with skeletons, tests, or thick pellicles; the form of colonies is determined by the material holding the colony together. Form may change within a life cycle and also may be modified by environmental conditions. Many protozoan species contain a single spheroid nucleus, whose fine structure resembles that of other nuclei (see CELL AND CELL DIVISION). Some groups, however, have two distinct kinds of nuclei (see Figure 5). The ciliates, for example, have a smaller micronucleus and a larger macronucleus, differing in function and structure. The micronucleus participates in conjugation and autogamy. The macronucleus exerts genetic influence. In addition. ***Paramecium*** cannot survive without a macronucleus,' even with extra micronuclei equalling the macronuclear mass. This may be correlated with the known synthesis of RNA in the macronucleus. A macronucleus is sometimes elongated, branched, or nodular and contains many small granules (chromatin, possibly condensed chromosomes) and usually some nucleoli. Except for ciliates, foraminiferans, and possibly some radiolarians, multinucleate species have nuclei of one kind.

Cytoplasmic inclusions

The cytoplasm. Protozoan cytoplasm contains various particles and organelles. Mitochondria are tiny bodies whose intense enzymic activity is associated with cellular respiration and protein synthesis. The mitochondrion of trypanosomids (*e.g.*, ***Trypanosoma conorhini)*** is tubular and relatively long. Mitochondria1 structure and activity vary with conditions. During growth of ***T. pyriformis,*** mitochondria are normal, peripheral, and may divide; in old populations, they are shorter and smaller, drift into the endoplasm, lose ribosomes, and accumulate lipid globules. Blood stream trypanosomes (T. ***brucei*** group) have poorly developed mitochondria, but cultured forms have well-developed mitochondria with considerable enzymic activity.

From (left) Biolooical Sciences Curriculum Study, Biological Science: An *Inquiry Into Life*, 2nd ed. (1968); Harcourt Brace Jovanovich. (right) *The Science* of *Zoology* by P.B.Weisz, Copyright 1966. Used with permission of McGraw-Hill Book Co.

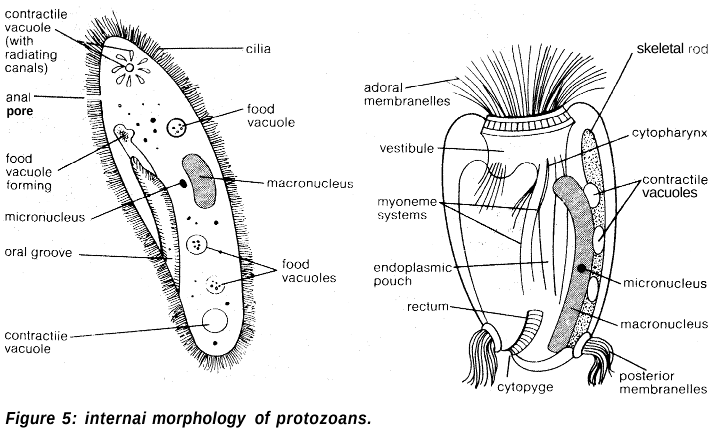


Figure 5: internai morphology of protozoans. (Left) Paramecium. (Right) Spirotrichan ciliate.

Portions of the membranous endoplasmic reticulum are smooth, others rough. Glycogen granules may be synthesized in the reticulum, and rough vesicles underlying the pellicle also may take part in secretion. Ribosomes may be free or associated with the endoplasmic reticulum; they are often clumped into polyribosomes.

Golgi material often consists of stacks of flattened vesicles (cisternae), but single vesicles occur. During acetate starvation of bleached Euglena gracilis Golgi bodies expand, producing lysosome particles that in a short time contain mitochondria and cytoplasmic material and also show enzymic activity. Coccoliths are formed in association with Golgi material in *Cricosphaera cai terae.*

Pigment granules (violet to red) occur in many phytoflagellates and certain other Protozoa. Carotenoids accumulate in certain euglenoids and phytomonads, the predominant type varying with the species. Reddish pigment in Euglena rubra moves to the periphery of the cell with increases in light intensity or temperature. The pink pigment of *Blepharisma* is changed by bright light into a derivative that is toxic to Protozoa (including *Blepharisma*) and rotifers.

Vacuoles. Water elimination vesicles (contractile vacuoles) maintain osmotic equilibrium by eliminating excess water in freshwater species and in certain marine and parasitic species. Hydrostatic pressure on the vacuole causes a discharge of fluid and the collapse of the vesicle. The vesicle is refilled with water from a variety of collecting devices.

Gastrioles

Digestive vacuoles (gastrioles) are formed during ingestion. In Amoeba, part of the cell covering apparently is brought in to line the gastriole. In ciliates with a mouth, a gastriole grows out from the cytopharynx. In Suctoria, gastrioles develop at the bases of tentacles. Flotation vacuoles form a foamy zone of thin-walled vesicles. So-called sensory vacuoles of certain ciliates contain granules (statoliths) of uncertain function.

Pigment bodies. Chromatophores, limited to phytoflagellates (but lacking in many), contain chlorophylls and accessory pigments, some abundant enough to mask the green chlorophyll. Specific pigments vary in different orders. Chromatophores in *E. gracilis* contain DNA and ribosome-associated RNA that are structurally different from nuclear DNA and cytoplasmic RNA. Mutant strains of E. *gracilis* without chromatophores lack a satellite DNA found in normal flagellates. Species of Euglena differ in the presence or absence of pyrenoids, bodies within or attached to chromatophores. Euglenoid pyrenoids are usually covered with paramylum, which may act as a food reserve.

A stigma, or eyespot, common in green flagellates, occurs also in a few colourless species. In Chlamydomonas, the stigma lies inside a chromatophore; in Euglena, it lies in the reservoir from which the flagellum emerges. Composed of one or more plates of reddish globules, the stigma is involved in light reception. The stigma in Euglena periodically shades the paraflagellar body, resulting in flagellar activity directed toward maintaining a favourable orientation to the light source.

Protective coverings. Some phytoflagellates have a covering (theca) composed of cellulose or pectin, resembling a plant cell wall. Mineral-impregnated thecae form rigid coverings and occur as fossils. The dinoflagellate theca, composed of three to four membranes, often shows characteristically arranged plates between the second and third membranes. A conical or tubular rigid covering (lorica), with a wall and an opening for partial extrusion of the organism, occurs in certain flagellates and ciliates. Loricae may adhere to produce colonies.

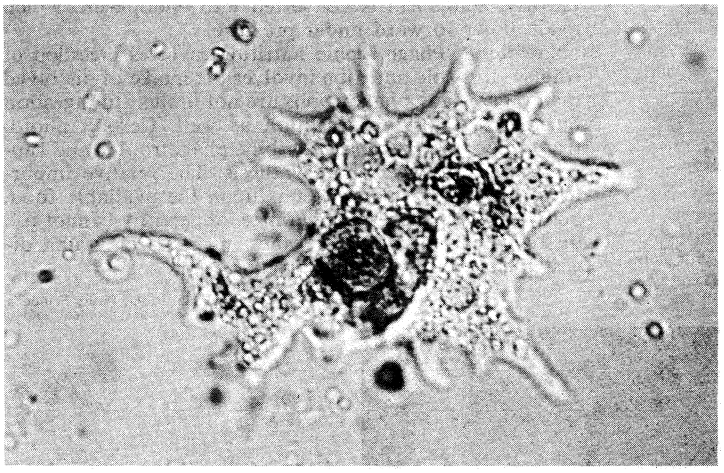
Formation of shells

Foraminiferan shells, or tests, are usually calcareous (sometimes siliceous), often with perforate walls; but *Allogromia* and related types produce a proteinaceous test. Others bind extraneous particles with a glycoprotein. Cytoplasm extruded through the pores may cover the test, or the pores may be filled with a secretion continuous with the calcite layers. Primitive foraminiferans have one-chambered tests with a large aperture; more specialized ones build several to many chambers, in specific patterns. For each new chamber some of the pseudopodia coalesce to form a template on which the chamber wall is secreted. Many radiolarians have siliceous skeletons; those of acantharians are mostly made of strontium sulfate.

So-called protective structures include toxicysts of predatory ciliates (*e.g.,* Dileptus). These organelles contain toxins that can paralyze rotifers and Protozoa, and actually disintegrate them. In Paramecium, a few filamentous trichocysts at a time are partially extruded for anchorage while feeding. Protozoa are covered with a pellicle complex of two or three layers. The pellicle may be decorated with nipples or ridges; in certain gregarines (*Rhynchocystis*) long hairlike cytopilia extend from pellicular ridges. The cortex of ciliates and gregarines, including the pellicle and underlying cytoplasm, is a specialized zone containing a variety of structures.

Appendages. Locomotor and feeding devices include pseudopodia, flagella, and cilia. Most pseudopodia are contraction-hydraulic or two-way sliding types. In the first (*e.g.,* Amoeba proteus), the outer layer is denser than the inner cytoplasm. The second type shows two distally continuous filaments, flowing in opposite directions on opposite sides. These sticky filaments make effective food traps, as in the extensive networks found in foraminiferans. In many heliozoans and acantharians, each axopodium, or "ray," contains an axial filament or cylinder.

Hugh Spencer



The constanfly changing form of Amoeba.

Flagella and cilia

Flagella and cilia have a sheath, continuous with the pellicular complex, and an axoneme, ending basally in a blepharoplast or kinetosome. The axoneme contains one central and nine peripheral pairs of fibrils (peripheral bases forming the blepharoplast). Sometimes the sheath bears lateral fibrils (mastigonemes) that increase the surface and, when rigid, may control direction of the effective force. In Crithidia fasciculata the sticky sheath makes the flagellum a holdfast organelle.

Flagella commonly arise anteriorly and extend forward. Dinoflagellates, however, usually have two fibrils arising near the middle, one extending posteriorly and the other twisted spirally around the body. A trailing flagellum, arising anteriorly, may beat freely or adhere to the edge of a membrane, as in Trichomonas. Flagella range from one to four in most free-living species to dozens in certain parasites. A flagellum may be associated with certain organelles, especially in parasites.

Cilia are arranged in rings or spiral rows. Cilia may be fused into membranes, membranelles, and spiny cirri. Fibrillar systems often seem to join kinetosomes. Contractile fibres, or myonemes, in certain ciliates, gregarines, and flagellates, typically cause changes in form, as when avoiding obstacles, and in the helical coiling of the vorticellid stalk. Myonemes of gregarines and certain flagellates may represent a basic gliding mechanism.

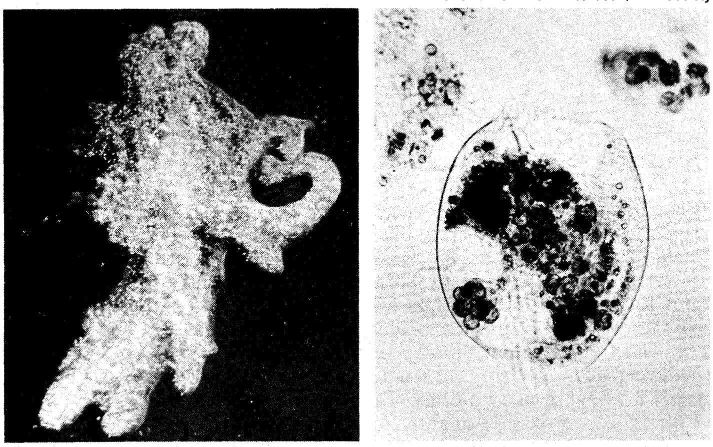
Locomotion. Flagellate locomotion usually involves flagellar activity. In flagellates such as Leptochloris and *Medusochloris,* myoneme-induced contractions expel water from a posterior cavity, driving the organism forward. In Mixotricha paradoxa, swimming results from the svnchronized activity of many spirochetes that attach to and propel the flagellate, whose flagella serve as rudders. Gliding in ***Petalomonas*** involves latero-posterior beating of one flagellum while a second acts like a sled runner. An anterior flagellum also pulls ***Peranema*** and ***Entosiphon.*** The characteristic gliding and creeping in ***Euglena*** (euglenoid movement) are not yet explained satisfactorily. Swimming involves various flagellar activities, while undulatory waves push or pull.

Ciliates use cilia or their derivatives for locomotion; hypotrichs can "walk" on coalesced cilia or cirri. A cilium of ***Paramecium*** beats in a helical wave from base to tip. In swimming forward the ciliary movements in a row pass posteriorly, each cilium slightly out of phase with its anterior and posterior neighbours (metachronal rhythm) but in phase with its lateral neighbours in adjacent rows (isochronal rhythm).

Movements involve pseudopodia in sarcodines and some stages in life cycles of certain mastigophorans and sporozoans. The two-way sliding type functions in locomotion but the mechanism is somewhat uncertain. Contractile-hydraulic movements are exemplified by ***Amoeba proteus.*** An interior fluid (endoplasm) flows forward under pressure from the exterior gel (ectoplasm); on reaching the tip of a developing pseudopodium, the fluid turns backward against the ectoplasm and changes into gel. Posteriorly, ectoplasm is converted into endoplasm, which again flows forward under pressure.

Nutrition. Phagotrophic nutrition involves ingestion of solids; saprozoic nutrition involves the intake of dissolved materials. Many phagotrophs are not limited to ingestion but utilize saprozoic nutrition as well. Certain phototrophs ***(e.g. Ochromonas)*** combine phagotrophic and saprozoic feeding with photosynthesis. The relative importance of phagotrophy depends upon the available food. Such rumen ciliates as ***Entodinium*** apparently cannot use dissolved carbohydrates, yet they ingest starch grains, digest them, and utilize the products.

(Left) Roman Vishniac. (right) Philip Feinberg. Fellow. New York Microscopical Society



(Left) Amoeba catching a ciliate. (Right) Euplotes digesting algae; movement of cilia has started a whirlpool of algae at right.

The process of ingestion

During ingestion, amoebas may form food cups or sometimes invaginations simulating a cytopharynx. In ***Actinophrys sol,*** a flagellate captured on a sticky pseudopodium is moved toward the body surface, where contact stimulates extension of a food cup. The pseudopodial net of foraminiferans is more effective, but the mechanism is similar. Some flageliates ingest food in amoeboid fashion. A few, such as ***Peranema*** and ***Heteronema,*** have a rod apparatus for puncturing prey, whose contents are then sucked down the cytopharynx. Ciliates that eat small organisms commonly have a buccal cavity. In many, an oral groove leads to this cavity. The buccal cavity contains a membrane and membranelles or bands of cilia, which drive particles to the mouth, or cytostome. Mucus in the oral groove or buccal cavity agglutinates particles on their way to the cytostome. Ciastrioles evaginate from the cytopharynx to accept ingested food. When a suctorian host contacts with a ciliate prey, the host's tentacle fuses with the prey, forming a tube down which the ciliate's protoplasm flows into a gastriole of the host. As digestion gets under way, smaller vacuoles are pinched off the gastriole, possibly distributing soluble materials throughout the organism.

Pinocytosis involves saclike infoldings of the cell surface, followed by pinching off of vacuoles internally. Pinocytosis may be a major method of feeding in saprozoic organisms.

Digestion of cellulose occurs in various free-living herbivores and certain parasites. Digestion of starches and glycogens is common, and many protozoan species can digest lipids.

Phototrophs need only light for energy, but a few are able to grow well in darkness in suitable media. Many phototrophs grow on a nitrate or an ammonium salt medium, with supplies of magnesium, calcium, phosphate, sometimes a chloride, and trace metals; in addition, some need one or more vitamins.

As the only carbon source, CO_2 is adequate for certain phytoflagellates but not for other Protozoa. Acetate is an excellent source of carbon for heterotrophic phytoflagellates and can be used to some extent by *Acanthamoeba* and certain other heterotrophs. Glucose is useless to ***Chilomonas*** and is used only by certain strains of ***Astasia*** and ***Euglena,*** yet various chrysomonads ***(e.g., Ochromonas)*** and certain others ***(e.g., Cryptothecodinium)*** use glucose readily.

Vitamin needs

With respect to vitamin requirements, two protozoan groups are distinguishable: (1) The phytoflagellate type usually needs vitamin B_{12}, thiamine, and biotin; (2) heterotrophs require most or all of the following: thiamine, biotin, pantothenic acid, pyridoxine (or pyridoxamine or pyridoxal), riboflavin, nicotinic acid, thioctic acid foi ciliates, folic acid, hematin for certain trypanosomids, and biopterin for ***Crithidia fasciculata.*** Certain heterotrophs also require lipids: a sterol, varying with the species; a tocopherol derivative; and a long-chain fatty acid. Determination of food requirements has encouraged use of Protozoa as sensitive bioassay organisms in medical and ecological research. The choice of organisms is determined by such practical matters as rate of growth and suitability for spectrophotometric measurement of growth. Vitamin B_{12} has been assayed, even in body fluids, with ***Euglena gracilis*** and with ***Ochromonas malhamensis,*** the former being a very sensitive indicator but slightly less specific than the latter in that it responds to pseudo–B_{12} compounds as well. ***Ochromonas*** *malhamensis* also is a sensitive assay organism for thiamine in body fluids, while thioctic acid can be assayed with ***Tetrahymena pyriformis.***

Oxygen requirements

There is a wide range in oxygen needs among Protozoa. Obligate aerobes are limited to habitats with adequate oxygen concentrations; certain aerobes, however, grow best at oxygen tensions below that of the normal atmosphere. Obligate anaerobes, on the other hand, are restricted to environments lacking oxygen—*e.g.*, natural waters loaded with putrefying material, the lower levels of deep lakes, Imhoff sewage tanks, the intestines of vertebrates, etc.

Typical aerobes utilize the cytochrome system—a terminal oxidative system characteristic of aerobic organisms—in their major oxidative pathway. The genus ***Trypanosoma,*** however, is interesting in that certain species are sensitive to cyanide poisoning, which blocks the cytochrome pathway, while others are rather insensitive.

Excretion and secretion. Excretion involves elimination of wastes through the surface. The major nitrogenous waste is ammonia (accounting for about 90 percent of the total in ***Paramecium caudatum).*** In addition, small amounts of a few amino acids and products of purines and pyrimidines are eliminated by certain ciliates. Waste products of carbon metabolism include CO_2 and, in fermentation of carbohydrates, organic acids and sometimes hydrogen (H_2). Secretions include materials for shells, mucous trails for gliding, cements, the matrix of colonies, and component plates.

EVOLUTION AND PHYLOGENY

The geological age of Protozoa is uncertain but estimates run as high as 1,500,000,000 to 3,000,000,000 years. Fossils dating from the Early Paleozoic, with ages of approximately 450,000,000 years, include skeletons of radiolarians, loricae of tintinnid ciliates, and thecae of dinoflagellates. Some phytoflagellate-like fossils may be several times as old. Each of the Paleozoic groups represents fairly specialized members of its class or order, so that protozoan evolution must have been well along when these organisms were alive.

Origin of Protozoa

Although fossil records supply data for estimating ages of a few protozoan groups, tracing the phylogenetic relationships of modern Protozoa is a different matter. Any phylogenetic tree is extremely hypothetical at best. It is based upon available collateral information, which forms a basis for speculation. A biochemical approach to estimating degrees of relationship among protozoans involves extracting DNA from different species and testing for shared nucleotide sequences, the inference being that the higher the number of shared nucleotide sequences the closer the relationship of the organisms in question. As for the origin of the Protozoa, it is often assumed that some ancient "plant-animal" group (ancestral phytoflagellates) split up into several different stocks, each of which gave rise to the more immediate ancestors of modern Protozoa. This theory of polyphyly, or origin from several ancestral forms, is inferred from the distinct protozoan lines represented in living groups: (1) the algal stocks that gave rise to modem algae and phytomastigophorans; (2) the amoeboid stocks that produced modern sarcodines, apparently representing several distinct lines; (3) the flagellate stocks leading to zoomastigophorans (possibly also to the opalinates); (4) the sporozoan stocks leading to gregarines, coccidians, and toxoplasmas; (5) the stocks that produced the myxosporidians, microsporidians, and haplosporans; and (6) the stocks leading to modern ciliate groups. The phylogenetic origins of Protozoa may be traced thiough an uncertain number of different stocks to a remote hypothetical group of ancestral "plant-animals" presumably representing unspecialized phytoflagellates able to undergo diversification.

CLASSIFICATION

Protozoans are eucaryotic organisms; that is, they have distinct nuclear material. Their disposition into taxa is based primarily on differences in locomotion, number of nuclei, organelles, life cycle, and mode of reproduction. Nutrition, whether plantlike (capable of photosynthesis) or animallike and whether free-living or parasitic, is important in the ranking of lower taxonomic groups.

Many protozoology textbooks will flatly state, "protozoans are unicellular animals." Others consign certain types to the botanists as unicellular plants. Proposals to assign the protozoans a place within the Protista because of structural and evolutionary closeness to slime molds, fungi, and algae have not met with whole-hearted endorsement. Protozoa, however, are not "animals" in the same sense that higher animals are. They are, nevertheless, the most animallike of protists. A collective term for the four major groups of protozoans is desirable as a matter of convenience. On such grounds, the continued use of the term Protozoa seems satisfactory.

The following system, proposed by the Society of Protozoologists in 1964, represents a compromise, proposed hopefully to stimulate further interest in protozoan classification.

PHYLUM PROTOZOA

About 30,000 living species described; fossil types are probably more numerous. Eucaryotic organisms, 1 to many nuclei; organisms single or associated in colonies; majority heterotrophic, but chromatophores present in many phytoflagellates; other organelles vary in complexity; many taxa contain a few to many (or only) parasitic species.

Subphylum Sarcomastigophora

Locomotor organelles are flagella or pseudopodia, sometimes both; nuclei are uniform (except particular stages in life cycles of certain Forarniniferida); reproduction by fission, budding, or plasmotomy.

Superclass Mastigophora (flagellates)

About 6,000 species (not including fossils) Solitary or colonial; with flagella, except in certain-stages of life cycle; mostly heterotrophic, but many phytoflagellates have chromatophores.

Class Phytamastigophorea (phytoflagellates)

Many with chromatovhores: food requirements simvler than those of other flagellates; rarely more than 1 or 2 flagella; mostly free-living; syngamy only in certain groups.

Order Chrysomonadida. Marine and freshwater; 1 or 2 active flagella, some with a third inactive flagellum; chromatophores, 1 or 2, yellow to brown; stored leucosin and lipids; cyst wall siliceous; life cycles may include flagellate, amoeboid, or palmella stages, any one of which may be dominant. (Ochromonas, *Dinobryon*, *Mallomonas*, Prymnesium.)

Order *Silicoflagellida.* Marine; uniflagellate; numerous greenish-brown chromatophores; leucosin stored; siliceous skeleton characteristic; many fossils known. (*Dictyocha.*)

Order Coccolithophorida (coccoliths). Mostly marine or brackish-water, very few freshwater types; biflagellate; chromatophores yellow to brown; peripheral zone of calcareous coccoliths. (Discosphaera, Hymenomonas, Rhabdosphaera.)

Order Heterochlorida. Flexible periplast; typically 2 unequal flagella; 1 to several yellow to greenish-yellow chromatophores; leucosin and lipids stored; cyst bivalved; some have amoeboid, plasmodial, or palmella stages. (*Chloromeson*, Nephrochloris, Myxochloris.)

Order Cryptomonadida. Marine and freshwater; 2 flagella, emerging from a ventral groove or pouch lined with refractile inclusions ("trichocysts"); usually 2 chromatophores—brown, blue-green, green, or red; starch and lipids stored. (Chilomonas, Cryptomonas, Nephroselmis.)

Order *Dinoflagellida* (dinoflagellates). Mostly marine, some freshwater; one flagellum typically trailing, the other lying in transversely spiral girdle; theca common, often with differentiated plates; chromatophores (if present) numerous, yellow-green to brown, sometimes blue-green; some are phagotrophs. (Ceratium, Gymnodinium, *Gonyaulax*, Peridinium, Noctiluca.)

Order Ebriida. Marine, planktonic; biflagellate; no chromatophores; typically with siliceous skeleton, internal; fossils fairly numerous. (Ebria, Hermesinum.)

Order *Euglenida.* Mostly freshwater, some marine; 1 or 2 flagella, rarely more, emerging from a pouch (reservoir); chromatophores (if present) fairly numerous, green; some species accumulate red pigment; species with thin pellicle complex undergo metaboly (euglenoid movement). (Euglena, *As*tasia, Peranema, Trachelomonas.)

Order Chloromonadida. Freshwater body usually flattened dorsoventrally; 2 flagella (one trailing) arising from anterior pouch or groove; numerous bright green chromatophores; lipids, glycogen stored. (*Gonyostomum*, *Vacuolaria.*)

Order Volvocida. Mostly freshwater; chromatophores green, also some species colourless; some accumulate red pigment; flagella apical, 2 or 4; solitary or colonial; starch and lipids stored; syngamy common, sometimes marked anisogamy (*Volvox*, etc.). (Chlamydomonas, Haematococcus, *Po*lytoma, Gonium, Pandorina.)

Class Zoomastigophorea (zooflagellates)

Without chromatoyhores; organelles other than flagella more varied than in Phytomastigophorea; syngamy and autogamy reported in a few.

Order *Choanoflagellida.* Mostly freshwater; anterior collar surrounds base of single flagellum; solitary or colonial forms, typically sessile, with or without stalk; lorica in some species. (Codonosiga, *Diplosiga*, Salpingoeca.)

Order Bicosoecida. Freshwater; loricate, colonial; 1 flagellum extends from lorica, the other anchors the flagellate to the base of its lorica. (Bicosoeca.)

Order Rhizomastigida. Mostly freshwater, some parasitic; more or less amoeboid, usually with 1 to 4 flagella; Histomonas causes enterohepatitis in chickens and turkeys. (Histomonas, Mastigella, Mastigamoeba.)

Order Kinetoplastida. Mostly parasitic; kinetoplast characteristic; 1 to 4 flagella; life cycles may involve more than 1 host. (Bodo, Trypanosoma, Leishmania, Crirhidia.)

Order Retorfomonadida. Parasitic; 2 to 4 flagella, one extending posteriorly in a cytostomal groove bordered by fibrils. (Chilomastix,Retortomonas.)

Order Diplomonadida. Mostly parasitic; binucleate, bilaterally symmetrical; organized as 2 karyomastigonts, each with four flagella and associated organelles. (Giardia, *Hexami*ta, Trepomonas.)

Order *Oxymonadida.* Parasitic; uninucleate or multinucleate; 1 or more karyomastigonts, each usually with 4 (sometimes 8 or 12) flagella; some attach to gut wall of termites and wood roaches by extensible rostellum; syngamy in certain species. (Dinenympha, Pyrsonynzpha, uninucleate; *Microrhopalodina,* Barroella, multinucleate.)

Order Trichomonadida. Parasitic; each mastigont contains 3 to **6** flagella, a parabasal body, and an axostyle; uninucleate or multinucleate. (Devescovina, *Calonympha, Trichomonas,* Snyderella.)

Order Hypermastigida. Parasitic; uninucleate, multiflagellate; parabasal apparatus multiple; blepharoplasts distributed in rows or plates, flagella in rows or bundles; some ingest wood chips; gut of termites, wood roaches, cockroaches. (Trichonympha, Holomastigotoides, Mixotricha.)

Superclass *Opalinata*

About 200 species. Parasites; resemble both Mastigophora and Ciliophora; cilia-like organelles forming oblique rows over body; 2 to many monomorphic nuclei; plane of fission oblique, not transverse as in typical ciliates; syngamy reported. (Opalina, Protoopalina.)

Order Opalinida. The only order under this superclass.

Superclass Sarcodina

About 11,500 living species (many fossils, about 28,000 Foraminifera alone); mostly free-living; some have flagella in certain stages of the life cycle. Syngamy in certain groups; pellicle usually thin and flexible; tests in certain group;; endoskeletal elements in some.

Class Rhizopodea

Temporary flowing appendages called pseudopodia; no foamy peripheral cytoplasm; tests are common.

Subclass Lobosia. Fingerlike pseudopodia (lobopodia).

Order Amoebida (amoebas). Marine and freshwater; mostly uninucleate; mostly free-living. (Amoeba, Thecamoeba, Entamoeba.)

Order Arcellinida. Freshwater; tests characteristic, secreted or arenaceous; pseudopodia extruded through aperture of test. (Arcella, Centropyxis, *Difflugia.*)

Subclass Filosia. Pseudopodia slender, branching, rarely anastomosing.

Order Aconchulinida. Without tests; little known.

Order Gromiida. Mostly freshwater; tests mostly chitinous, aperture distinct; flagellate stages in certain species; anastomosis of pseudopodia not marked. (Gromia.)

Subclass Granuloreticulosia. Pseudopodia delicate, showing moving granules.

Order *Athalamida.* No tests; little-known group.

Order Foraminiferida (forams). Mostly marine and brackish-water; about 4,200 living species, approximately 28,000 fossil types; tests with 1 to many chambers, secreted (calcareous predominantly, siliceous rarely) or arenaceous; during growth, multilocular types add new chambers in characteristic patterns; reticulopodia extend from aperture or partly from cytoplasm covering test; extensive reticulopodial network characteristic. (Allogromia, *Patellina,* Saccamina, *Cyclammina,* Bathysiphon.)

Order Xenophyophorida. Deep-sea; discoid or fan-shaped, reaching several centimetres across; consist of fine tubes in agglutinated foreign particles; finer tubules contain multinucleate protoplasm; iiving forms not yet investigated: relationships uncertain. (Psammetta, Stanomma.)

Subclass Mycetozoia. Young amoebas develop into aggregates or into large plasmodia; resemble fungi in production of "sporangia" enclosing "spores" (cysts), from which hatch amoeboid stages.

Order Acrasida. Soil, decaying plants; aggregates are pseudoplasmodia; flagellate stages unknown; free-living. (Acrasis, Dictyostelium.)

Order Eumycetozoida (slime molds). Rotting leaves, logs; plasmodium often shows many channels containing flowing endoplasm; many produce sporangia at maturity. (Fhysarum, Fuligo, Arcyria.) (See SLIME MOLDS.)

Order Plasmodiophorida. Parasitic; mature stage a plasmodium inside a plant cell; cysts ("spores") develop inside the plasmodium. (Plasmodiophora, Sorosphaera, Textramyxa.)

Subclass Labyrinthulia. Mostly marine, on eelgrass and certain algae; spindle-shaped organisms that secrete tubular "tracks" by means of which the organisms glide; swimming stages (1 anterior, 1 trailing flagellum) described as similar to fungal zoospores.

Order *Labyrinthulida.* The only order. (Labyrinthula.)

Class Piroplasmea

Parasitic; transferred by ticks; attacks vertebrate red blood cells; reproduction by fission; taxonomic position debatable.

Order Piroplasmida. The only order in the class. (Babesia.)

Class *Actinopodea*

Mostly freshwater; pseudopodia slender, radially arranged; true axopodia in some genera; test present or absent; may have skeletal elements of silica or strontium sulfate; a central capsule in many; syngamy (pedogamy) in certain genera.

Subclass Radiolaria. Marine; about 4,800 living, 2,400 fossil species; central capsule dividing cytoplasm into **2** zones; a siliceous skeleton, or separate spicules; pseudopodia may be filopodia, reticulopodia or, rarely, axopodia.

Order Porulosida. Pores distributed over surface of spherical central capsule; a few have no skeleton; in others, spicules, a perforated test, or both may be present; colonies formed in some genera. (*Collozoum,* Pipetta, Cenosphaera.)

Order Oculosida. Pores limited to a few fields in the capsule. (*Lithocircus,* Aulacnntha, *Challengeron.*)

Subclass Acantharia. Mostly marine; skeletons composed mainly of strontium sulfate; primary skeletal elements are rods extending radially from centre, passing through the central capsule; a latticework test at the surface may be fused with the radial rods (20, or some multiple of **20**), which form a characteristic pattern; outer layer of cytoplasm often joined to rods by contractile fibrils; axopodia present.

Order Acanthometrida. Skeleton composed of rods, no latticework test. (Acanthometra, Zygacanthidiurn.)

Order Acnnthophractida. Latticework test present. (*Dorotaspis,* Lithoptera.)

Subclass Heliozoia. Mostly freshwater; skeletal elements (if present) often limited to scales and spines, but test in some; no central capsule; axopodia or filopodia.

Order Actinophryida. Scales and spines lacking; outer cytoplasm often highly vacuolated; uninucleate or multinucleate; axopodia present. (Actinophrys, Actinosphaerium.)

Order Centrohelida. Central granule (centroplast) present; axonemes of axopodia often converge in centroplast; skeletal plates and spines often present. (Acanthocysiis, *Raphidocystis.*)

Order Desmothoracida. No centroplast; skeleton a nonsiliceous test, with pores through which slender granular pseudopodia extend. (Choanocystis, Hedriocystis, Clathrulina.)

Subclass Proteomyxidia. Parasitic; slender granular pseudopodia or filopodia present; no skeletal elements; many are invaders of Volvocida, filamentous algae, and cells of higher plants.

Order Proteomyxida. The only order. (Pseudospora, Vampyrella, Leptomyxa.)

Subphylum Sporozoa

Abont 3,600 species known. Parasites of invertebrates and vertebrates, or both in certain species; life cycle with asexual and sexual phases, reproduction in each (except in most gregarines); majority produce spores; syngamy in most, if not all, sometimes with flagellated gametes; locomotion by gliding, sometimes by pseudopodia; reproduction by schizogony typical.

Class Telosporea

Syngamy, followed by sporulation, just before or after transfer to a new host; pseudopodia obvious only in growth stages of certain groups; sporozoites (often encysted, forming a spore) transferred to a new host, or from vector to vertebrate.

Subclass Gregarinia (gregarines). Parasites of invertebrates, mostly in body cavity and digestive tract; gamonts (gametocytes) often associate in pairs before gametogenesis; mature organisms large, extracellular.

Order Archigregarinida. Parasites of annelids, sipunculids, hemichordates, urochordates; life cycle shows intracellular schizogony; only gamonts are extracellular. (Selenidium, *Schizocystis.*)

Order Eugregarinida. Parasites of annelids, insects, crustacea, echinoderms, sipunculids, urochordates; no merogony in asexual phase of cycle. (Gregarina, Monocystis, Porospora.)

Order Neogregarinida. Mostly parasites of insects; schizogony in asexual phase; each gametocyte produces 1 spore. (Ophryocystis, Merogregarina.)

Subclass Coccidia. Parasites of invertebrates and vertebrates; trophozoites mature intracellularly, much smaller than those of gregarines; reproduction in asexual and sexual phases of cycle.

Order Protococcida. Elongated nematode-like organisms; no intracellular schizogony; rare. (Selenococcidium, first reported from marine crab.)

Order Eucoccida. Parasites mostly in epithelial or blood cells; intracellular merogony characteristic. (Eimeria, Adelea, *Isospora,* Plasmodium.)

Class Toxoplasmea

lntracellular parasites, common in man and other vertebrates; reproduction by fission or a peculiar internal budding; locomotion (without flagella or pseudopodia) involves a gliding similar to that of Plasmodium sporozoites; life cycle completed in 1 host; a number of parasites accumulate in host cell, remnants of which form a pseudocyst.

Order Toxoplasmida. The only order in the class. (*Toxoplasma*, Besnoita, similar to Toxoplasma in morphology and antigenic composition, probably belongs here.)

Subclass *Haplosporea.* Parasites of fishes and tunicates, and of mollusks, annelids, and other invertebrates; mature stage a plasmodium; flagella absent; infective amoeba stage may invade tissue cell or develop in body cavity; syngamy described in Coelospora, Minchinia, unknown in others. (*Coelosporidium*, Coelospora, Haplosporidiunz, Minchinia.)

Subphylum Cnidospora

About 1,100 species known. Parasites producing spores (cysts) with 1 or more polar filaments and 1 or more infective stages (sporoplasms); young trophozoite typically develops into a plasmodium; structure of spore differentiates classes.

Class Myxosporidea

Spore membrane composed of 2 or 3 valves; spore of "multicellular" origin, containing 1 or more sporoplasms.

Order Myxosporida. Parasites of lower vertebrates; spore membrane composed typically of 2 valves; spore contains 1 or 2 sporoplasms and usually 2 (sometime 1 to 6) polar filaments. (*Unicapsula,* Myxidium, Ceratomyxa, Myxobolus.)

Order Actinomyxida. Parasites of invertebrates (mostly annelids); spore membrane composed of 3 valves, each sometimes drawn out into a spine; 3 polar filaments; 1 to many sporoplasms. (Triactinomyxon, Sphaeractinomyxon, Guyenotia.)

Order Helicosporida. Parasites of arthropod larvae; mature spore contains 3 sporoplasms surrounded by spirally wound filaments; spore membrane not differentiated into valves. (Helicosporidium.)

Class Microsporidea

Parasites mostly in arthropods and fishes (usually skin, muscles); spores small (about 2-20 μm long), membrane not differentiated into valves; one polar filament.

Order Microsporida. The only order in this class. (*Coccomyxa*, *Nosema,* Telomyxa.)

Subphylum Ciliophora (Ciliates)

About 6,000 living species. Marine and freshwater; cilia or derivatives in at least 1 stage of the life cycle; 2 kinds of nuclei characteristic; conjugation (also autogamy in some) rather than syngamy; plane of fission typically transverse to major axis.

Class *Ciliatea*

Characteristics of the subphylum.

Subclass Holotrichia. Most free-living, few parasitic; less specialized than other Ciliatea; somatic ciliature often uniform.

Order Gymnostomatida. Marine, freshwater, some predatory; no buccal organelles; cytostome opens at surface; circumpharyngeal trichites often present; ciliature often uniform, but cilia limited to ventral surface in 1 group of flattened types; some feed on algae, protozoans, rotifers. (Coleps, *Dileptus*, Didinium, Chilodonella.)

Order Trichostomatida. Fresh-water, some commensal or parasitic; no buccal organelles, but vestibular ciliature may be specialized; cytostome often at base of groove or pit; somatic ciliature usually uniform; found in the stomach of certain ruminants and in the intestine of many vertebrates, including man. (Colpoda, Conidiophrys, Balantidium, Trichospira.)

Order Chonotrichida. Marine (except one genus), commensals mostly on crustacea; somatic ciliature absent in mature forms; vase-shaped body typical; peristomial funnel equipped with cilia derived from those of larval stage. (*Stylochona*, Trichochona, Chilodochona.)

Order Apostomatida. Parasites mostly of marine crustacea; ciliature spiral at maturity; life cycles typically polymorphic, minute cytostome near peculiar "rosette" of uncertain function. (Chromidina, Foettingeria, Polyspira, *Gymnodinioides*.)

Order Astomatida. Parasitic, mostly in earthworms, a few in snails, flatworms, and amphibians; no cytostome; somatic ciliature usually uniform; anterior holdfast organelle common; chains common in some families. (Anoplophrya, Corlissiella, Maupasella, Haptophrya, Hoplitophrya.)

Order Hymenostomatida. Marine, freshwater, some parasitic; buccal organelles (tetrahymenal type) include membrane (on right) and typically three membranelles (on left); somatic ciliature essentially uniform; buccal cavity ventral. (Tetrahymena, Colpidium, Glaucoma, Paramecium.)

Order Thigmotrichida. Generally parasitic in bivalve mollusks;-tuft of thigmotactic cilia commonly present near anterior end; bucca organelles (if present) are ventral and in posterior half (or at posterior end). (*Boveria*, *Conchophthirius*, Hypocoma, Heterocinefa.)

Subclass Peritrichia. Marine, freshwater, some parasitic; somatic ciliature absent at maturity, but apical ciliature winds counterclockwise to cytostome; mostly sessile at maturity, attached by a stalk (product of a scopula) or adhesive basal disk; larval stages migratory, with ciliary girdle; many are colonial.

Order *Peritrichida.* The only order in this subclass. (*Vorticella, Carchesium,* Trichodina, *Epistylis.*)

Subclass Suctoria. Marine, freshwater, free-living predators; ciliature (but not kinetosomes) absent in mature stage; mostly sessile, attached by stalk (scapula-produced); astomatous larva, produced by budding, develops sparse ciliature from inherited kinetosomes; adults feed mostly on other ciliates by means of suctorial tentacles.

Order *Suctorida.* The only order in this sublcass. (*Tokophrya*, Podophrya, Ephelota, *Dendrosoma.*)

Subclass Spirotrichia. Marine, freshwater, some commensal or parasitic; somatic ciliature sparse (except in Heterotrichida); conspicuous buccal organelles include many membranelles, in zone winding clockwise to cytostome.

Order *Heterotrichida.* Marine, freshwater, some parasitic; ciliature commonly uniform, but restricted to ventral surface in 1 family; membrane often present at right of peristome; a few sessile types loricate, with a migratory larva. (*Bursaria, Stentor, Blepharisma, Spirostomum, Folliculina.*)

Order Oligotrichida. Mostly marine; somatic ciliature reduced or lacking (persisting somatic ciliature often arranged in tufts); conspicuous buccal membranelles often form apical zone around end of body. (Strombidium, Halteria, *Strobilidium*.)

Order Tintinnida. Nearly all marine; all loricate; buccal membranelles conspicuous when extended from lorica. (*Tintinnus,* Codonella, Favella, Tinrinnidium.)

Order Entodiniomorphida. Commensals; ciliature much reduced, limited to organelles arranged in tufts or bands; adoral membranelles extend spirally to cytostome lying in apical disk; the firm pellicular complex may be drawn out into posterior spines; found in the rumen of ruminants and in the intestine of certain other herbivores. (Diplodinium, *Entodinium*, *Ophryoscolex*, Cycloposthium.)

Order *Odontostomatida.* Nearly all freshwater; small laterally compressed, wedge-shaped; somatic ciliature much reduced; only eight buccal membranelles; firm pellicular complex, often drawn out into spines. (Saprodinium, Mylestoma, Epalxis, Atopodinium.)

Order Hypotrichida. Marine, freshwater, few parasitic; cilia replaced by cirri, arranged in specific patterns on ventral surface (also near margin of body in some genera); adoral zone of membranelles conspicuous; typically flattened dorsoventrally, often with firm pellicle. (*Euplotes,* Stylonychia, Uronychia.)

Critical Appraisal. Protozoa as a taxon meets the criterion of convenience in current classifications. Troubles begin primarily when the essential artificiality of the Phylum Protozoa is disregarded. The system outlined above recognizes four major groups of Protozoa, each ranked as a subphylum. Whatever their ranks, the groups differ in degree of homogeneity. The Sarcomastigophora, the least homogeneous, include two groups of Mastigophora that differ considerably in morphological and biochemical features, although sharing the ability to produce flagella. The Phytomastigophorea, essentially, are a heterogeneous collection of algae that protozoologists choose to call Protozoa. The Zoomastigophora are clearly animallike in many characteristics. A subphylum containing both groups must be considered polyphyletic. Furthermore, the Sarcomastigophora also include the Sarcodina, and there are sound reasons for concluding that this subdivision likewise contains two major groups (differing in their mechanisms of locomotion). The other subphyla—Ciliophora, Sporozoa, and Cnidospora — are more nearly homogeneous than the Sarcomastigophora. The degree of relationship between any two of the four is quite uncertain at present; and the only safe conclusion, if a common phytoflagellate ancestry is assumed, is that the four groups have evolved independently for a long time. A collective term for the four major groups is desirable as a matter of convenience. On such grounds, continued use of the taxon Protozoa seems satisfactory.

One recent proposal would assign most of the Protozoa to a Kingdom Protista, elevating a number of presently subordinate taxa to the rank of phylum. This would result in taxonomic equivalence of such groups as the Order Euglenida (Euglenophyta) and the Subphylum Ciliophora (presently including 15 orders). The average protozoologist will find little philosophical or utilitarian advantage in such a system, which obviously would require extensive revision of our current concepts of taxa.

So far as the Protozoa are concerned, the interests of taxonomy would probably be served best by continued attempts to collect pertinent data and to devise less subjective methods for estimating degrees of relationship. There seems to be little fundamental benefit to be derived from a mere reshuffling of taxonomic names.

BIBLIOGRAPHY. T.T. CHEN (ed.), ***Research in Protozoology,*** 4 vol. (1967–), reviews on contemporary investigations of protozoan structure and function; J.O. CORLISS, ***The Ciliated Protozoa*** (1961), a modified classification and a guide to the literature; P.P. GRASSE (ed.), ***Traité de Zoologie,*** vol. 1 (1952), a classic work in French that covers morphology, physiology, biochemistry, biophysics, and taxonomy; R.P. HALL, ***Protozoology*** (1953), an introductory college textbook on protozoa stressing principles and surveying the group, and ***Protozoa*** (1964), an introduction, including life cycles and processes, taxonomic problems, and an abridged classification, and ***Protozoan Nutrition*** (1965), *a* detailed treatment of nutritional requirements of protozoans; S.H. HUTNER and A. LWOFF (eds.), ***Biochemistry and Physiology of Protozoa,*** 3 vol. (1951–64), contributions dealing with nutrition, metabolism, and growth of the protozoa, and the use of protozoa as biochemical tools; L.H. HYMAN, ***The Invertebrates: Protozoa Through Ctenophora,*** vol. *1* (1940), an important advanced source book; T.L. and F.F. JAHN, ***How to Know the Protozoa*** (1949), an elementary book, with accurate descriptions and an identification key with pictures; R.R. KUDO, ***Protozoology,*** 5th ed. (1966), a comprehensive reference book containing detailed information on common and representative Protozoa; D.L. MACKINNON and R.S.J. HAWES, ***An Introduction to the Study of Protozoa*** (1961), a textbook with detailed descriptions of representative species of important Protozoan groups, and excellent figures and notes on laboratory techniques; R.D. MANWELL, ***An Introducton to Protozoology*** (1968), a readable, well written text; D.P. PITELKA, ***Electron-Microscopic Structure of Protozoa*** (1963), an advanced study of the fine structure of protozoa; S.H. HUNTER ***et al.*** in R.N. FIENNES (ed.), ***Nutrition of Lower Organisms*** (1971), a chapter of which is devoted to the nutrition and metabolism of protozoans.

(R.P.H.)

Proudhon, Pierre-Joseph

Pierre-Joseph Proudhon, libertarian Socialist, is celebrated as the first man to call himself an Anarchist. Mikhail Bakunin, the Russian aristocrat who gained more notoriety in that role, once remarked that "Proudhon was the master of us all," and in the sense that Anarchism as a movement derived from Proudhon's teachings and his immediate disciples, the statement was correct. Yet Proudhon was not the first to enunciate the doctrine that is now called Anarchism; before he claimed it, it had already been sketched out by, among others, the English philosopher William Godwin in prose and by his follower, Percy Bysshe Shelley, in verse.

There is no evidence, however, that Proudhon, who was born in Besançon, France, January 15, 1809, ever studied the works of either Godwin or Shelley; and his characteristic doctrines of Anarchism (society without government), Mutualism (workers' association for the purpose of credit banking), and federalism (the denial of centralized political organization) seem to have resulted from an original reinterpretation of French revolutionary thought modified by personal experience.

Early life and education

Proudhon was born into poverty as the son of a feckless cooper and tavern keeper, and at nine he worked as a cowherd in the Jura Mountains. His country childhood and peasant ancestry influenced his ideas to the end of his life; and his vision of the ideal society almost to the end remained that of a world in which peasant farmers and small craftsmen like his father could live in freedom, peace, and dignified poverty, for luxury repelled him and he never sought it for himself or others.

Proudhon at an early age showed the signs of intellectual brilliance, and he won a scholarship to the college at Besançon. Despite the humiliation of being a child in sabots (wooden shoes) among the sons of merchants, he developed a taste for learning and retained it even when his family's financial disasters forced him to become an apprentice printer and later a compositor. While he learned his craft, he taught himself Latin, Greek, and Hebrew; and in the printing shop he not only conversed with various local liberals and Socialists but also met and fell under the influence of a fellow citizen of Besançon, the utopian Socialist Charles Fourier.

Proudhon and his children, oil painting by Gustave Courbet, c. 1865. In the Musée du Petit Palais, Paris.
Giraudon

With other young printers, Proudhon later attempted to establish his own press, but bad management destroyed the venture; and it may well have been compounded by his own growing interest in writing, which led him to develop a French prose difficult to translate but admired by writers as varied as Flaubert, Sainte-Beuve, and Baudelaire. Eventually, in 1838, a scholarship awarded by the Besançon Academy enabled him to study in Paris. Now, with leisure to formulate his ideas, he wrote his first significant book, ***What Is Property?*** (1840). This created a sensation, for Proudhon not only declared, "I am an anarchist"; he also stated, "Property is theft!"

His first book, ***What Is Property?***

This slogan, which gained much notoriety, was an example of Proudhon's inclination to attract attention and mask the true nature of his thought by inventing striking phrases. He did not attack property in the generally accepted sense but only the kind of property by which one man exploits the labour of another. Property in another sense—in the right of the farmer to ***possess*** the land he works and the craftsman his workshop and tools—he regarded as essential for the preservation of liberty; and his principal criticism of Communism, whether of the utopian or the Marxist variety, was that it destroyed freedom by taking away from the individual control over his means of production.

In the somewhat reactionary atmosphere of the July monarchy in the 1840s, Proudhon narrowly missed prosecution for his statements in ***What Is Property?;*** and he was brought into court when, in 1842, he published a more inflammatory sequel, ***Warning*** to ***Proprietors.*** On this first of his trials, Proudhon escaped conviction because the jury conscientiously found that they could not clearly understand his arguments and therefore could not condemn them.

In 1843 he went to Lyons to work as managing clerk in a water transport firm. There he encountered a weavers' secret society, the Mutualists, who had evolved a protoanarchist doctrine that taught that the factories of the dawning industrial age could be operated by associations of workers and that these workers, by economic action rather than by violent revolution, could transform society. Such views were at variance with the Jacobin revolutionary tradition in France, with its stress on political centralism. Nevertheless, Proudhon accepted their views and later paid tribute to his Lyonnais working-class mentors by adopting the name of Mutualism for his own form of Anarchism.

As well as encountering the obscure working-class theoreticians of Lyons, Proudhon also met the feminist Socialist Flora Tristan and, on his visits to Paris, made the acquaintance of Karl Marx, Mikhail Bakunin, and the Russian Socialist and writer Aleksandr Herzen. In 1846 he took issue with Marx over the organization of the Socialist movement, objecting to Marx's authoritarian and centralist ideas. Shortly afterward, when Proudhon published his *System of Economic Contradictions: or the Philosophy of Poverty* (1846), Marx attacked him bitterly in a book-length polemic, *The Poverty of Philosophy* (1847). It was the beginning of a historic rift between libertarian and authoritarian Socialists and between Anarchists and Marxists which. after Proudhon's death, was to rend Socialism's First International apart in the feud between Marx and Proudhon's disciple Bakunin and which has lasted to this day.

Rift with Karl Marx

Early in 1848 Proudhon abandoned his post in Lyons and went to Paris where in February he started the paper *Le Représentant du peuple.* During the revolutionary year of 1848 and the first months of 1849 he edited a total of four papers; the earliest were more or less regular Anarchist periodicals and all of them were destroyed in turn by government censorship. Proudhon himself took a minor part in the Revolution of 1848, which he regarded as devoid of any sound theoretical basis. Though he was elected to the Constituent Assembly of the Second Republic in June 1848, he confined himself mainly to criticizing the authoritarian tendencies that were emerging in the revolution and that led up to the dictatorship of Napoleon III. Proudhon also attempted unsuccessfully to establish a People's Bank based on mutual credit and labour checks, which paid the worker according to the time expended on his product. He was eventually imprisoned in 1849 for criticizing Louis-Napoleon, who had become president of the republic prior to declaring himself Emperor Napoleon III, and Proudhon was not released until 1852.

His conditions of imprisonment were——by 20th-century standards—light. His friend could visit him, and he was allowed to go out occasionally in Paris. He married and begat his first child while he was imprisoned. From his cell he also edited the last issues of his last paper (with the financial assistance of Herzen) and wrote two of his most important books, the never translated *Confessions d'un révolutionnaire* (1849) and *The General Idea of the Revolution in the Nineteenth Century* (1851). The latter —in its portrait of a federal world society with frontiers abolished, national states eliminated, authority decentralized among communes or locality associations, and with free contracts replacing laws—presents perhaps more completely than any other of Proudhon's works the vision of his ideal society.

After Proudhon's release from prison in 1852 he was constantly harassed by the imperial police; he found it impossible to publish his writings and supported himself by preparing anonymous guides for investors and other similar hack works. When, in 1858, he persuaded a publisher to bring out his three-volume masterpiece *De la justice dans la Rkvolution et darts l'église,* in which he opposed a humanist theory of justice to the church's transcendental assumptions, his book was seized. Having fled to Belgium, he was sentenced *in absentia* to further imprisonment. He remained in exile until 1862, developing his criticisms of nationalism and his ideas of world federation (embodied in *Du Principe fkdkratif,* 1863).

On his return from Paris, Proudhon began to gain influence among the workers; Paris craftsmen who had adopted his Mutualist ideas were among the founders of The First International just before his death on January 19, 1865. His last work, completed on his death bed, *De la capacité politique des classes ouvrières* (1865), developed the theory that the liberation of the workers must be their own task, through economic action.

Assessment

Proudhon was a solitary thinker who refused to admit that he had created a system and abhorred the idea of founding a party. There was thus something ironical about the breadth of influence that his ideas later developed. They were important in the First International and later became the basis of Anarchist theory as developed by Bakunin and the Anarchist writer Peter Kropotkin. His concepts were influential among such varied groups as the Russian populists, the radical Italian nationalists of the 1860s, the Spanish federalists of the 1870s, and the syndicalist movement that developed in France and later became powerful in Italy and Spain. Until the beginning of the 1920s, Proudhon remained the most important single influence on French working-class radicalism, while in a more diffuse manner his ideas of decentralization and his criticisms of government had revived in the later 20th century, even though at times their origin was not recognized.

BIBLIOGRAPHY. There are four books in English on Proudhon that together give a comprehensive view of his life and theories: D.W. BROGAN, *Proudhon* (1934); HENRI DE LUBAC, *Proudhon et le christianisme* (1945; Eng. trans., *The Un-Marxian Socialist: A Study of Proudhon,* 1948); GEORGE WOODCOCK, *Pierre-Joseph Proudhon* (1956); and ALAN RITTER, *The Political Thought of Pierre-Joseph Proudhon* (1969). Relatively little of Proudhon's own writing has been translated into English; in addition to *What Is Property?* (including *Warning to Proprietors), The General Idea of the Revolution in the Nineteenth Century,* and *System of Economic Contradictions,* there are only two volumes of selections: *Proudhon's Solution of the Social Problem* (1927) and *Selected Writings of Pierre-Joseph Proudhon* (1969).

(G.W.)

Proust, Marcel

The entire climate of the 20th-century novel has been affected by Marcel Proust, whose novel *À la recherche du temps perdu (Remembrance of Things Past)* is one of the supreme achievements of world literature. "My instrument," Proust said, "is not a microscope but a telescope directed upon Time." Taking as raw material the author's past life, *À la recherche* is ostensibly about the irrecoverability of time lost, about the forfeiture of innocence through experience, the emptiness of love and friendship, the vanity of human endeavour, the triumph of sin and despair; but Proust's conclusion is that the life of every day is supremely important, full of moral joy and beauty, which, though man may lose them through faults inherent in human nature, are indestructible and recoverable. Proust's style is one of the most original in all literature, unique and self-engendered in its union of speed and protraction, precision and iridescence, force and enchantment, classicism and symbolism. His reputation as a writer of genius continues to rise.

Permission S.P.A.D.E.M. 1971 by French Reproduction Rights, Inc.; photograph J.E. Bulloz

Proust, oil painting by Jacques-Emlie Blanche (1861–1942). In a private collection.

Life and works. Marcel Proust, born on July 10, 1871, at Auteuil, then a still rural suburb of Paris, was the son of Adrien Proust, an eminent physician of provincial French Catholic descent, and his wife Jeanne, *née* Weil,

Influence of Bergson

of a wealthy Jewish family. After a first attack in 1880, he suffered from asthma throughout his life. His childhood holidays were spent at Illiers and Auteuil (which together became the Combray of his novel) or at seaside resorts in Normandy with his maternal grandmother. At the Lycée Condorcet (1882–89) he wrote for class magazines, fell in love with a little girl named Marie de Benardaky in the Champs-Élysées, made friends whose mothers were society hostesses, and was influenced by his philosophy master Alphonse Darlu. He enjoyed the discipline and comradeship of military service at Orléans (1889–90) and studied at the Ecole des Sciences Politiques, taking *licences* in law (1893) and in literature (1895). During these student days his thought was influenced by the philosopheis Henri Bergson (his cousin by marriage) and Paul Desjardins and by the historian Albert Sorel. Meanwhile, via the bourgeois salons of Madames Straus, Arman de Caillavet, Aubernon, and Madeleine Lemaire, he became an observant *habitué* of the most exclusive drawing rooms of the nobility. In 1896 he published *Les Plaisirs et les jours,* a collection of short stories at once precious and profound, most of which had appeared during 1892–93 in the magazines *Le Banquet* and *La Revue Blanche.* From 1895 to 1899 he wrote *Jean Santeuil,* an autobiographical novel that, though unfinished and ill-constructed, showed awakening genius and foreshadowed A *la recherche.* A gradual disengagement from social life coincided with growing ill health and with his active involvement in the Dreyfus affair of 1897–99, when French politics and society were split by the movement to liberate the Jewish army officer Alfred Dreyfus, unjustly imprisoned on Devil's Island as a spy. Proust helped to organize petitions and assisted Dreyfus's lawyer Labori, courageously defying the risk of social ostracism. (Although Proust was not, in fact, ostracized, the experience helped to crystallize his disillusionment with aristocratic society, which became visible in his novel.) Proust's discovery of John Ruskin's art criticism in 1899 caused him to abandon *Jean Santeuil* and to seek a new revelation in the beauty of nature and in Gothic architecture, considered as symbols of man confronted with eternity: "Suddenly," he wrote, "the universe regained in my eyes an immeasurable value." On this quest he visited Venice (with his mother in May 1900) and the churches of France and translated Ruskin's *Bible of Amiens* and *Sesame and Lilies,* with prefaces in which the note of his mature prose is first heard.

Involuntary revival of a childhood memory

The death of his father in 1903 and of his mother in 1905 left him grief stricken and alone but financially independent and free to attempt his great novel. At least one early version was written in 1905–06. Another, begun in 1907, was laid aside in October 1908. This had itself been interrupted by a series of brilliant parodies—of Balzac, Flaubert, Renan, Saint-Simon, and others of Proust's favourite French authors—called "L'Affaire Lemoine" (published in *Le Figaro*), through which he endeavoured to purge his style of extraneous influences. Then, realizing the need to establish the philosophical basis that his novel had hitherto lacked, he wrote *Contre Sainte-Beuve,* attacking the French critic's view of literature as a pastime of the cultivated intelligence and putting forward his own, in which the artist's task is to release from the buried world of unconscious memory the ever-living reality to which habit makes us blind. In January 1909 occurred the real-life incident of an involuntary revival of a childhood memory through the taste of tea and a rusk biscuit (which in his novel became *madeline* cake); in May the characters of his novel invaded his essay; and, in July of this crucial year, he began À *la recherche du temps perdu.* He thought of marrying "a very young and delightful girl" whom he met at Cabourg, a seaside resort in Normandy that became the Balbec of his novel, where he spent summer holidays from 1907 to 1914; but, instead, he retired from the world to write his novel, finishing the first draft in September 1912. The first volume, *Du côté* de *chez Swann,* was refused by the best-selling publishers Fasquelle and Ollendorff and even by the intellectual *La Nouvelle Revue Française,* under the direction of the novelist André Gide, but was finally issued at the author's expense in November 1913 by the progressive young publisher Bernard Grasset and met with some success. Proust then planned only two further volumes, the premature appearance of which was fortunately thwarted by his anguish at the flight and death of his secretary Alfred Agostinelli and by the outbreak of World War I.

During the war he revised the remainder of his novel, enriching and deepening its feeling, texture, and construction, increasing the realistic and satirical elements, and tripling its length. In this majestic process he transformed a work that in its earlier state was still below the level of his highest powers into one of the profoundest and most perfect achievements of the human imagination. In March 1914, instigated by the repentant Gide, the *La Nouvelle Revue Française* offered to take over his novel, but Proust now rejected them. Further negotiations in May–September 1916 were successful, and in June 1919 A *l'ombre des jeunes filles en fleurs* appeared simultaneously with a reprint of *Swann* and with *Pastiches et mélanges,* a miscellaneous volume containing "L'Affaire Lemoine" and the Ruskin prefaces. In December 1919, through Léon Daudet's recommendation, À *l'ombre* received the Prix Goncourt, and Proust suddenly became world famous. Three more installments appeared in his lifetime, with the benefit of his final revision, comprising *Le Côté de Guermantes* and *Sodome et Gomorrhe.* He died in Paris on November 18, 1922, of pneumonia, succumbing to a weakness of the lungs that many had mistaken for a form of hypochondria and struggling to the last with the revision of *La Prisonnière.* The last three parts of À *la recherche* were published posthumously, in an advanced but not final stage of revision: *La Prisonnière, Albertine disparue* (originally called *La Fugitive,* though the new title may well have had Proust's authority), and *Le Temps retrouvé.*

Raw material of his novel

His autobiography in the novel. Proust's enormous correspondence (3,000 letters have appeared in print; many more await publication), remarkable for its communication of his living presence, as well as for its elegance and nobility of style and thought, is also highly significant as the raw material from which a great artist built his universe. For À *la recherche du temps perdu* is the story of Proust's own life, told as an allegorical search for truth.

Its fictional narrative. At first, the only childhood memory available to the middle-aged narrator is the evening of a visit from the family friend, Swann, when the child forced his mother to give him the goodnight kiss that she had refused. But, through the accidental tasting of tea and a *madeleine* cake, the narrator retrieves from his unconscious memory the landscape and people of his boyhood holidays in the village of Combray. In an ominous digression on love and jealousy, the reader learns of the unhappy passion of Swann (a Jewish dilettante received in high society) for the courtesan Odette, whom he had met in the bourgeois salon of the Verdurins during the years before the narrator's birth. As an adolescent the narrator falls in love with Gilberte (the daughter of Swann and Odette) in the Champs-Élysées. During a seaside holiday at Balbec, he meets the handsome young nobleman Saint-Loup, Saint-Loup's strange uncle the Baron de Charlus, and a band of young girls led by Albertine. He falls in love with the Duchesse de Guermantes but, after an autumnal visit to Saint-Loup's garrison-town Doncières, is cured when he meets her in society. As he travels through the Guermantes's world, its apparent poetry and intelligence is dispersed and its real vanity and sterility revealed. Charlus is discovered to be homosexual, pursuing the elderly tailor Jupien and the young violinist Morel, and the vices of Sodom and Gomorrah henceforth proliferate through the novel. On a second visit to Balbec the narrator suspects Albertine of loving women, carries her back to Paris, and keeps her captive. He witnesses the tragic betrayal of Charlus by the Verdurins and Morel; his own jealous passion is only intensified by the flight and death of Albertine. When he attains oblivion of his love, time is lost; beauty and meaning have faded from all he ever pursued and won; and he renounces the book he has always hoped to write. A long absence in a sanatorium is interrupted by a wartime visit to Paris, bombarded like Pompeii or Sodom from the skies. Charlus, disintegrated by his vice, is seen in Jupien's infernal brothel, and Saint-Loup, married to Gilberte and turned homosexual, dies heroically in battle. After the war, at the Princesse de Guermantes's afternoon reception, the narrator becomes aware, through a series of incidents of unconscious memory, that all the beauty he has experienced in the past is eternally alive. Time is regained, and he sets to work, racing against death, to write the very novel the reader has just experienced.

The notion of time regained

Its structure and meaning. Proust's novel has a circular construction and must be considered in the light of the revelation with which it ends. The author reinstates the extratemporal values of time regained, his subject being salvation. Other patterns of redemption are shown in counterpoint to the main theme: the narrator's parents are saved by their natural goodness, great artists (the novelist Bergotte, the painter Elstir, the composer Vinteuil) through the vision of their art, Swann through suffering in love, and even Charlus through the Lear like grandeur of his fall. Proust's novel is, ultimately, both optimistic and set in the context of human religious experience. "I realized that the materials of my work consisted of my own past," says the narrator at the moment of time regained. An important quality in the understanding of A *la recherche* lies in its meaning for Proust himself as the allegorical story of his own life, from which its events, places, and characters are taken. In his quest for time lost, he invented nothing but altered everything, selecting, fusing, and transmuting the facts so that their underlying unity and universal significance should be revealed, working inward to himself and outward to every aspect of the human condition. *À la reclzerche* is comparable in this respect not only with other major novels but with such creative and symbolic autobiographies as Johann Wolfgang von Goethe's *Dichtung und Walrrheit* and the Vicomte de Chateaubriand's *Mémoires d'outre-tombe*, both of which influenced Proust.

Proust's projection of his own sexual inversion upon his characters is aesthetically justified by his literary use of this propensity and of snobbism, vanity, and cruelty as a major symbol of original sin. His insight into women and the love of men for women (which he himself experienced for the many female originals of his heroines) remained unimpaired, and he is among the greatest novelists in the fields of both heterosexual and homosexual love.

Reputation and influence. Other frequent charges against Proust are as misleading as those of pessimism, antireligion, or obsession with homosexuality. He has been thought idle, unproductive, weak willed, corrupted by snobbism, ill equipped in philosophy, and absorbed in microscopic detail. In fact, research has shown that he worked unceasingly, prolifically, and with iron will from early youth, always in the direction of his great novel. He was an anti-snob with genuine interest in a brilliant and dying culture, a metaphysician with academic training and individual genius, and an impressionist whose detail is an imagery subordinated to the totality of hie creation. In general, hostile criticism of Proust, being based on assumptions or on misuse of the biographical approach, has tended to reveal deficiencies not in Proust but in the scholarship or sensibility of its exponents. Proust's mature prose is remotely influenced by such varied precursors as the 17th-century moralist La Bruyère, the harsh social memoirist Saint-Simon, the innovating arch-Romantic Chateaubriand, the novel cycle of Balzac, the humanism of Renan and Ruskin, and by such contemporary acquaintances as the novelist Anatole France, the symbolist poets Stéphane Mallarmé, Anna de Noailles, and Francis Jammes, and the decadent Robert de Montesquiou (the chief model for Baron de Charlus). Even so, his style remains one of the most original in all literature. His own direct influence is evident in such works as André Gide's *Les Faux-Monnayeurs (The Counterfeiters)*, Jacques de Lacretelle's *Silbermann*, Virginia Woolf's *The Waves*, and Anthony Powell's novel sequence *The Music of Time*. His reputation, though still imperfectly liberated from the superficial and semihostile assessments of the first generation of Proustian criticism. has always been safe with the general reader, and continues to rise. A new critical tendency, in which Proust's work is considered sympathetically from within and on its own terms, may well predominate in the final judgment.

Proust's interest in a dying culture

MAJOR WORKS

NOVELS: *Jean Santeuil* (first published 1952; Eng. trans. by Gerard Hopkins, 1955); A *la recherche du temps perdu*, consisting of *Du côté de chez Swann* (1914), *A l'ombre des jeunes filles en fleurs* (1918–19), *Le Côté de Guermantes I* (1920), *Le Côté de Guermantes II* together with *Sodome et Gomorrhe 1* (1921), *Sodome et Gomorrhe II*, 3 vol. (1922), *La Prisonnière* (published posthumously 1923), *Albertine disparue* (posthumously 1925), and *Le Temps retrouvé* (posthumously 1927), standard text, ed. by Pierre Clarac and André Ferré, 3 vol. (1954); *Remembrance of Things Past*, 12 vol., trans. by C.K. Scott Moncrieff (1922–30), vol. 12 by Stephen Hudson (1931) and Andreas Mayor (1970), consisting of *Swann's Way, Within a Budding Grove, The Guermantes Way, Cities of the Plain, The Captive, The Sweet Cheat Gone*, and *Time Regained.*

OTHER WORKS: *Les Plaisirs et les jours* (1896; *Pleasures and Regrets*, trans. by Louise Varbse, 1950), short stories; *La Bible d'Amiens* (1904), and *Sésame et les lys* (1906), translations from Ruskin; *Pastiches et mélanges* (1919), parodies and prefaces to the Ruskin translations; *Contre Sainte-Beuve* (written 1908–09, first published 1954; *By Way of Sainte-Beuve*, trans. by Sylvia Townsend-Warner, 1958), critical studies.

BIBLIOGRAPHY. Bibliographies on Proust may be found in RENE DE CHANTAL, *Marcel Proust, critique littéraire*, 2 vol. (1967); in Proust's *Textes retrouvés*, ed. by PHILIP KOLB and LARKIN B. PRICE (1968), and *Lettres à la N.R.F.* (1932); in D.W. ALDEN, *Marcel Proust and His French Critics* (1940), listing books and articles in French, and continued for the period after 1940 and for all languages in his *Bibliography of Critical and Biographical References for the Study of Contemporary French Literature* (1949–53; as *French VII Bibliography*, 1954–).

Manuscript collections: The Bibliothèque Nationale, Paris, possesses the major Proust archive, including manuscripts and workbooks of A *la recherche*, manuscripts of *Jean Santeuil, Conire Sainte-Beuve*, and minor pieces and letters. The University of Illinois, Urbana, Illinois, has a further extensive archive. The Société des Amis de Marcel Proust et des Amis de Combray, Illiers, France, issues an annual *Bulletin* (1950–), including new texts and letters of Proust.

Letters: Correspondance (1970–), ed. by PHILIP KOLB; *Correspondance générale*, 6 vol. (1930–36); letters to Antoine Bibesco (1949; Eng. trans. 1953), Robert de Billy (1930), Lucien Daudet (1929), André Gide (1949), Reynaldo Hahn (1956), Georges de Lauris (1948; Eng. trans. 1949), Paul Morand (1949), Marie Nordlinger (1942), Mme Proust (1953; Eng. trans. 1956), and Jacques Rivière (1955); selected letters in *Letters of Marcel Proust*, trans. and ed. by MINA CURTISS (1950); *Choix de lettres* (1965) and *Lettres retrouvies* (1966), both ed. by PHILIP KOLB; bibliography, chronology, and index of letters in PHILIP KOLB, *Correspondance de Marcel Proust* (1949).

Biography: G.D. PAINTER, *Marcel Proust*, 2 vol. (U.S. titles, *Proust: The Early Years* and *Proust: The Later Years;* 1959–65), a full-scale biography; shorter biographies include ANDRE MAUROIS, A *la recherche de Marcel Proust* (1949; Eng. trans., *The Quest for Proust*, 1950); RICHARD H. BARKER, *Marcel Proust* (1958); P.L. LARCHER, *Le Parfum de Combray* (1945), on Proust at Illiers; MARTHE BIBESCO, *Au bal avec Marcel Proust* (1928; Eng. trans., *Marcel Proust at the Ball*, 1956), a friend's memories.

Criticism: LEO BERSANI, *Marcel Proust: The Fictions of Life and Art* (1965); GERMAINE BREE, *Marcel Proust and Deliverance from Time*, 2nd ed. (1969); HOWARD MOSS, *The Magic Lantern of Marcel Proust* (1962); ROGER SHATTUCK, *Proust's Binoculars: A Study of Memory, Time, and Recognition in A la recherche du temps perdu* (1963)—the above are creative studies of complementary insight and importance; P. HANSFORD JOHNSON, *Six Proust Reconstructions* (1958), radio plays of critical value; P.A. SPALDING, *A Reader's Handbook to Proust* (1952); GERMAINE BREE, *The World of Marcel Proust* (1966), a general study.

(G.D.P.)

Providence, Religious Doctrines and Myths of

Providence is the quality in divinity on which man bases his belief in a benevolent divine intervention in human affairs and the affairs of the world he inhabits. The forms that this belief takes differ, depending on the context of the religion and the culture in which they function.

In one view the concept of Providence, divine care of man and the universe, can be called the religious answer to man's need to know that he matters, that he is cared for, or even that he is threatened, for in this view all religions are centred on man, and man is individually and collectively in constant need of reassurance that he is not an unimportant item in an indifferent world; if he cannot be comforted, to be threatened is better than to be alone in an empty void of nothingness. According to J. van Baal, a Dutch anthropologist,

> Man experiences his universe as a universe full of intentions, a universe which holds a claim on him, addressing him with something undefined, urging him to act or to be in some way or another. The experience is strongest in moments of crisis, when events turn up with such an overwhelming force that it is as if they address their victim, delivering a message to him.

In answer to such a universe, religions must offer a coherent view of God or gods, world, and mankind and must give man and his physical or psychical well-being, or both, a prominent place within this world view. Thus, in all religions Divine Providence or its equivalent is an element of some importance.

NATURE AND SIGNIFICANCE

Basic forms of Providence. Basically, there are two possible forms of belief in Providence. In the first, man believes in more or less divine beings that are responsible for the world generally and for the welfare of man specifically. Although omnipotence as an attribute of gods is rare, it is true that, as a rule, gods and other divine beings have considerable power not only over man but also over nature. The gods take care of the world and of mankind, and their intentions toward mankind are normally positive. The capricious and arbitrary gods of paganism exist for the most part only in the imagination of those Christian theologians who attempt to denigrate the pagan religions. Gods and men are generally connected into one community by reciprocal duties and privileges. The belief in evil spirits does not contradict this belief in Providence but, on the contrary, strengthens it, just as in Christianity the belief in the devil might serve to strengthen the belief in God.

Belief in a cosmic order

In the second form, man believes in a cosmic order in which the welfare of man has its appointed place. This cosmic order is usually conceived as a divine order that is well intentioned toward man and is working for man's well-being as long as he is willing to insert himself into this order, to follow it willingly, and not to upset it by perversion or rebellion; the firmness of the order, however, may become inexorable and thus lead to fatalism, the belief in an impersonal destiny against which man is powerless. In that case a clash between the concepts of Providence and fatalism is inevitable. In most religions, however, both views are combined in some way.

Etymological history of the term Providence. The English word Providence is derived from the Latin term ***providentia,*** which primarily means foresight or foreknowledge but also forethought and Providence in the religious sense; thus, Cicero used the phrase the "Providence of the gods" ***(deorum providentia).*** The Stoic philosophers thoroughly discussed the significance of the term Providence, and some of them wrote treatises on the subject. A hymn to Zeus written about 300 BC by Cleanthes, a Greek poet and philosopher, is a glorification of the god as a benevolent and foreseeing ruler of the world and of mankind. God has planned the world in accordance with this Providence:

> For thee this whole vast cosmos, wheeling round
> The earth, obeys, and where thou leadest
> It follows, ruled willingly by thee.

The author asserts that "naught upon Earth is wrought in thy despite, O God" and that in Zeus all things are harmonized. Seneca, a Roman Stoic philosopher, formulates the belief in Providence in one of his dialogues as follows: man should believe "that Providence rules the world and that God cares for us." The Stoic school disagreed with those who believed that the world was ruled by blind fate; they did not deny that a controlling power exists, but, as everything happens according to a benevolent divine plan, they preferred to call this power Providence. According to the Stoic emperor Marcus Aurelius, God wills everything that happens to man, and for that reason nothing that occurs can be considered evil. Stoic ideas about Providence influenced Christianity.

In later Latin after the emperor Augustus, the word Providence was used as a designation of the deity. Seneca, for example, wrote that it is proper to apply the term Providence to God. Finally, Providence was personified as a proper goddess in her own right by Macrobius, a Neoplatonic Roman author, who wrote in defense of paganism about 400.

Epicurus, a 4th–3rd-century-BC Greek philosopher, contested the Stoic belief in Divine Providence, but the objections of his followers could not change the spiritual climate of the Greco-Roman world. More eloquent, perhaps, than the dissertations of the learned Stoic philosophers were the many stories found in a work by Aelian, an early-3rd-century-AD Roman rhetorician, about strange events and miraculous occurrences ascribed to Providence. Aelian, however, was more interested in sensational stories than in historic accuracy.

Latin meanings of the word

The several meanings of the Latin word ***providentia*** exactly mirror those of its Greek equivalent, ***pronoia.*** Herodotus, the historian of the 5th century BC, was the first Greek author to use the word in a religious sense when he mentioned Divine Providence as the source of the wisdom that keeps nature in balance and prevents one kind of creature from prevailing over all others. Writers such as the historian Xenophon and the biographer Plutarch used the word for the watchful care of the gods over mankind and the world.

The belief in the existence of a blind and inexorable fate can lead to a conflict with the belief in a benevolent Providence. In the Greco-Roman world, where fatalistic belief was strong and where it found a popular expression in astrology, the belief that the whole world, but particularly man, is governed by the stars was contested by Judaism and Christianity. The Talmud, the authoritative collection of Jewish tradition, teaches that Israel is subject to no star but only to God. An example of this conflict is also found in the novel ***The Golden Ass*** by Apuleius, a 2nd-century-AD philosopher and rhetorician deeply interested in Hellenistic mystery cults, which taught a faith that liberated man from the power of the stars. In the novel the hero is converted to the goddess Isis; then, the priest of the goddess addresses him:

> "Lucius, my friend," he said, "you have endured and performed many labours and withstood the buffetings of all the winds of ill luck. Now at last you have put into the harbour of peace and stand before the altar of loving-kindness. Neither your noble blood and rank nor your education sufficed to keep you from falling a slave to pleasure; youthful follies ran away with you. Your luckless curiosity earned you a sinister punishment. But blind Fortune, after tossing you maliciously about from peril to peril has somehow, without thinking what she was doing, landed you here in religious felicity. Let her begone now and fume furiously wherever she pleases, let her find some other plaything for her cruel hands. She has no power to hurt those who devote their lives to the honour and service of our Goddess's majesty."

Christian use of the term

The Christian use of the term Providence, besides being profoundly influenced by Greek and Roman thought, is based on the Old Testament story of the patriarch Abraham's sacrifice of his son Isaac, which is found in the book of Genesis. Abraham tells Isaac, "God will provide himself with a young beast for a sacrifice, my son." The Hebrew language lacks a proper word to express the notion of Providence, but the concept is well known in the Old Testament.

In the New Testament the word pronoia and related words are used rarely, but in no case are they used in the later Christian sense of Providence. This is of interest because the idea of Providence as such is far from foreign to the religious thinking of the New Testament. In the Gospel According to Matthew, for example, Jesus says:

> Are not two sparrows sold for a penny? And not one of them will fall to the ground without your Father's will. But even the hairs of your head are numbered. Fear not, therefore; you are of more value than many sparrows.

Providence as used in Christianity is thus a dogmatic term rather than a biblical term; it indicates that God not only created the world but also governs it and cares for its welfare. A well-known German reference work, Religion in Geschichte und Gegenwart ("Religion in History and the Present"), gives a more elaborate and more theological definition of Providence that, freely translated, is as follows:

> God keeps the world in existence by his care, he rules and leads the world and mankind deliberately according to his purpose, and he does this in his omnipotence as God the Creator, in his goodness and love as revealed by his son Jesus Christ, and to further the salvation of mankind through the Holy Spirit.

BASIC CONCEPTS AND SCOPE

Qualities of **the** divinity. The concept of Providence is rooted in the belief in the existence of a benevolent, wise, and powerful deity or a number of beings that are benevolent and that are either fully divine or, at least, appreciably wiser and more powerful than man (*e.g.*, ancestors in many religions). Benevolence is the primary requirement. In northern Malawi, death in later life is usually ascribed to the will of the ancestors, but a miscarriage or the death of a very young child is not considered to be their work because such an act would be in contradiction with their benevolent and helpful attitude toward their offspring. The three attributes, however, are all essential for the concept of Providence: the divine being or beings must be well intentioned toward man, must have the necessary wisdom to know what is good for mankind, and must have the power to act on this intention and insight. Benevolence does not exclude the possibility of punishment of men in cases of transgression. There is probably no god in existence who only rewards and helps and never punishes his believers.

Providence, however, need not operate in a direct way; it may operate through many intermediary beings—*e.g.*, the ancestors and various kinds of spirits in several nonliterate religions or the angels in Christian and Muslim belief—or the concept may be implicit in and expressed by a fixed world order, a cosmic order that makes human life possible biologically, socially, and spiritually and that guarantees its existence in the future. Thus, Providence may become a more or less impersonal principle of cosmic order as instituted and maintained by a divine being, but, if the starting point of a benevolent and just divine being is completely lost sight of or if it is consciously denied, then Providence becomes fate.

Cosmic order. ***Notion*** of cosmic order. Although the introduction of intermediary beings brings no essential change in the idea of Providence as the divine watchful care for the benefit of mankind, the notion of a cosmic order changes the picture profoundly. Even if the cosmic order is conceived as a benevolent order in which man is able to feel safe and whose very existence reassures him, such an order is different from the personal relationship between man and his god or gods. The concept of an unchangeable world order requires a different reaction. A persona1 god may, perhaps, be moved by prayer and sacrifice to give or to prevent events; when the order of the world is fixed, however, the course of events cannot be changed by these or any other means. There is probably no religion that acknowledges an all-embracing world order without any exceptions at all. Generally, man has such an important function in the order of the world that he also has a certain opportunity to manipulate this order, at least to a certain extent, for instance by sacrifice or other ritual acts. One opening is presented by the fact that the cosmic order is valid for everything of a more general character, but as a rule the divine will or the free will of man or chance operates on the level of the common occurrences and daily life of the individual. Though in theory the order may govern everything, a large field is left open for different concepts to function. In some cases even uncertainty and chance have their proper place within a determined order. In Yoruba religion (Nigeria), for example, the god Eshu represents the principle of chance and uncertainty and of all that cannot be foreseen. He is one of the gods of the pantheon and has his own sanctuaries and priests.

Manipulation of order and predestination

Another possibility for combining the idea of a personal divine will with a fixed course of events is the concept of predestination best known from Islām and some forms of Calvinism (derived from the thought of John Calvin, a 16th-century French Protestant Reformer) and also important in the theology of Augustine of Hippo, a 4th–5th-century Church Father. Although predestination essentially is concerned with salvation—the question of whether a certain individual will be saved or damned—it is a concept that easily lends itself to a more general application. In a few religions the idea that the individual chooses his own destiny before birth is encountered; *e.g.*, the Batak of Sumatra and some West African tribes. In this conception free will and predestination merge.

In all religions that acknowledge the existence of a more or less impersonal cosmic order, man is expected to work with the cosmos, to insert himself into the cosmic order. Man's behaviour in all fields is governed by a set of rules that are all based on the same principle: to act and to be in harmony with the order of the world, which is natural and divine at the same time.

The cosmic order is given with the creation of the world, but it is possible to question the relation of the Creator to the world after creation. On one hand, there is the belief that God will not abandon the world he has created; on the other, the belief that God created the world and the cosmic order in such a manner that to a great extent the course of the world is fixed from the first beginning and he is no longer involved in it. The latter was, in fact, the thesis of the 17th- and 18th-century Deists in Europe (see DEISM). The fact of creation helps man to believe in Providence because it would be inconsistent for the creator god or gods not to care for the further existence of the created world. Only persistent disobedience and open rebellion can then furnish **a** reason for the Creator to abandon or destroy the world. This situation is expressed in the myths of a great flood or some other form of destruction sent as a punishment. There is, however, never a total destruction of the world in these myths, although this final solution may be threatened for the eschatological (ultimate end) future. It may also be promised, if the eschatological events are construed as the definitive institution of a world order that is perfect for all eternity and will never deteriorate.

Cosmic order and ethical principles

The cosmic order is often clearly contrasted with the disorder of chaos. The cosmic order is a total order; it comprises not only all natural things but also social and ethical rules. This does not mean that cultures and religions centred on a cosmic order have no clear idea of distinctive ethical principles but that ethics is considered as one function of the total cosmic order and as such can never be quite independent. The rules of ethics depend on and are derived from the more general rules that govern the cosmos in its totality; they are no more than special manifestations of these general rules. An example of this attitude can be found in the Greek hymns in praise of the goddess Isis. She is honoured as the queen of heavens; she divided the earth from the heaven, showed the stars their paths, and ordered the course of the sun and the moon. But the same hymn says that she ordained that children should love their parents, that she taught men to honour the images of the gods, and that she made justice stronger than gold and silver. She established penalties for the people practicing injustice and taught that men should have mercy with suppliants. She is also praised because she invented writing, devised marriage contracts, invented navigation, and watches over all men who sail on the sea.

Personal and impersonal forms. The cosmic order can appear in a personalized form, as, for example, the Egyptian goddess Maat; but this personification of the cosmic order is not general: the Iranian Asha, the Indian *ṛta,* and the Chinese Tao are all to a high degree impersonal. Maat represents truth and order; her domain includes not only the order of the nature, but also the social and ethical orders. She plays an important role in the judgment of the dead: the heart of the deceased is weighed against the truth of Maat. She is often called the daughter of Re. In this case, Re is the creator god who not only created the world but also founded the cosmic order as represented by Maat. Her importance is also apparent in the conception of the Maat sacrifice. In Egypt sacrifice is not so much a gift of men to the gods as a sacral technique that enables man to contribute to the maintenance and, if necessary, the restoration of harmony and order in the world. Not only must man live according to Maat but also the gods must live by her truth and order; according to Egyptian texts, the goddess Maat is the food by which the gods live.

Asha, *ṛta,* and Tao

The idea of a determined cosmic order that is natural as well as ethical is an important concept in the Persian religion of Zoroastrianism (also called Mazdaism and, in India, Parsiism) founded during the late 7th and early 6th centuries BC by Zoroaster (Zarathustra). This idea is called Asha and is the counterpart of Drug, which represents evil and deceit and the disorder connected with these. Asha is connected with the sacred element fire. The Indian concept of *yta* forms the Indian counterpart of Asha. The gods, especially the Adityas, protect the world against chaos and ignorance and maintain the world order, which, however, exists independently from the gods. Although the power of *ṛta* operates according to its own principles and laws, man is able, provided he knows the right methods, to manipulate this power to some extent for his own benefit. The proper means for this manipulation is found especially in older Hindu sacrifice. The gods are generally benevolent and friendly toward men who follow *ṛta,* and they punish their own enemies and those of the world order, which in India, too, embraces the social ethical rules.

The concept of Tao is of great importance in Chinese religion, especially in Taoism, founded by Lao-tzu according to tradition in the 6th century BC. Lao-tzu is the author of the *Tao-te Ching* ("Classic of the Way and Its Power") in which he expounds this concept in a manner that is more mystical than philosophical. Tao, literally translated "road," is a difficult and complex concept. It certainly represents the cosmic order, but in Taoism it is even more than that. It is also the concept that gives existence meaning; it is the primeval power that forms the foundation of all that is; and, in some cases, it is even used to designate some kind of high god. Taoism is a mystic religion, and the ***Tao-te Ching*** is a mystic treatise in which the essence of the Tao is expounded in many parables and metaphors because it cannot be expressed rationally.

Manv related concepts exist. The Greek Moira, for instance: is comparable to Asha and *ṛta*; it lacks, however, the mystic overtones of Tao. The Moira in classical Greek religion is not yet fate as this idea was found in Greco-Roman times. The concept of cosmic order may function either in a religious or in a philosophic context; *e.g.,* the pre-established harmony ***(karmonia*** *praestabilita)* in the philosophy of Gottfried Wilhelm Leibniz, a German Rationalist, is the cosmic order that holds together and unifies the innumerable individual units, called monads by Leibniz.

Particular objects of Providence. Although cosmic order is necessarily a general idea comprising the whole of the world and all that exists in it, the concept of Providence may be more particular: the benevolent aspect of Providence may be confined to a special group of people or at least be specially related to that group; or a number of patron gods or saints may watch over some specific activity or smaller group. This accounts for the idea of a chosen people watched over and led by a just and loving God. The ancient people of Israel is, perhaps, the best known example; the concept, however, is widespread. Patron gods and patron saints who are particularly charged with caring for some small group, craft, or activity or who operate in special circumstances, such as during illness or war, occur in most religions and are popular in many.

Although Providence in most religions operates primarily for the welfare and the salvation of the community as a whole, it may also be experienced as personal guidance. This latter phenomenon is common in some diverse cultures—*e.g.,* that of the Plains Indians of North America and in some forms of Protestantism in which generally each person is expected to have a private experience of divine guidance. In other cultures and religions, personal guidance is often a prerogative of some person or persons singled out for some reason by God or the gods.

CRITICAL PROBLEMS

It is clear that the concept of Providence by its central position in many religions is connected with numerous other aspects of religion. In monotheistic religions Providence is a quality of the one divinity; in polytheistic religions it may be either a quality of one or more gods or it may be conceived as an impersonal world order on which the gods, too, more or less depend. In the latter case, Providence may lose its aspect of benevolence and become inexorable fate or fickle chance. Most religions show a certain ambivalence; for fate and Providence do not always form a clear-cut contradiction.

Still another form of ambivalence occurs between fate or divine will and the will of man when the latter is conceived as free, or at least free to a certain degree. In some religions the benevolent aspect of Providence appears as grace, and a discussion may arise about the relationship between free will and grace. Perhaps the most difficult problem connected with the notion of Providence is the existence of evil; men have perennially coped with the question of how to reconcile the idea of a provident God or gods with the evident existence of evil in the world.

The problem of evil

BIBLIOGRAPHY. No general introduction to the subject written from the point of view of the science of religion exists, but useful articles are found in some specialized encyclopaedias, such as ***Hustings' Encyclopaedia of Religion and Ethics*** (1919). Further information has to be gathered from monographs about specific problems related to Providence; *e.g.,* JOHN BOWKER, ***Problems of Sufferingin the Religions of the World*** (1970). Much information on cosmic order is found in MIRCEA ELIADE, ***Traité d'histoire des religions,*** new ed. (1964; Eng. trans., ***Patterns in Comparative Religion,*** 1963). *A* great amount of literature on ancient philosophy is available: REINOUT BAKKER, ***Lot en daad, geluk en rede in het Griekse denken von Solon tot Marcus Aurelius*** (1957); *c.* PARMA, ***Pronoia und Providentia, der Vorsehungsbegriff Plotins und Augustins*** (1972); MAX POHLENZ, ***Die Stoa, 2*** vol. (1948–49); ROBERT M. WENLEY, ***Stoicism and Its Influence*** (1924).

(T.P.v.B.)

Psilopsida

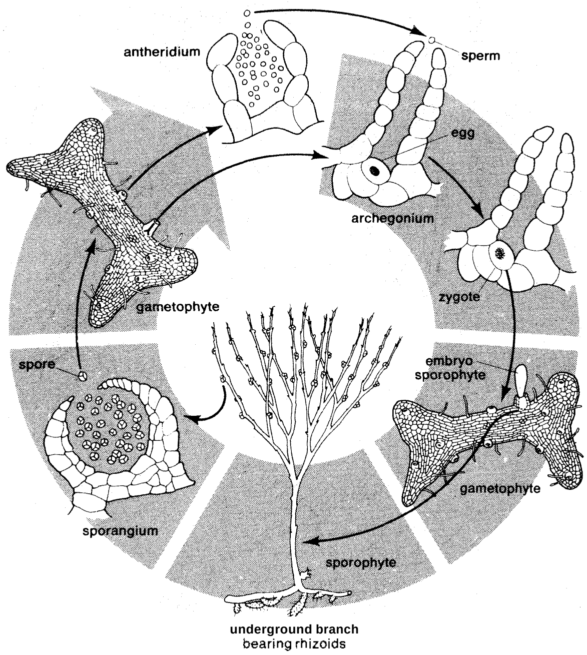
The Psilopsida is a class of primitive vascular plants (division Tracheophyta) with two living genera and several extinct members, present as early as the Silurian Period (430,000,000 years ago). Psilophytales, the extinct order of the Psilopsida, is academically interesting as the apparent ancestor of the living order Psilotales and is further considered by many botanists to be ancestral to all other vascular plants—*i.e.,* plants with specialized cells for conducting food and water.

General features. Early psilopsids ranged from plants a few centimetres tall, and Rhynia-like plants 25 to 40 centimetres (ten to 15 inches) tall, to ***Asteroxylon*** and others, more than one metre (about 40 inches) high. Unlike the leafy shoot systems of present-day flowering plants, psilopsids had spinelike leaves instead of true leaves along the stem, and the stems were capable of photosynthesis (based on the presence of holes, called stomates, and air spaces).

The living genera, ***Psiloturn*** (whisk ferns) and ***Tmesipteris,*** are terrestrial or epiphytic (living on other plant surfaces). Psiloturn, an upright, green plant that looks like a leafless shrub about 30 centimetres high, is tropical and subtropical in distribution, reaching as far north as Florida and Hawaii; Tmesipteris, a hanging green epiphyte, is found in Australia, New Caledonia, New Zealand, the Philippines, and other islands of the South Pacific area. The extinct genera grew in lowlands or were aquatic. *Psilotum* is sometimes grown in greenhouses.

The extant genera

Certain features are common to both the living and extinct psilopsids. The conspicuous green plant, the sporophyte, consists of an aerial system and an underground stem (rhizome) system, both of which repeatedly fork dichotomously—*i.e.*, into two equal branches. No roots are present, the rhizome performing the functions of a root. In Psiloturn the "leaves" are very small without vascular tissue. In Tmesipteris the "leaves" are larger, flattened structures, each with one unbranched midvein. (It is thought that the "leaves" of the psilopsids may not have had the same evolutionary development as the true leaves of ferns and seed plants.)



Life cycle of a psilopsid.

Life cycle. Spores germinate to form gamete-bearing plants (gametophytes), or prothallia, as they are commonly known when they are newly developed. These plants are very small, measuring about one millimetre in diameter and several millimetres in length. The gametophytes may grow on trunks of trees or underground. Like the sporophyte, the gametophyte branches repeatedly. It is devoid of chlorophyll, however, and lives a saprophytic existence (*i.e.*, on decaying organic matter), aided presumably by a fungus living within it. These plants are so similar to the underground rhizomes that they can be identified only if gamete-producing structures (gametangia) are present on the gametophyte. Two types of gametangia are scattered over the surface and are intermingled. The egg-producing structure, the archegonium, consists of a sterile jacket of four rows of cells enclosing another row of cells, the lowest one of which is the egg. The sperm-producing structure, the antheridium, is globose and emergent on the surface of the gametophyte; it consists of a jacket of sterile cells enclosing many spermatocytes, which upon release become swimming multiflagellated sperm. To effect fertilization a sperm must swim to the archegonium and down a canal-like passage to the egg—the passage being created by the disintegration of the cells adjacent to the nonmotile egg. The first division of the zygote (fertilized egg) is transverse (perpendicular to the long axis of the archegonium). The outer cell continues to divide and forms the axis of the developing sporophyte. The inner cell gives rise to a multicellular foot, which is in close contact with the gametophyte and presumably functions in the transfer of nutrients from the gametophyte to the young sporophyte. Frequently the young sporophyte axis branches very early during its existence; one branch may become a green aerial shoot; the other becomes a rhizome. The rhizome continues to branch, but any of the underground branches may turn up and develop into a green aerial shoot and form leaves and sporangia.

Form and function. All psilopsids have a simple internal organization. The fossil members had smooth stems or stems with spines or small "leaves." The spore case (sporangium) is three-lobed in Psiloturn and two-lobed in Tmesipteris. The sporangia in *Psilotum* appear generally to be axillary in position (in the upper angles of the "leaves") but are actually terminal on short branches; the leaves associated with the sporangia are appendages on the short branch. The presence of vascular tissue at the base of a sporangium, or even in partitions separating compartments in a sporangium, supports the concept of terminality of the sporangium.

The protostele form of vascular tissue

The vascular tissue is in the form of a slender cylinder occupying the centre of the stem (protostele) in the rhizomes of both living genera. The protostelic condition is true of the upper branches in Psiloturn, whereas in the lower branches of *Psilotum* and in the stems of *Tmesipteris* the vascular tissue is not central but forms a siphonostele, a cylinder of strands surrounding nonvascular cells. The cells of the outer layers of the stem contain chloroplasts, and in *Psilotum* this region accounts for most of the photosynthesis. One interesting feature of stem anatomy in Psiloturn is the presence of an endodermis—a layer of cells surrounding the vascular cylinder in which each cell has a substance (lignin) deposited in the form of a band in the cell wall. An endodermis is a characteristic feature of roots in all vascular plants but occurs only rarely in stems of seed plants.

Spores formed in the sporangia have nuclei with half the number of chromosomes found in nuclei elsewhere in the plant. Only one type of spore is formed; consequently these plants are said to be homosporous. Spores are released by the splitting of sporangia along longitudinal slits.

Evolution and classification. Unquestionably the extinct order Psilophytales is an ancient and primitive group of vascular plants. The Psilotales also are ancient, but some botanists question the assumption that the two orders have a close linear relationship. The extinct order takes its name from Psilophyton, a fossil from Quebec, Canada, described in 1859 by Sir John William Dawson. It was almost 60 years later before the real significance of this plant was understood. At that time some extremely well-preserved ancient vascular plants (Rhynia and *Asteroxylon*) from deposits of the Devonian Period (about 395,000,000 years ago) in Scotland were found to resemble Psilophyton, and the order Psilophytales was established. Some botanists believe that there is enough variation in the extinct Psilophytales to warrant the establishment of at least three orders or even classes. Future systems of classification may well reflect this change. Groups indicated with a † (dagger) in the classification below are extinct and known only from fossils.

CLASS PSILOPSIDA

Primitive seedless vascular plants, lacking roots and leaves, the upright stem serving as photosynthetic organ and the horizontal stem serving to anchor and to absorb nutrients. Only 2 living genera but many extinct forms.

†Order Psilophytales

Wholly extinct group, whose members were especially abundant in the Devonian Period (345,000,000–395,000,000 years ago). Rhynia, Asteroxylon, and Psilophyton are well-known examples.

Order Psilotales

Two living genera: *Psilotum*, like a leafless shrub, in the tropics and subtropics; and Trnesipteris, a hanging epiphyte, with flattened leaflike structures, in Oceania and the South Pacific area.

BIBLIOGRAPHY. The following general textbooks have sections dealing with the psilopsids: T.E. WEIER, C.R. STOCK-

ING, and M.G. BARBOUR, Botany: An Introduction *to Plant* Biology, 4th ed. (1970); R.F. SCAGEL et al., An Evolutionary Survey of the Plant Kingdom (1965); H.N. ANDREWS, Studies in Paleobotany (1961); and THEODORE DELEVORYAS, Morphology and *Evolution* of Fossil Plants (1962), both of which give detailed accounts of fossil forms and their relationships with living plants; A.S. FOSTER and E.M. GIFFORD, *Comparative* Morphology of Vascular Plants (1959); and D.W. BIERHORST, Morphology of Vascular Plants (1971), both of which provide detailed treatments of vascular plants together with theory and interpretation; and H.P. BANKS, Evolution and Plants of *the* Past (1970), a modern and readable account of the earliest evidence of plant life and evolution. Journals that regularly publish articles on vascular plant morphology include the *American* Journal of Botany, the Botanical Gazette, Phytomorphology, the Annals of Botany, and *Palaeontographica*.

(E.M.G.)

Psittaciformes

The avian order Psittaciformes contains more than 300 species of generally brightly coloured, noisy, tropical birds, to which the general name parrot may be applied. Various species are known as keas, cockatoos, cockatiels, lories, lorikeets, parrotlets (or parrolets), parakeets, budgerigars, rosellas, conures, lovebirds, amazons, and macaws. Although the first accurate written reference to a parrot is frequently credited to the Greek historian Ctesias, of the 5th century BC, who described clearly what is now called the blossom-headed parakeet (*Psittacula* cyanocephala) of India, there is no doubt that parrots were associated with man much earlier, for aborigines on all continents had parrots as pets when first visited by explorers. Parrots of many kinds have been long transported to Europe for zoos and private collections. Affluent citizens of early Rome often kept parrots in their homes and even esteemed them as delicacies of the dinner table.

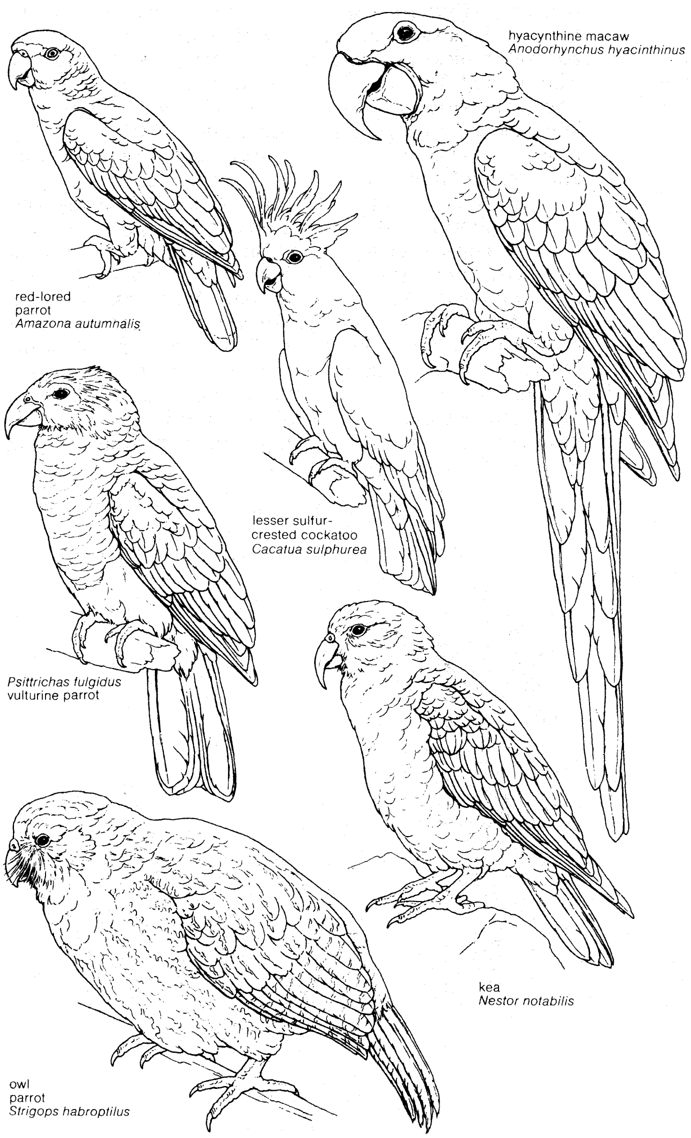


Figure 1: Body plans of representative larger Psittaciformes.
Drawing by G. Tudor

GENERAL FEATURES

Appearance. Parrots vary in total length from eight to 100 centimetres (three to nearly 40 inches), the latter in longtailed forms. The short neck and sturdy body, along with the stout feet and thick bill, give them a bulky appearance. The broad wings are often pointed; the tail is highly variable in both length and shape. In some species the tail is short and rounded or square; in others, such as the macaws (Ara), it is long and pointed. In numerous species the central tail feathers are very long, surpassing the body in total length. In the five species of racket-tailed parrots (*Prioniturus*) the central tail feathers are longer than the others and are spatulate, the middle part of the feather shaft being bare. No parrot has a forked tail. Pointed wings and a long tail usually are found in species that fly great distances; rounded wings and blunt tails typify the more adept climbers. Most parrots are swift on the wing, although they normally fatigue quickly.

Distribution. Parrots are primarily birds of the tropics. Their distribution encompasses the tropical and south temperate regions of the world, including Madagascar, New Zealand, and the West Indies. In Asia they occur throughout almost all of India but extend northward only to the Himalayas and southern China. They are absent from Europe. In the early 1970s in North America they ranged north only to the extreme southwestern United States (one species, the thick-billed parrot, Rhynchopsitta pachyrhyncha). Prior to the early 1900s, however, the Carolina parakeet (Conuropsis carolinensis) inhabited most of the eastern United States; it was brought to extinction by human persecution. The last captive died in the Cincinnati Zoological Garden in 1914, but the last generally accepted observation in the wild was a flock seen in Florida in 1904, although it has been claimed that they existed in South Carolina until 1938. In the Southern Hemisphere a number of parrots range to Tasmania and New Zealand, and in South America one species is found on Tierra del Fuego, but none are found on the southern tip of Africa.

Although parrots are found in most parts of the three southern continents (excluding the Sahara), they are not evenly distributed. Of the 81 genera recognized in a 1937 revision of the taxonomy of parrots, the Neotropical region (South and Central America) has 28, not one of which is found elsewhere. The Nearctic region (temperate and arctic North America) has had only the now-extinct genus Conuropsis. The Australo-Papuan region, which encompasses Australia, New Zealand, New Guinea, Celebes, and many of the Pacific Ocean islands, has 44 genera, three of which are shared with the Oriental region. The Oriental region, which, in addition to Asia, includes the Philippines and most of Malaysia, has nine genera, including the three shared with the Australo-Papuan region and one shared with the Ethiopian region. The Ethiopian region, comprising Africa, Madagascar, and nearby islands, has only six genera, one of which is the extinct Mascarinus from Reunion Island, 650 kilometres (400 miles) east of Madagascar. The Palearctic region (Europe and northern Asia) has none. Most parrot genera contain only a few species; only six include more than ten species, four of which occur in the Neotropics.

Importance *to* man. The qualities of parrots, especially the ability of many species to imitate human sounds, make them popular as pets. The African gray parrot (Psittacus erithacus) and some species of amazons (Amazona) from the New World tropics are particularly good mimics. There is no evidence to suggest, however, that talking parrots realize what they are saying.

Parrots as pets

Another appealing attribute of parrots is their display of affection, not only to others of their own species but also to man. Pairs of many species, especially the lovebirds (Agapornis), are together almost constantly, nibbling each other's feathers with seeming affection; if one bird disappears its mate sometimes dies, apparently of loneliness. Many parrots seem to delight in being petted and scratched, which is rare among birds. Parrots have extremely powerful jaws, however, and an indiscriminate attempt to pet them can result in a severe bite. The use of the toes for climbing and food handling, in much the same manner as man uses his hands, also makes parrots appealing. Their longevity, bright colours, intent gaze, ability to learn tricks, and willingness to remain on a perch instead of fluttering about contribute to man's fondness for various kinds of parrots as pets. Finally, most species are vegetarians and thrive on a varied diet. This circumstance, as well as the fact that their droppings typically are dry and compact, means that parrots require little care.

Most species of parrots have been kept in captivity at one time or another, and most have been bred, with the exception of the pygmy parrots (Micropsitta). A large zoological garden may have more than 100 species on display at one time. No parrot has been domesticated in the sense of gallinaceous birds and waterfowl, although in recent decades breeders have produced a variety of colour strains of the budgerigar (Melopsittacus *undula*tus), commonly called shell parakeet. In the mid-1950s "budgies" became popular household pets in the United States; within a decade more than 5,500,000 people had at least one in their homes, Captive parrots, especially the larger species, are long-lived. Claims of 80 or even 100 years are frequent and perhaps true, although thus far impossible to document. A convincing record of a 56-year-old greater sulfur-crested cockatoo (Cacatua galcrita) exists, however.

In the early 1930s the importation and sale of parrots in the United States was drastically curtailed by quarantine laws designed to combat psittacosis, or parrot fever (better called ornithosis, as the disease by no means is restricted to psittacine birds), a respiratory virus that can infect humans. Antibiotics now available reduce the severity of the disease, and most of the restrictions on the parrot trade have been relaxed.

The primary economic importance of parrots derives from their popularity with aviculturists. So popular are some species that the Australian government passed laws forbidding export because wild populations were being decimated. Unfortunately, only a few of the many individuals captured ever reach the comparative safety of a comfortable cage, because of mishaps en route. Some parrots, especially certain of the Australian seed-eating species, damage crops and therefore are hunted and killed.

NATURAL HISTORY

Habitat and food choice. Most parrots inhabit forests, although a few live in grasslands. Of the forest-inhabiting species many forage along the forest edge and on the ground. Some parrots live in the mountains, especially in the Himalayas and Andes; the New Zealand kea (*Nestor* notabilis) is a mountain inhabitant but obtains much of its food in the forested valleys; it nests either in high elevation forests or near the forest edge. Many Australian parrots, such as members of the genera *Neophema* (grass parakeets) and Psephotus, are found in dry, open grasslands, although typically where trees are scattered through the habitat. The budgerigar and the rare night parrot (Geopsittacus occidentalis) also are Australian grassland birds.

Parrots feed almost entirely on plant materials. The smaller species tend to utilize grass seeds, berries, fruits, and the juices of blossoms; the larger forms obtain fruits and nuts from trees, and bulbs, tubers, and roots from the ground. When digging, many parrots also capture larval and adult insects, and raven, or black, cockatoos (Calyptorhynchus) gnaw through bark to obtain wood-boring beetles. Many kinds of nectar-eating birds suck juices through tubelike tongues, but brush-tongued parrots feed on nectar by crushing flowers and licking the juices. The tiny pygmy, or woodpecker, parrots feed on fruit, arboreal termites, and fungi. The kea feeds on dead sheep and carrion and will attack sick, injured, or trapped individuals, but rarely will it harm healthy sheep.

Aggregations of parrots

Social behaviour. Typically, parrots are gregarious and noisy, often forming small groups—sometimes huge flocks—flying rapidly high overhead and screeching. Their seemingly conspicuous bright colours are somewhat misleading, for a group of parrots in foliage is difficult to discern. The grassland-inhabiting parrots are nomadic and often occur in flocks of tens or even hundreds of thousands. The development of agriculture in the interior of Australia, particularly the increased availability of water. has resulted in larger populations of several species, such as the corella (Cacatua *sanguinea*) and budgerigar.

The vocalizations of most parrots are loud, raucous screeches; generally the larger the species the more ear-splitting the calls. The voices of some of the smaller ones include chattering and twittering notes pleasant to the human ear. About 12 different calls, each announcing a different mood, have been identified for the greater sulfur-crested cockatoo. The amazing mimetic abilities of many parrots mentioned above are expressed only in captivity.

Reproduction. Parrots are monogamous. Some species breed colonially; others space themselves through the nesting habitat. Based on the fewer than ten species that have been studied extensively, courtship and behaviour to maintain the pair bond may include vocalizations, bill caressing, mutual preening, bowing, wing-raising, tail-spreading, and feeding the mate.

With few exceptions, parrots nest in holes in trees. Some species add nest material such as leaves, fibres, and bark strips; others lay their eggs on the floor of the cavity. Some lovebirds cut leaves into strips, which are then tucked into the feathers of the back for transportation to the nest. Several parrots, including the pygmy parrots and the orange-fronted parakeet (*Aratinga* canicularis), hollow out cavities in termite nests. Exceptional in the family is the monk parakeet (Myiopsitta *monachus*) of South America, which builds a communal stick nest in trees. Several species nest in rock crevices or earthen caves; examples include the burrowing parrot (*Cyanoli*seus patagonus), kea, night parrot, and the flightless owl parrot (Strigops habroptilus). The ground parrot (*Pezoporus wallicus*) lays its eggs in a shallow cup on the ground.

Parrot eggs are white and usually nearly spherical. Generally the larger species lay only two eggs and produce one brood per year; the smaller species lay up to six or even eight or nine eggs and may breed two or three times per year. Incubation time, which generally varies directly with size, ranges from 16 to **30** days in the few species for which there is documentation. Either both parents or else only the female may incubate the eggs. The young hatch naked or with very sparse down. They are altricial—that is, they are helpless and require complete parental care—and they are also nidicolous—that is, they remain in the nest for some time after hatching. The young are fed by regurgitation, typically by both parents. Care of the young may continue for several weeks after they have left the nest.

FORM AND FUNCTION

The yoke-toed foot

Foot structure. Parrots can be distinguished from other birds by the structure of the feet and bill. Most birds have the four toes arranged with three directed forward—the inner (11), middle (111), and outer (IV) —and one backward, the hallux (I). This condition, called anisodactyl, literally means without equal toes, referring to the unequal arrangement. Parrots have two toes (the inner and middle) directed forward, and two directed backward; this arrangement is called zygodactyl, which literally means yoke-toed and refers to the occurrence of toes in pairs. Zygodactyl also occurs in

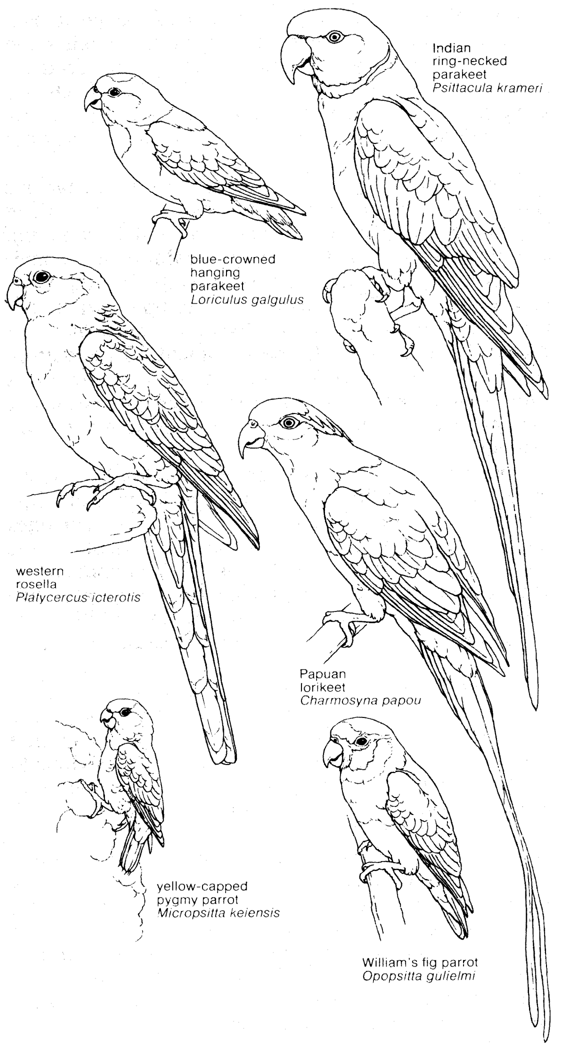


Figure 2: Body plans of representative smaller Psittaciformes.
Drawing by G. Tudor

woodpeckers and their allies (Piciformes), cuckoos (Cuculiformes), and some other birds. The proximal (upper) bone of a bird's foot, the tarsometatarsus (commonly considered the lower leg), lies between the elevated heel joint and the toes. In parrots it is short and stout, and at least one toe is always longer. It is the characteristic short, thick tarsometatarsus—or tarsus, as the entire region is called—and the zygodactylous, long, strong toes that enable parrots to climb and manipulate objects so ably. The entire foot is encased in tough skin covered with small scales.

Bill and skull. The most distinctive morphological trait of parrots is the strongly hooked, powerful bill. Superficially the bill resembles that of the hawks and owls, but the upper and lower mandibles of parrots normally have a stronger and more uniform curve. Often the anterior edge of the lower jaw is broad and truncate. The under surface of the upper bill usually possesses transverse or oblique file-like corrugations where the lower jaw occludes. These file-like ridges, along with the highly manipulative tongue, assist in holding seeds as the bird uses the chisel-shaped cutting edge of the lower bill to peel away a seed cover.

The use of the bill for manipulating objects, cracking hard nuts, and as a third "foot" in climbing are all possible because of a highly kinetic (movable) upper jaw. Most living birds have such a kinetic upper jaw, which is connected to the skull dorsally by a hinge, and is able to be moved independently by swinging on this hinge, but nowhere among birds is this kinesis better expressed than in parrots. The raising of the upper jaw can be described as follows: all movement of the upper jaw originates at the point of attachment between the skull and the quadrate bone, which forms the hinge between the skull and the lower jaw. Two series of bones (the quadratojugal-jugal series and the pterygoid-palatine series), both of which lie in the roof of the mouth, are situated between the quadrate and the upper jaw. Wnen the quadrate is swung forward, the two series of bones slide forward, causing the upper jaw to swing upward on its hinge with the skull.

The short, thick, and fleshy tongue of parrots shows a variety of specializations at the tip; one found in several groups is a brushlike fringe. Primarily, the tongue functions to manipulate and hold food. Those parrots with brushlike terminal papillae (nipple-like projections) on the tongue use them to hold juices, as a brush holds paint.

All parrots possess a cere, an area of soft skin surrounding the nostrils; it may be naked or covered with small, soft feathers. In adult budgerigars the cere is blue in males and tan in females.

The orbits (eye sockets) of some, but not all, parrots are ringed with bone. Other features of the parrot skeleton include a prominent keel on the sternum (breastbone), except in the flightless owl parrot of New Zealand, and a highly variable furcula (wishbone), which may be normal, weak, unfused medially, or absent.

Plumage coloration

Skin and plumage. Parrots have relatively few feathers, which are hard in texture and normally gaudy in colour. Many species are bright green with patches of red, orange, yellow, blue, or white; the plumage of others is predominated by the latter colours. A few parrots are brown or all green. Sexes are alike or nearly so, with a few notable exceptions. One, the eclectus parrot (Lorius roratus), was for many years thought to be two separate species until it was noted that only males were known for the predominantly green "species" and only females for the wine-red "species." The head is crested in a few parrots, especially among the cockatoos (Cacatuinae).

Powder downs, which occur in a variety of birds including some parrots, are specialized feathers, usually found in well-defined patches, that produce a powdery substance used to clean and waterproof the other feathers. They are well developed in cockatoos, in which they occur primarily as a pair of lateral rump patches.

Skin glands, which are abundant in mammals, are almost entirely lacking in birds, with the exception of the oil gland. The oil produced in this gland—also known as the uropygial gland because of its location at the base of the tail and as the preen gland because of its function—is used, like the powder down, to clean and waterproof the feathers. Oil is squeezed from the gland, and birds either use the bill to apply it to the feathers or rub their heads directly over the gland. The nipple of the gland, which protrudes through the skin at the base of the tail dorsally, is surrounded by a tuft of feathers in parrots. Not all parrots have an oil gland; for example, it is absent from the *Amazona,* Brotogeris, and *Pionus* parrots and greatly reduced in some others.

EVOLUTION AND CLASSIFICATION

The extinction of island species

Evolution. Dispersed on a multitude of small islands, parrots have always been vulnerable to extinction, and in recent times the group has suffered increasingly in this regard. From 1680 to the early 1960s at least 16 species disappeared entirely, and another 14 became endangered. Most of the extinct species lived on small islands or on large islands in the West Indies; only the Carolina parakeet had an extensive continental range.

The place of evolutionary origin of the parrot family is not known. The greatest structural diversity is exhibited by parrots from the Australo-Papuan region, but the greatest number of species occur in the Neotropical region. Some authorities doubt that either was the original home of this ancient and distinct group, although it probably did arise somewhere in the Old World. The scanty fossil record is of no assistance in determining the original home of parrots, but it does provide an indication that the group is very old. Typical parrots are known from fossils from the Miocene epoch (about 20,000,000 years ago) of France and from North America. The fact that the Old and New World share no genera and that five of the six Ethiopian genera are not found elsewhere supports the antiquity of the group. An American biologist, Philip J. Darlington, has suggested that parrots and pigeons (Columbiformes), which show similar characteristics of distribution, may have been dominant everywhere before the rise of passerine, or perching, birds (Passeriformes), which may now be replacing them.

Annotated classification. The following classification follows the scheme proposed in 1959 by the German ornithologist Hans von Boetticher. He treated the parrots, as did most of his predecessors, as one family, with seven subfamilies, the two largest of which are further subdivided into tribes. Boetticher reduced the 81 genera recognized by U.S. ornithologist James Lee Peters in 1937 to 75, a figure that reflects the modern view of broader genera.

ORDER PSITTACIFORMES (parrots and allies)

Chunky, primarily tropical birds, with short necks, tarsi, and wings. Distinctive bill, short and strongly hooked, the upper mandible extending down over the tip of the up-curved lower mandible. Feet zygodactyl. Most brightly coloured; often gregarious: highly vocal, About 75 genera, 310–320 species; length 8 to 100 cm; found in the tropics and subtropics of the world and the temperate regions of the Southern Hemisphere.

Family Psittacidae

Characteristics of the order.

Subfamily Nestorinae (kea and kaka)

Bill rather long, narrow, less curved than in other parrots. Plumage greenish-brown. Two species: kea (*Nestor notabilis*) and kaka (N. meridionalis); length about 45 cm; New Zealand and adjacent islands.

Subfamily Psittrichasinae (vulturine parrot, or bristlehead)

Bill rather elongated, but more curved than in Nestorinae; face largely devoid of feathers, those present stiff and narrow. Plumage dark brown to black, with large patches of red. One species, Psittrichas fulgidus; length about 45 cm; New Guinea and adjacent islands.

Subfamily Cacatuinae (cockatoos)

Bill strongly curved (and massive in the palm cockatoo, Probosciger aterrimus); lower mandible wider than upper. Plumage black, gray, pink, or white, sometimes tinged with yellow or pink; often a prominent erectile crest, sometimes yellow or red. Four (or 5) genera, about 16 species; length about 30–80 cm; Australo-Papuan region and Philippine Islands.

Subfamily Micropsittinae (pygmy parrots)

Tiny parrots. Tail feathers short, with stiff shafts and pointed tips. One genus, 6 species; length about 8 cm; New Guinea and adjacent islands.

Subfamily Trichoglossinae (lories and lorikeets)

Bill relatively small; tongue brush-tipped. Mostly brightly coloured and gregarious. Feed on fruit, buds, pollen, and nectar; may be important pollinators of trees. Two tribes, Psittaculirastrini and Trichoglossini, about 14 genera, 62 species; length 11–32 cm; Australo-Papua, Polynesia, Indonesia, and the Philippines.

Subfamily *Strigopinae* (owl parrot or kakapo)

Large, flightless parrot with soft greenish plumage and an owl-like facial disk. Nocturnal and vegetarian. One species, Strigops *habroptilus*; length about 60 cm; New Zealand.

Subfamily Psittacinae (all other parrots)

Range of characteristics encompasses that of order. About 52 genera, 234 species.

Tribe Platycercini (rosellas and allies). Most with long tails, many with bright plumage colours. Twelve genera, 31 species; length about 17–40 cm; Australo-Papua, especially in interior of Australia. Tribe includes budgerigar (*Melopsittacus*), grass parakeets (*Neophema*), rosellas (*Platycercus*), night parrot (Geopsittacus), and ground parrot (Pezoporus).

Tribe Loriini (wax-billed parrots). Bill typically smooth and waxlike. Many species with bright plumage, sexes often differing. About 14 genera, 47 species; length about 14–45 cm; Australia and southern Asia to Africa and Madagascar. Tribe includes racket-tailed parrots (Prioniturus), ring-necked parakeets (Psittacula), the eclectus parrot (Lorius), and lovebirds (Agapornis).

Tribe Loriculini (bat parrolets or hanging parakeets). Small, short-tailed, green parrots. Rest hanging upside down. One genus (Loriculus), 9 species; length about 12 cm; India to Philippines and New Guinea.

Tribe Psittacini (blunt-tailed parrots). Medium-sized parrots; many brightly coloured, especially in green. Many are good mimics, especially in genera Psittacus and Amazona. Twelve genera, 66 species; length about 16–65 cm; New World tropics (majority), Africa and Madagascar.

Tribe Araini (wedge-tailed parrots). Tail usually graduated (*i.e.,* centre feathers longest, side feathers progressively shorter), often long and pointed. Some, especially the macaws (Ara) and blue macaws (Anodorhynchus), brilliantly coloured in red, yellow, and green, or blue. Many species with bare skin around eyes. Thirteen genera, 81 species; length about 12–100 cm; New World tropics. Three genera collectively include 54 species: Pyrrhura and Aratinga (conures and parakeets) and Ara (macaws).

Critical **appraisal.** With the present paucity of knowledge about the evolution of the orders of birds, supposed relationships of the Psittaciformes, which are highly specialized, are purely hypothetical. Various authors have suggested, with little supporting evidence, relationships to the hawks (Falconiformes), owls (Strigiformes), game birds (Galliformes), pigeons (Columbiformes), woodpeckers (Piciformes), and cuckoos (Cuculiiormes), especially the turacos (Musophagidae).

Relationships within the parrot order have proved just as perplexing. In the 1890s British ornithologist Alfred Newton commented on the state of knowledge of parrot classification in A Dictionary of Birds by writing:

> It is a reproach to ornithologists that so little satisfactory progress has been made in this direction, and the result is all the more disheartening seeing that there is no group of exotic birds that affords equal opportunities for anatomical examination, since almost every genus extant, and more than two-thirds of the species, have within recent times been kept in confinement in one or another of our zoological gardens, and at their death have furnished subjects for dissection.

Since Newton's time at least six classifications of the order have been proposed, many of which show basic similarities, but his statement was still basically true as of the early 1970s.

BLBLIOGRAPHY

Popular accounts: J.M. FORSHAW, Australian Parrots (1969); and W.R. EASTMAN, JR. and A.C. HUNT, ***Parrots of*** Australia (1966), are large, lavish books describing and depicting in colour all Australian parrots. H. BATES and R. BUSENBARK, Parrots and Related Birds (1959); and the DUKE OF BEDFORD, Parrots and Parrot-like Birds (1954), treat all of the parrots, with particular reference to aviculture. H.S. ZIM, Parakeets (1953), is devoted to the aviculture of the budgerigar. W.C. DILGER, "The Behavior of Lovebirds," *Scient.* Am., 206:88–98 (1962), is an excellent summary of the author's research into the behaviour and evolution of *Agapornis.*

Technical accounts: H. VON BOETTICHER, Papageien (1959), is a 116-page booklet, in German, containing ten pages of introductory information, followed by a new classification based on external morphology, especially of the cere, with all species described briefly. A. NEWTON, A Dictionary of Birds (1893–96), contains a seven-page summary of parrot biology, as known at the end of the last century. J.L. PETERS, Check-List of Birds of the World, vol. 3 (1937), presents a classification of all parrots down to subspecies, with information on geographic ranges and scientific names. A. REICHENOW, *Vogelbilder* aus *fernen* Zonen-Papageien, 2nd ed. (1955), includes colour paintings and brief descriptions, in German, of almost all parrots.

(G.E.W.)

Psychiatric Treatment, Concepts of

Basic goals of treatment in psychiatry are the same as those sought from healers of all kinds since the earliest recorded time—simply, the means to relieve distress. When the results of injury or illness are felt in the body, the distress may be called pain or discomfort; or it may have less apparently palpable origins or manifestations and be known by less tangible names such as depression, anxiety, or schizophrenia. The ultimate threat posed by distress or suffering of any kind due to illness or injury is death.

In so-called primitive societies the role of the medicine man, or healer, is often identified or overlapping with that of religious leaders, such terms as witch doctor (or shaman) exemplifying this natural–supernatural duality. No comparable duality persists in modem medicine, which is nevertheless not exclusively scientific (since much remains to be explained through science). The place of psychiatry in modern medicine rests upon the assumption that the treatment of what seems to be intangible (but excluding what metaphysics and religion refer to as the spirit, or supernatural soul) can be carried out through natural means.

Psychiatry is a branch of medicine primarily concerned with disorders of thought, feeling, and behaviour. To extend this definition to efforts to solve moral and religious issues is to take psychiatry out of its depth. In just the same way, to expect psychiatry to provide the ultimate answer to the human predicament in terms of its sum of suffering is to ask a part of medicine to become a parcel of magic. The temptation to demand magic, or to deny the limitations of psychiatry, is bound up with the difficulty of assuming mind and matter as aspects of the same reality; but, unless they are so regarded, the physician cannot begin to fathom or treat them.

Primitive approaches

Modern psychiatry has discarded many of the theories of primitive medicine on the grounds that they have been tried and found wanting. Efforts to treat psychiatric symptoms by letting demons escape through holes (trephines) gouged in the skull, by seeking to restore the sufferer's wandering soul, by sorcery, or by propitiating the gods represent some of these disease theories. Overall, the evidence is that, when it is successful, much of primitive medicine as still practiced among some tribal groups owes its power to belief in magic and the normal suggestibility of people. By contrast, this article will from here on be concerned essentially with the efforts of modern medicine to mitigate distress by techniques openly taught and carefully verified by observation. It is essentially by this openness and constant exploration and validation, as distinct from mere polishing of technical skills, that modern medicine deserves its name.

In one sense all illness or injury, indeed suffering of any kind, involves disorder of feeling; thus, the primary concerns of psychiatry with disorders of feeling, thought, and behaviour are vital to a proper understanding of medicine and surgery in general. But the special province of psychiatry is the understanding of those disorders of subjective (inner, private) experience or of objective (observable, public) behaviour that are themselves a cause of disability. A great deal more has been written about the nature of modern psychiatry. Suffice it to observe that the discipline remains subject to misunderstanding, even among those who consider themselves reasonably well educated; hopefully, a consideration of the concepts of psychiatric treatment will serve to promote fuller understanding.

For as long as psychiatric disturbance has been recognized, it has been treated either by attempts at transformation of the more or less passive sufferer or by efforts to communicate with him.

Treatment by transformation. With a history as long as that of any form of medical treatment, transformation has the underlying aim of changing the sufferer by some means that either does not depend at all or depends only in the most concrete and limited sense upon his own cooperation, understanding, and interest. Under this heading come all the various forms of physical treatment that have been employed to relieve or cure psychiatric disorders. Such techniques often are identified as somatotherapy (that is, bodily treatment).

The history of psychiatry repeatedly records the past practice of inflicting some kind of physical unpleasantness or pain, almost as though the person had to be made to expiate the crime or sin of being mentally distressed. This shadow lingers in occasional theories offered to account for the effectiveness of such forms of physical treatment as the use of drugs, electric currents to induce convulsions, and brain surgery. But none of these methods has anything in common with the whips and chains to which disturbed people were subjected in former times, nor is the production of fear any part of the rationale of their use. There are instances of behavioral treatment in which the deliberate production of discomfort, with the individual's full knowledge and consent, leads to successful retraining, but terror plays no part in the experience. Yet, with a history of almost 2,000 years of continuous violence in the treatment of the mentally disturbed, it is not surprising that even today some refer to such methods collectively as "shock" treatments, implying an emotionally shocking experience rather than any beneficial physiological changes in the body.

Electric- shock treatment

Electric-shock treatment was so named originally by theorists who assumed that the passage of electricity through the brain produces emotional shock or dread. Yet modern treatment of this kind, successfully carried out under general anesthesia, is neither shocking nor unpleasant. Once widely used, insulin shock is a technical term covering the physiological response to large doses of insulin, also being produced during a period when the patient was unconscious and knew nothing about it. Indeed, a major aspect of the doctor's task in discussing these forms of treatment with a patient is to disabuse him of the idea that fear or pain or horror play any part in them whatsoever.

Treatment by communication. Covering all forms of exchange of ideas, discussion, reasoning, and emotional response, treatment by communication is an effort to make the individual's world comprehensible to him, even to help him learn to see it with new insight, and to modify his disturbed behaviour along lines governed by his changed understanding and by his increased confidence. Conimunication is the basis of all varieties of psychotherapy, whether they rest on psychodynamic (insightful) changes or on behavioral retraining. The term psychotherapy means literally the treatment of the mind; by such a definition, however, it might even be applied to surgical methods of treatment for mental illness. Nevertheless, use of the term in practice is restricted to those methods that rely for their effect upon exchange between patient and doctor through sensory avenues of communication.

BIOLOGICAL APPROACHES (SOMATOTHERAPY, TREATMENT BY DIRECT PHYSIOLOGICAL METHODS)

General methods of bodily treatment in psychiatry include rest, diet, and protective, reassuring hospital observation leading to whatever symptomatic measures seem likely to be helpful (*e.g.*, warm, relaxing baths to relieve symptoms of tension). More specific techniques depend on drugs (chemotherapy), other neurophysiological approaches, and, still occasionally, neurosurgery.

Chemotherapeutic, or pharmacological, treatment. *Sedation.* The use of sedative drugs by day has the aim of reducing subjective distress (*e.g.*, anxiety and objectively disturbed behaviour); by night it can combat insomnia, an otherwise constant and distressing feature of many forms of psychiatric disorder. Generally speaking, a class of drugs called barbiturates (*e.g.*, phenobarbital) still includes the most efficient sedatives; when properly used, some are possibly the safest and least complicated to use. Popular apprehension and medical division of opinion about them are based largely on their misuse, as when they are taken in excessive quantities and without sufficient medical supervision (see also DRUG PROBLEMS). No effective drug, however, is free of harmful effects if misused; under proper control, barbiturates continue to be widely used when sedation is required.

Chloral hydrate

As a possible alternative to barbiturates, a drug called chloral hydrate is a safe and reliable sedative. When taken by some people, however, it produces irritation of the lining of the stomach (gastritis). As a precaution, it should be taken with some palatable vehicle such as fruit juice or milk or in dilution with sugar and water.

The regularity with which overdoses of sedatives are used for suicidal attempts has prompted a number of drug manufacturers to produce combinations, such as barbiturates mixed with emetics (to induce vomiting when overdosed) or with antidotes (such as Megimide).

The day of bromides and paraldehyde for sedation in psychiatry is past. Bromide preparations have cumulative toxic effects and readily lead to confusion and disturbance of sleep rather than to its improvement. Paraldehyde, while less pernicious than the bromides, is a malodorous and nauseating drink; in intramuscular injection it carries a significant risk of abscess formation.

Sedatives are sometimes employed to maintain continuous sleep for up to 12 hours a day for two to five days or longer. This special technique occasionally is used in the treatment of severe anxiety and tension, acutely agitated chases of depression, and other states of abnormal excitement. Prolonged sedation of this sort carries considerable risks of secondary complication unless nursing care is extremely efficient and physicians who have employed the method are readily available (see also SEDATIVE–HYPNOTIC DRUGS).

Tranquillization. Tranquillizing drugs act primarily to reduce anxiety and tension, as opposed to the major sleep-inducing effects of the sedatives. The great value of a tranquillizer is to produce calmness and relaxation while maintaining a degree of alertness during waking hours. Thus, such drugs are widely used to control symptoms of anxiety that otherwise might hamper an individual in his daily work or other activities. These drugs can also relieve some of the symptoms of tension and agitation associated with depression or obsessive-compulsive disorder (see PSYCHONEUROSES). There is a wide range of them in current clinical use, of which three groups are central examples: (1) meprobamate, (2) chlordiazepoxide and diazepam, and (3) phenothiazine compounds (see also TRANQUILLIZER). In addition to providing a basic chemical structure from which many other psychotropic drugs (those acting on mental function) are derived, phenothiazines are of particular value in the treatment of many severe psychiatric disorders that have psychotic characteristics (see below; see also PSYCHOSES).

Types of tranquillizers in clinical use

Stimulation. Stimulant drugs, besides combatting sluggishness, have rapid symptomatic antidepressant action; they may be distinguished from a newer group of specific antidepressant compounds (see below) that tend to reverse depressive disorder without producing immediate relief. Symptomatic antidepressants include a class of chemical compounds called amphetamines, which have fallen into a phase of distinct unpopularity, owing to their undoubted risk of addiction and the ease with which they are misused by potential addicts. This group includes amphetamine, dextroamphetamine, methamphetamine, and other derivatives of dl-amphetamine. While their use as symptomatic antidepressants has been largely discarded, they occasionally may be prescribed for varieties of epilepsy in which abrupt fits of compulsive sleep are prominent features. Another amphetamine-related stimulant that may still be used, particularly in the treatment of old people with chronic lethargy, is methylphenidate. Although it has relatively fewer disadvantages in terms of addiction or misuse, like all other psychotropic drugs it should never be entrusted to the unsupervised self-administration of the patient. Careful monitoring and supervision are indispensable to its proper use (see also STIMULANT).

Antidepressants. Various newer, nonstimulant preparations have been found to produce specific reversal of depressive symptoms in about 60 percent of patients treated. These include so-called tricyclic compounds, such as amitriptyline and imipramine, and another group of drugs classed as monoamine oxidase (MAO) inhibitors, such as nialamide. The exact mode of action of these preparations is not yet entirely understood, but it seems probable that they enhance or prolong the action of substances normally produced in the body, such as adrenaline and serotonin (see also DRUG AND DRUG ACTION).

As a general rule the tricyclic antidepressants are safer to use but less spectacular in their effect than are the MAO inhibitors; the latter act powerfully on a number of normal physiological systems in the body, affecting many amines (protein constituents) in the metabolic cycle. Careful restrictions therefore are placed upon the diet and intake of any other drug by patients who are using monoamine oxidase inhibitors.

Restrictions with the use of MAO inhibitors

Anti-excitement and *antihallucinatory* preparations. Haloperidol is a relatively new preparation that has proved valuable in controlling excitement (mania or less intense hypomania). Abnormal excitement and overconfidence also have been controlled with a simple chemical compound (lithium carbonate) in appropriate doses monitored by weekly blood tests.

The specific roles within the body of magnesium compounds other than haloperidol are recognized in the treatment of toxic confusional states and other forms of delirium. The exhausted brain is frequently depleted of vitamins B and C, and, by saturating the circulation with them through injection, as well as by restoring blood fluids and salts, immediate benefit can be obtained for the delirious patient.

Many phenothiazine derivatives, in addition to their action as tranquillizers, share varying potentials for reducing the general level of excitement, disorientation, confusion, and thought disorder. Some of these drugs appear to have a specifically tranquillizing and antihallucinatory effect in schizophrenia, making communication possible with otherwise inaccessible patients; they also may be combined profitably with electrical treatment (see also TRANQUILLIZER).

Antirepressive ("mind-releasing") drugs. Such drugs as mescaline, cannabis (*i.e.*, marijuana), and lysergic acid diethylamide (LSD) have not fulfilled hopes once placed in them for psychiatric treatment, and they would appear to have no supportable medical uses. The hypothesis was that their action on the brain might facilitate nerve pathways and thereby effect the release of partly forgotten (repressed) memories or the loosening of obsessive-compulsive rituals. This hypothesis has not proved valid; disturbing, long-term side effects and undesirable direct consequences (*e.g.*, so-called bad trips) of some of these drugs have contraindicated their use in psychiatric treatment. Immediate release and subsequent exploration of emotionally charged memories remain feasible by the use of injectable barbiturates of proved safety. This overall process of explosive recall and subsequent emotional disinhibition is called abreaction and can form part of psychotherapy in appropriate cases (see also HALLUCINOGEN).

Physiological methods. Physiological techniques of psychiatric treatment (that is, those that produce a suspension of consciousness) demand for their administration the attention of a team of professional experts. Special physiological methods have radically altered the outlook for some types of mental disorder previously beyond the power of psychological medicine to relieve, cure, or control. These empirically useful techniques share a capacity to change the function of the nervous system and ductless glands by the use of electricity, substances such as insulin, and surgery.

Electroconvulsive therapy. The history of electroconvulsive treatment can be said to have begun with the discovery that the function of the brain can be altered immediately or over a period of time by the effects of electrical energy introduced from outside. Initially, hypotheses for the immediate use of electricity to induce seizures were based on evidence that some convulsion-producing drugs, such as pentylenetetrazol, also had beneficial psychiatric effects. Subsequent years of research and refinement have shown that, even without convulsions, induced alterations in the electrical and biochemical activity of the brain over a limited period of time can relieve symptoms of depression, delirium, stupor, and schizophrenia. Limits also have been established within which electrical current can be safely used without damage to the brain or harm to the patient.

Induced seizures

Electroconvulsive therapy (ECT) depends on the passage through the brain of a small electrical current, on the order of 200 to 500 milliamperes, for less than one second at a level not exceeding 150 volts. The current is now usually administered to only one side of the brain

—the nondominant side. Typically, this is the right half of the brain, most people showing signs of dominance (*e.g.*, location of speech centres) in the left brain hemisphere. The advantage of such unilateral ECT is simply that it diminishes undesirable side effects formerly frequent when the bilateral form of this treatment was common. Included among these side effects are loss of immediate memory, an element of confusion, and a sense of dislocation in time and awareness; all tend to be temporary, but they are relatively disagreeable.

Although bilateral ECT is still used occasionally, future methods of electrical treatment will probably be increasingly confined to the nondominant hemisphere of the brain, and the advantages to be sought from it will be increased and the side effects diminished. The appropriate indications and detailed techniques for this treatment are still a matter for continuing relearch. Usually administered at a frequency of two to four times per week, the treatment is given while the patient is under a combination of anesthesia and muscle relaxant so that no convulsions ensue.

The way in which this remarkable treatment works is still incompletely understood, but recent research into the electrophysiological and biochemical changes accompanying it suggests that its effects involve several parts of the brain (*e.g.*, cortex, midbrain, hypothalamus), the pituitary gland, and other autonomic parts of the nervous system (see also NERVOUS SYSTEM, HUMAN; ENDOCRINE SYSTEM, HUMAN). It can further be postulated that the convulsion originally associated with the treatment was always merely incidental, a by-product of the stimulation required to trigger off other, more directly therapeutic, physical processes. Although it is clear that spontaneous fits among sufferers of epilepsy produce no comparable benefit, the administration of ECT to carefully selected patients can yield spectacular and highly satisfactory results.

Whatever the indication for the treatment, it is routinely combined with direct explanation and communication with the patient. As improvement progresses, the degree to which personal problems remain to be dealt with will vary among individual patients and will affect the nature of any psychotherapy then undertaken.

Electrically induced sleep

Electronarcosis. This form of treatment is given more frequently in Britain and in the Soviet Union than anywhere else. It was pioneered as much in Australia as in the United Kingdom and has rarely found much clinical support in the United States. Essentially, the immediate aim in electronarcosis is to produce an artificial state of electrically induced sleep, which may be initiated under an anesthetic. Such sleep has been subsequently maintained for periods varying from a few minutes up to 15 to 20 minutes or longer, depending on previous experience. The eventual routine use of this induced-sleep therapy will rest on the outcome of research into its effectiveness and its appropriate indications.

Insulin treatment. Injections of insulin lower the level of sugar in the blood, the low-level state being called hypoglycemia. In so-called deep-insulin treatment, sugar is depleted to the point that the patient loses consciousness. This treatment, with its hypoglycemic coma (reversed by injections of sugar), played a historical and epochal part in transforming the outlook for sufferers of some varieties of mental disorder (particularly schizophrenia) who hitherto could be offered, at best, only protracted custodial care; it is now, however, obsolete in the treatment of schizophrenia.

Modified insulin treatment (in which the induced hypoglycemia is mild enough to avoid deep coma) continues to be a useful, practical method of treatment for conditions characterized by chronic anxiety, anorexia (loss of appetite), loss of weight, profound tension, and exhaustion or by restless, irritable resistance to treatment and management of any kind. Modified insulin treatment requires that the patient be nursed in bed, at least for the first half of every day; this provides a foundation for a regular regime of nursing care and attention, with specific concern for quiet, rest, and progress. As such it has undoubted benefit in certain cases by reducing physiological signs of disturbances and by improving the patient's sense of well-being.

In the modified method, the patient begins with a minimal test dosage of insulin every morning; the dose then is increased to an arbitrary maximum that usually does not exceed *50* units of soluble insulin. A balancing meal (containing sugar or other carbohydrate) is given within an hour of the injection, and such side effects as drowsiness, occasional faintness, and some confusion are nursed symptomatically. A favourable response is indicated by reappearance of normal appetite, gain in weight, and, most of all, a calming effect and an ability to accept other forms of treatment (*e.g.*, psychotherapy) and improvement as reasonable goals.

Neurosurgery. Brain surgery as part of the modern treatment of mental illness often is said to have begun with the speculation and experimental work of a brilliant Portuguese neurologist, António Egas Moniz. Moniz, however, had an international tradition of clinical observation and perceptive hypothesis in neurophysiology on which to build. He was familiar with the observations of the great English neurologist John Hughlings Jackson upon localization of function in the brain; she pioneer work of the noted American surgeon Harvey Cushing in brain surgery and its effects upon brain function; and German theories of the role of the frontal lobes of the brain. He had also considered the 19th-century medical history of a patient in the United States who showed major personality changes after his frontal lobes had been virtually destroyed by a crowbar that had been driven upward through the front of his skull and emerged from the top of his head (but without killing him).

There were also reports on the behavioral effects of experimental work with the living brain in otherwise intact, intelligent chimpanzees whose frontal lobes were removed or severed (leucotomized) from the rest of the brain. Moniz accepted the responsibility of transferring the operation to men when he was confronted with sufferers of apparently incurable mental disturbance, in some of whom emotional tension and depression were crippling. The operation was called prefrontal leucotomy (or lobotomy) because it depended for its effect upon severing the connecting nerve fibres (white matter) between the frontal lobes and the rest of the brain. These white fibres run beneath the surface of the brain (gray matter).

Prefrontal lobotomy

There can be no doubt that, in both conception and performance, the operation was in its earliest day a remarkable combination of technical brilliance and concentual crudity. Prefrontal lobotomy proved remarkably successful in certain carefully selected types of illness, but it was truly disastrous in others. Study of the first series of Moniz' cases demonstrated that the operation undoubtedly produced changes in the mental state of the patients. These differed, however, from the theoretical expectations of Moniz; for while delusions and hallucinations were unaffected, what was changed was the attitude displayed by the patients toward their symptoms and their overall state. Like lobotomized chimpanzees, they had become more placid, tranquil, and phlegmatic—too often, excessively so.

Following this initial series, clinical observation and research were stimulated, and the operation proved more enduring than the opposition that it first had provoked. The thesis that eventually emerged was that a great deal of emotional tension and the patient's conscious control over and preoccupation with behaviour that might lead to such tension are in some way of neurological origin. Specifically, the symptoms seem connected with the function of nervous pathways connecting the frontal lobes with the rest of the brain (particularly with the thalamus, a mass of nervous tissues about half the size of a golf ball situated in the lower core of the blain). The role of the thalamus appears to be that of a central reception station for incoming sensory messages of all kinds, and its special connections with the frontal lobes and with an associated structure (the hypothalamus) appear in turn to be concerned with the elaboration of conscious awareness of, and emotional response to, stimuli. In this sense the frontal connections to other parts within the brain seem to play a fundamental role in shaping the way the patient feels about present reality and past experience.

This is the basis for the subsequent extension of the use of the operation in certain cases of otherwise intractable pain, as in cancers that are not accessible to surgery. Whether the origin of the suffering seems to be primarily mental or physical, the relief appears to depend upon the patient's altered attitude to his pains, rather than upon their total disappearance. The general indication for these operations, however undertaken (whether by open surgery or by the introduction of radioactive implants in the body), can best be summed up as otherwise intractable distress in which tension and emotional stress are paramount and in which apathy and withdrawal from the environment are not major features.

The human frontal lobes (and the rest of the brain cortex) are substantially more highly developed than are those of most other animals. It seems probable that, as these structures have evolved and enlarged, so has their capacity for inhibiting more primitive types of activity in the nervous system. It has been pointed out that the great development of intelligence and flexible behaviour in man may be linked to a superior ability to inhibit immediate reflex reactions to disagreeable experience. Planning and thought (associated with the cortex) are involved in putting off actions until an opportune moment and in the deferment of short-term satisfactions for the sake of long-term goals.

Reduction of neural inhibition

Thus, one fundamental principle seems to emerge: lobotomy in psychiatry seems to be a way of reducing the abnormal or exaggerated effects of inhibition among different brain parts. This kind of brain surgery shows no promise of ever adding anything; rather, it is by subtracting or reducing brain functions that lobotomy relieves the patient's distress (see also EMOTION: The ***frontal lobes and emotion).***

TREATMENT BY COMMUNICATION (PSYCHOTHERAPY)

The theories underlying psychiatric treatment by communication are broadly classifiable as either psychodynamic or behavioral (see also PERSONALITY, THEORIES OF).

The concept of psychodynamics in psychotherapy owes its existence to the hypothetical postulate that each person has an unconscious mind. Versions of this theory have been advanced throughout history (*e.g.*, by Socrates in ancient Greece); but they were later considerably developed and expanded by Sigmund Freud. Subsequent streams of dynamic therapy evolved from and eventually separated from the Freudian school of psychoanalysis, under such names as Jungian, Adlerian, and Rogerian (nondirective) therapy; other dynamic variations are called Gestalt therapy and Existential analysis (originally called Daseinsanalyse by its founder, Binswanger).

By contrast, the fundamental concept underlying behavioral psychotherapy is that of the conditioned reflex. Ivan Pavlov, whose work with dogs pioneered the concept of the conditioned reflex (see LEARNING THEORIES), strenuously maintained that his was physiological rather than psychological research. Nevertheless, this mechanistic, neurological concept of learning (reflexology) as a basis for mental disturbance has become associated with the names of many psychologists. It would be generally agreed upon that its first formulation came from the American psychologist John Broadus Watson, who first wrote of it in the ***Encyclopædia Britannica*** in ***1913.***

All other systematic forms of psychotherapy are essentially derived from one of these two mainstreams; or they represent breakaway movements that tend to become mystical, religious, or heavily authoritarian in concept and so resist open, scientific scrutiny. Until they permit objective, experimental evaluation, such breakaways remain beyond the scope of scientific psychiatry.

Psychodynamic treatment. The principles of psychodynamics are ultimately mechanistic, using as their analogy the idea of physical forces in motion. Their hypothetical source can be traced to three basic theoretical principles or assertions:

1. Human behaviour is prompted chiefly by emotional considerations, but insight and self-understanding are necessary to modify and control such behaviour and its underlying aims.

2. A significant proportion of human emotion, together with the action to which it leads, is not normally accessible to one's personal awareness (or introspection: literally, looking into one's self), being rooted in the mind beneath the level of consciousness.

3. It follows logically (and is supported by experimental and clinical work) that any process that makes available to individual consciousness the true significance of emotional conflicts and tensions hitherto repressed (in the unconscious mind) will thereby produce both heightened awareness and increased stability and emotional control.

The application of these principles in practice hopefully can lead not only to improved psychiatric health in its widest sense but also to a more mature and developed personality. Classical dynamic psychotherapy is a relatively intensive type of talking treatment aimed at providing the person with insight into his conscious and unconscious life and into the deeper relationship of psychiatric symptoms to emotional conflict; it is often uncomfortable for the patient himself, at least in the early stages. Yet it promises the considerable advantage of enabling him to achieve an awareness of himself that may not only help to arm him against present difficulties but also make it easier for him to control at least a part of his fate in the future. The effectiveness of dynamic psychotherapy depends ultimately upon the patient's acceptance of a significant share of the responsibility for the success of the treatment, by experiencing the disagreeable effect of examining his emotionally tender spots. The process also imposes a considerable strain upon the therapist, who is not supposed to relieve his own anxiety about the patient by taking responsibility for the patient's difficulties away from him.

Factors in effectiveness

The original, classical, and most elaborate form of dynamic treatment was that developed by Freud and called by him psychoanalysis. Analytical psychotherapy, developed by the Swiss psychiatrist Carl Gustav Jung, and individual psychotherapy, by the Austrian psychiatrist Alfred Adler, are two further specialized psychotherapeutic techniques worked out by these former pupils of Freud who later diverged from him to create their own approaches to the goal of self-awareness. Extremes of subsequent psychodynamic theories are exemplified by the nondirective therapy of Carl Rogers, at one pole, and Existential analysis, or Daseinsanalyse, as expounded originally by Binswanger, at the other (see below).

The advantage claimed exclusively for these traditional methods of analytic treatment by their proponents is that when completed they accomplish radical and exhaustive personality change. What has been spent in time is held to be repaid in depth and degree of awareness and consequent personal stability. In practice a full, classical analysis demands not less than three to five interviews a week for anything from two to four years, excluding only the inevitable breaks caused by absence of doctor or patient. Any procedure that demands so much expenditure of skilled time is inevitably costly. Moreover, such procedures are suitable only for patients whose intelligence, determination, and resources of time or money are sufficient to enable them to complete and profit from so intensive a course of treatment. Indeed, on these grounds alone, failures in analytic treatment are quite common.

Problems of older analytic methods

These older analytic methods of psychotherapy are thus highly selective and self-limiting forms of treatment; their practice, moreover, is entrusted only to those few therapists whose training has fulfilled particularly rigorous conditions. But, despite these limitations, they are important procedures. Their importance today derives less from their practical results (gained in individually suitable patients, who probably make up not more than 5 percent of all patients with psychiatric problems of any kind) than from the body of understanding that these methods have contributed to what is called psychopathology (the study of psychiatric symptoms). Most important of all are modifications of these deep, intensive analytic techniques that have led to the evolution of various briefer varieties of dynamic psychotherapy devoted to the attainment of insight within a matter of weeks or months.

Psychoanalysis: the rnethod* of *Freud. The initial interviews in psychoanalysis follow the same general pattern as those of any other psychotherapeutic procedure. As far as possible, the history of the patient's life is obtained and noted in an orderly and comprehensible way. Thereafter it is the task of the patient to produce, out loud and absolutely without suppression or censorship of any kind, all his thoughts and feelings about whatever is uppermost in his mind. Such a talking procedure is called free association. This may sound difficult, but in practice it is even more challenging than it sounds, at least in the early stages of treatment. To speak of one's innermost (often socially unacceptable) thoughts can be most trying after years of practice in selecting what one will say to others.

In response to the hitherto inconceivable opportunity of acknowledgment and expression without fear of social disapproval, many of the ideas that rise into the patient's mind are thrust down almost before he has had time to become aware of them; such readily available ideas are said to be preconscious rather than truly unconscious. More difficult still is the patient's eventual acknowledgment of underlying unconscious sources of such preconscious ideas, which can only be achieved with time, practice, and, above all, a confidence that at the outset no patient can be expected to have.

The psychoanalytic procedure

The one essential rule of psychoanalytic procedure, that nothing shall be suppressed and nothing selected, demands the adoption of every possible measure that helps the patient to do his part of the work. It is for this reason alone that the patient is encouraged to lie down, to relax completely, to look at nothing more exciting than the ceiling, and to talk without the visible presence of another person to distract him. The analyst sits out of the range of the patient's vision and, while still able to do the therapist's part of the job, may have no cause for intervention of any kind. All the time the analyst is recording, sorting, and studying what the patient says, bearing in mind its possible interpretations and its correlation with what has been said before. As the analysis proceeds, he strives to accumulate deeper sources of understanding of the patient, but he does not obtrude into the process unless, without his help, the patient cannot go further.

The study of dreams forms an important part of psychoanalysis, being based upon theory as well as experimental evidence that dreams have a latent (hidden or disguised) content different from the manifest (literal) content by which they are remembered. This conclusion emerges from studies in which manifest dream content is subjected to analysis of the dreamer's free association about it. Thus, the ogre in one's dream may be found to be a disguised representation of a hostile, punishing parent. The production of the manifest content from the latent content is an unconscious process that Freud called the dream work, and it is the reversal of this process that constitutes the analysis of the dream.

Many of the experiences the patient resists remembering are those of personal relationships (in the recent or more remote past) that have been connected with intense and conflicting emotion and that have never been finally resolved or settled. The release of such emotion in the course of treatment, before its unconscious source has reached awareness or been understood and accepted, is in itself intensely disturbing. The patient often seeks to rationalize or to project (attribute to something other than himself) the emotion he feels. Under the conditions of treatment, the person who almost invariably is selected for such rationalization or projection is the analyst; that is, the patient is likely to blame his own emotional distress on the analyst.

In this way, during the course of treatment the patient comes to feel love and hatred, dependence and rebellion, and rivalry or rejection toward the analyst. These are the same attitudes he has felt but never fully acknowledged for other people in his life whose earlier impact has been inescapably close; they may have been his parents, his first love, or the friends, enemies, heroes, and villains of his childhood. The patient's uncritical and barely understood shift of earlier emotional attitudes to the therapist is called the transference. Control of this transference and regulation of its depth and intensity are to a great extent in the hands of the analyst, who strives to have its underlying interpretation available for use as the occasion demands. To interpret or interfere with the transference too early is to deprive it of the strength necessary to enable the patient to carry on with his already difficult contribution to the treatment. Any tendency by the analyst to escape prematurely from his obligations in transference may spring from his own anxiety or insecurity, especially if this has not been adequately dealt with in the course of his training. To interpret transference inadequately or too late or to fail to deal with the transference situation at all is to risk the development of a patient who is heavily dependent on the therapist. This may be a more serious complication than the symptoms that first brought him to seek treatment; at this point the patient is said to sufferan interminable transference neurosis.

Expression of emotions during analysis: transference

The handling of the transference situation is thus of vital importance in the course of psychoanalysis or, indeed, in the course of any form or psychotherapy. It is through transference that the patient discovers the nature of his underlying feelings and then becomes able to acknowledge them. Once this has been done, he often finds himself able to regard them in a far more tolerant and dispassionate light and to be liberated from their influence upon his future behaviour.

Jungian analysis: analytical psychology. The technique of analytical psychology as formulated by Jung differs from the original Freudian technique of psychoanalysis in that far less emphasis is placed upon free association. Much more importance is attached to an analysis and interpretation of certain aspects of the patient's fantasies and dreams. These aspects are handled in a way quite different from that propounded by Freud in his approach to the analysis of dreams by association. Perhaps the essential difference is between Jung's and Freud's concepts of unconscious mental life. The Freudian view maintains that the unconscious mind contains essentially only instinctual drives and repressed conflictive complexes. The Jungian thesis, however, is that there is an enormous reservoir of shared unconscious wisdom and ancestral experience transmitted throughout the generations to the whole of humanity. This so-called collective unconscious is said to permeate each individual unconscious mind and to enter consciousness only in symbolic form to influence thought and behaviour in powerful but indirect ways. The symbolic forms taken by these powerful archaic mental processes, regarded as common to the entire human race, are called archetypes; and it is in the archetypal nature of myths, stories, and dreams that the Jungian analyst seeks clues to his patient's problems and their interpretation.

Adlerian analysis: individual psychology. Adler's special contribution to individual psychotherapy was to stress the importance of the drive toward power, which lies at the heart of so much human endeavour.

He observed that for many people the achievement of power, status, and prestige was at least as important as the search for sexual gratification and could produce great impact upon the individual's life and emotional development. It was Adler who was responsible for the now-common phrase "inferiority complex," meaning not a conscious feeling of inferiority but, rather, an unconscious constellation of ideas of personal worthlessness. These ideas generate an intolerable sense of insecurity and inadequacy, the origins of which the sufferer is largely unaware. The effect of such unconscious feelings of inferiority upon behaviour is (by compensation) to produce a somewhat assertive attitude and to drive the individual into challenging situations within which he strives to prove his deeper misgivings about himself to be false.

The "inferiority complex"

Adlerian analysis concentrates attention upon the patient's idealized, partly conscious concept of himself and upon the goals that he has set himself. The therapist seeks to relate self-concept and goals to the patient's actual personality and to the practical circamstances, opportunities, and limitations of his life. In the inevitable conflicts and frustrations implicit in the gulf between aspiration and achievement, Adlerians perceive the foundations either of pathological emotion and behaviour or of renewed and constructive endeavour. Through the therapist's and patient's interpretations and insight, Adlerian analysis aims at enabling the individuai to become aware of the true springs of his own needs, drives, and impulses. When these urges are compensatory, in reaction to some real or imagined weakness in himself, the patient is helped to evaluate their implications consciously and, by discriminating between feasible and impractical goals, to free himself to pursue the one and discard the other.

The content of an Adlerian analysis is therefore essentially concerned with the relationship of the patient's personal aspirations and goals of achievement, both conscious and unconscious, to his social setting and total life situation. Less emphasis is placed on the more instinctual need for sexual fulfillment, which Freud saw as the ultimate driving force in human experience, or on the integration of the personality, which remains the Jungian concept of human fulfillment.

Rogers and Binswanger. In 1942 Carl R. Rogers in the United States and Ludwig Binswanger in Switzerland published two fuither (and in many ways polar opposite) contributions to psychodynamic psychotherapy: Rogerian, or nondirective psychotherapeutic counselling, and Daseinsanalyse, or existential analysis. The former relied entirely upon the inner resources of "the client" for its outcome. In Rogers' own words,

> Effective counselling consists of a definitely structured permissive relationship which allows the client to gain an understanding of himself to a degree which enables him to take positive steps in the light of his new orientation.

Binswanger's thesis was essentially inspired by Heidegger's Daseinsanalytik, a philosophical approach to the structure of human existence in general. Binswanger published it as a 726-page book, *Grundformen und Erkenntnis menschlichen Daseins* ("Basic Forms and Knowledge of Human Existence"), in essence an attempt to analyze every aspect of being-in-the-world for every human individual, embracing both shared or outwardly real experiences and the hallucinatory experiences of psychotic patients.

Gestalt therapy, a further development, can be seen in one way as an attempt to combine the best of all these various approaches by at first comprehending and then as far as possible reconstructing the whole of the subject's total inner and environmental predicament.

Brief methods of interpretive (or insight) therapy. The short summary, given above, of the analytical techniques does not present all details of the psychodynamic theories underlying them. But the highly complicated nature of this kind of treatment should be clear. Thus, it is most misleading to apply the term psychoanalysis or even analysis to psychiatric interviews that may have as their goals nothing more profound than the clarification of some immediate personal problem or domestic situation. Indeed, the vast majority of psychiatric interviews are of this relatively limited nature, and treatment by psychoanalysis, Jnngian analysis, Adlerian analysis, or any of the other varieties is suitable for only a small proportion of patients who may require psychotherapy in some form or another. For the remaining bulk of such patients, however, the practical possibilities of brief methods of dynamic treatment are most encouraging and appropriate.

Elements of psycho-analysis

All brief methods of insight therapy depend for their brevity upon reduction of elements that normally comprise a considerable part of psychoanalytic procedure and absorb much of the time involved. These elements include the repetition involved in working through problems against the patient's resistance, as heavy reliance is placed upon the patient's free associations to bring these problems back into the field of exploration every time they are evaded or shelved. In addition, much analytic time and effort is consumed in dealing with comparatively irrelevant patient comments until he finally begins to make associations that seem unquestionably important but that were formerly too emotionally charged for him to handle.

Brief psychotherapy selects those aspects of the patient's life that are most relevant to the specific problems he faces and the symptoms he displays, and thereafter it centres the entire procedure upon dealing with them.

All such methods have in common the aim of paring down the time and effort of analytical procedures by dealing almost exclusively with immediate or feasibly alterable sources of patient complaint. The criticism is sometimes offered that such brief methods tend to patch rather than to re-create analytically the personality of the patient. One practical rebuttal is that the limited resources of medicine as a whole inexorably compel the practitioner to do just so much and no more for as many people as possible. Moreover, the bright promise implicit in full-scale orthodox analysis cannot claim support by observable results in any but a minute pioportion even of those few patients for whom it is deemed practicable.

In practice, a useful framework for brief psychotherapy can be constructed by planning eight to 12 interviews at weekiy intervals, each of 30 to 45 minutes' duration, for which are set the following targets:

1. Establishment of rapport (effective doctor–patient relationship of trust and confidence) and collection of an adequate patient history (first two interviews);
2. Selection of goals; formulation of overall plan of treatment; and exploration of goals (subsequent three or four interviews);
3. Interpretation of information developed (begun at appropriate stages of exploratory interviews but completed within context of overall treatment); integration of these interpretations with the patient's own concept of his life situation and its implications; recognition and acceptance by the patient; support and reassurance by the doctor (final three or four interviews).

The phases, methods, and special techniques involved are indicated in the Table.

Outline of Brief Insight Therapy

phase of treatment	methods		special techniques
	patient	doctor	
Ventilation	describes symptoms; discharges emotions	listens; accepts; encourages	abreactive
Exploration	recalls traumas; discovers connections	questions; interprets	analytic
Guidance	listens, comments	reassures; explains	counselling; suggestion; social aid

As a general rule, it is reasonable to devote about 70 percent of the time to ventilation, 20 percent to exploration. and 10 percent to guidance in the total course of treatment. The initial interviews should consist essentially in the acceptance by the doctor of the patient's need to communicate and to be understood. In the early interviews it is not what the doctor tells the patient but what the doctor permits the patient to tell him that is likely to be decisive in the success or failure of this kind of brief treatment.

Supportive and interpretive psychotherapy compared

Supportive psychotherapy. When patients are encountered for whom a more ambitious approach is judged impracticable or unproductive, supportive psychotherapy may be employed. Supportive psychotherapy differs from interpretive psychotherapy only in that its goals are usually more modest and less ambitious; often, however, it is more difficult in execution. Supportive psychotherapy plays a great part in the relatively informal management of nonhospitalized patients who might otherwise receive no psychiatric help at all from their doctor. It can include direct, simple, and sympathetic advice and sheer reassurance and encouragement. This kind of supportive treatment may be combined with practical intervention in the social circumstances of the patient. Emotional complications of his homelife or his work, for example, may be constructively modified by making personal contact with the family or employer, usually through a trained psychiatric social worker under the supervision of the psychiatrist concerned.

The foundation for supportive psychotherapy of this kind, which may be combined with appropriate medicinal symptomatic treatment, is a thoroughgoing history of the patient. After such a history has been gathered (as early as possible), the doctor should be able to gain an understanding of the patient in some ways more objective and comprehensive than that perhaps ever gained by any other single person, including the patient himself. At this early stage, a comprehensive picture of the patient and the general direction and significance of the tendencies revealed in his life should help the doctor form a sound idea of the genesis of the patient's symptoms and the ways in which the stresses to which he is subjected may be lightened or relieved. Thereafter, such advice or explanation as a doctor has to offer can be based not upon his own personal feelings but on the objective desirability of any particular solution as revealed by the history. His advice will depend still less upon a projection on his part of what he would do if he were in the patient's shoes. Thus, a carefully reconstructed case history is a safeguard against the shortcomings from which much well-intentioned lay advice so often suffers.

Psvchotherapy of this limited but invaluable supportive kind is within the capacity of every practicing psychiatrist and other trained physicians. The otherwise unavailable help provided by expenditure of a limited amount of time in this way is not only humane, but economical. By the skilled use of a few hours spread out over weeks or months, the trained physician may avert the chronic invalidism and demoralization of the neurotic patient. He may also spare himself the bitterness and frustration that assail a medical practitioner who is continually confronted with psychiatric aspects of human suffering but has never acquired the interest or understanding necessary to deal with it effectively.

Group therapy. This technique is based on the capacity of groups of individuals to recognize and then learn to understand patterns of feeling and behaviour in each other that either resemble similar feelings in themselves or evoke reactions for which they have not previously had insight. Membership in such a group can increase the individual's sense of support and his readiness to accept others and to be accepted by them, factors that form an important part of the process of treatment.

Groups of various sizes have been treated, but probably the best size is about eight to 12 people. Such people meet together with a doctor as a member but not as the leader of the group to talk about their feelings and their problems. As their confidence increases, they attempt to offer and to accept among themselves awareness, courage, and compassion and to contribute to this experience as well as to gain from it. The doctor's role in the group is largely that of observer and occasionally interpreter. His interpretations should be confined to the meaning of the emotional interaction that is taking place and to the ways in which everyone concerned can accept and learn from it. He neither seeks nor assumes formal leadership, because this would interfere with his capacity to reflect an awareness of the struggle for power within the group or the fear or anticipation of rejection by it—both among the powerful emotional responses that arise.

Psycho-drama

Other extensions of group dynamics include various specialized types of groups in which role playing or specific social support form the goals. So-called sensitivity groups have become popular in the United States, one object being to make an individual aware of differences between the image he has of himself and how he is perceived by others. Psychodrama is the name given to a type of group activity in which the group members act out roles relevant to the problems of each of them in turn. An example of specific supportive group therapy is provided by Alcoholics Anonymous (see ALCOHOL CONSUMPTION).

Behavioral therapy. The basic principle of the modern technique of behavioral therapy rests upon the assumption that many neurotic disorders (particularly states of phobia and anxiety) can be regarded as learned responses that have been built up into conditioned reflexes. The treatment is, therefore, one of retraining by deconditioning; in one variety, the conditioned response (*e.g.*, fear) is replaced by a learned pattern of inner calmness, confidence, control, and well-being. Such feelings, for example, may be induced by hypnosis or by methods of relaxation that the patient readily learns from the therapist. Thus, the sufferer from a phobia for spiders is trained to respond to them with pleasant feelings rather than with his earlier fright.

An alternative approach to behaviour modification is the deliberate conditioning of negative responses to established but unwelcome patterns of behaviour; for example, an alcoholic may be conditioned to experience nausea when he tastes whiskey. Or, the tendency of transvestites to don clothing of the opposite sex may be extinguished with mild electric shocks to the hand. This is generally termed negative conditioning or aversion therapy.

The role of hypnosis in deconditioning has been recognized to a growing degree. There are a number of techniques available for the production of profound muscular relaxation, which may lead to hypnosis of varying depth. These hypnotic techniques vary from one practitioner to another, but the essential preliminary phase is one of suggestion of relaxation, drowsiness, comfort, and well-being. This is followed by more specific suggestions of hypnotic trance phenomena. Sometimes tests of the depth of hypnosis obtained at various stages are introduced, partly to inform the hypnotist and partly to reinforce the hypnotic suggestion in the mind of the subject. Tests of this kind tend, however, to introduce a gratuitous challenge in the proceedings that may increase any conscious or unconscious apprehension on the part of the patient that hypnosis involves some kind of battle of wills between him and the hypnotist (for an evaluation of this common attitude, see HYPNOSIS).

Use of hypnosis

Experience with patients who have attempted behavioral therapy with specialists in medical hypnosis suggests that failures to pass tests can impair confidence in the hypnotist and his methods. Confidence can very often be restored, however, by an approach to relaxation during normal wakefulness in which hypnosis itself may not even be mentioned. The ultimate role of hypnosis should be as an aid rather than as a requirement for successful reconditioning in psychiatric practice.

Whenever any alterations in human relationships are brought about by communication, individually or through groups, psychodynamic or behavioral, evidence of unconscious motives and of transference can only be neglected at the peril of all concerned. No form of psychiatric treatment is free of hazard. As in all forms of medicine and, indeed, in all human transactions, the higher the level of professional competence, integrity, and compassion, the better is the outlook for individual and society alike.

BIBLIOGRAPHY. Key theoretical works concerned with concepts that underlie psychiatric treatment include D.O. HEBB, "On the Meaning of Objective Psychology," *Trans. R. Soc. Can.*, Series 3, 55:81–86 (1961); GILBERT RYLE, *The Concept of Mind* (1949); HENRI F. ELLENBERGER, *The Discovery of the Unconscious* (1970); C.G. JUNG, *Collected Papers on Analytical Psychology* (1916); ALFRED ADLER, *The Practice and Theory of Individual Psychology*, trans. by P. RADIN (1924); R.D. LAING, *The Divided Self* (1960); and D.H. MALAN, *A Study of Brief Psychotherapy* (1963). Works that deal with pharmacological or physical methods of treatment include THOMAS A. BAN, *Psychopharmacology* (1969); and WILLIAM SARGANT and ELIOT SLATER, *An Introduction to Physical Methods of Treatment in Psychiatry*, 4th ed. (1963). For discussions of the underlying psychodynamics of group interaction, see DAVID STAFFORD-CLARK, "Supportive Psychotherapy," in *Modern Trends in Psychological Medicine*, 2nd Series (1970); ERIC BERNE, *Games People Play* (1964); and G. SEABORN JONES, *Treatment or Torture: The Philosophy, Techniques and Future of Psychodynamics* (1968). Elaborations of the concept of psychiatry as a whole are found in DAVID STAFFORD-CLARK, *Psychiatry for Students*, 3rd ed. (1969); *Psychiatry Today*, 2nd ed. (1963); and *What Freud Really Said* (1965). See also JOSEPH WOLPE and A.A. LAZARUS, *Behavior Therapy Techniques* (1966).

(D.S.-C.)

Psychology

On the walls of the Temple of Delphi, built by Phoebus Apollo on the jagged cliffs of Mount Parnassus, was inscribed the most famous of all Greek precepts—"Know thyself!" Man has been preoccupied with such admonitions, and his partially successful response is reflected in the discipline of contemporary psychology.

In the process of knowing themselves, some Western men from the times of the ancient Greek philosophers have divided each person into elements of "mind" (or "soul") and "body." About 1594–96, Otho Casmannus formalized this doctrine by coining words from Greek roots: anthropoiogy (the study of man), psychology (the study of mind), and somatology (the study of body). Casmannus discussed psvchology in his work Psychologia anfhropologica, sive animae humanae doctrina . . . and somatology in his *Secunda* pars anthropologiae, *hoc* est, *Fabrica humani* corporis methodice descripta . . . . This distinction between mind and body, so apparently manifest to such writers, seems capricious and unnecessary to many modern scholars.

The term mind has been eliminated gradually from scientific scrutiny, most present-day psychologists using it only with intentional vagueness. In the 17th century, influential philosophers were emphasizing views that rejected concepts of souls exercising free will and that regarded organisms as automatic machines whose actions were fully determined by internal or external stimuli. Psychology has come to mean the study of those actions and their origins, generally defined as the science of behaviour in man and other animals.

SUBJECT MATTER OF MODERN PSYCHOLOGY

Several basic topics constitute the subject matter of modem psychology: sensation–perception, motivation, emotion, innate patterns, learning, thinking, intelligence, personality, group dynamics, and behaviour pathology.

Sensation–perception. Sensations may be understood as simple experienced correlates of physical stimulation of sensory receptors, with perceptions being defined as meaningful interpretations of sensations. Subtopics include vision (mediated by such structures as the eyes and occipital lobes of the brain), audition (involving the ears and temporal lobes of the brain), olfaction (smell) and gustation (taste), somesthesis and kinesthesis (*e.g.*, experiences of temperature, pressure, and position arising from the skin and muscles), and vestibular proprioception (sensations of position and movement arising from structures of the inner ear; see SENSORY RECEPTION, HUMAN; PERCEPTION).

Motivation. Older ideas, such as "striving faculties" (the "push" behind behaviour), yielded to the more contemporary language of motives. So-called primary motives (those necessary for the preservation of the individual) include hunger, thirst, need for sleep, and avoidance of pain. Other posited primary motives are seen to be necessary for continuing the species (*e.g.*, sex and maternal behaviour). All primary motives are defined as physiological and innate. Secondary personal or social motives (those held not to be necessary for individual or species survival) include acquisitiveness, gregariousness, aggressiveness, and achievement seeking. Many so-called motives substantially depend on social groups and concern social dominance, conformity to societal norms (fads, fashions, customs, and mores), and obedience to authority. Among human beings, secondary motives seem almost always to be acquired (usually by learning), but are believed to be usually innate among other animals. The same superficially identical secondary motive (*e.g.*, social dominance) often is held to be acquired by people but to be innate in such animals as baboons (see MOTIVATION).

Emotion. Affective or emotional processes are identified with exciting or inhibiting states generated by mechanisms within the organism and that coexist with other behaviours. Weak emotions, enduring over long time periods, have been called feelings. Emotions traditionally have been classified according to such subjective, hedonistic states as pleasant, neutral, and unpleasant. Their study includes experimental aesthetics (*e.g.*, the effects of music or odours on them), their facial and postural expression, and their physiological aspects. Emotions clearly may play a motivational role (see EMOTION).

Innate patterns. Behaviour deriving directly from biological heritage potentially is exhibited by all living beings. A relatively new term in behavioral science is ethology, a discipline (with strong roots in Europe) that emphasizes these inborn sequences.

Biological bases of behaviour

Innate patterns have been classified as a form of taxis, or a simple orientation movement such as upward righting of the body in reaction to gravity; as reflexes or brief responses of limited scope, such as pupillary change to light; and as instincts or complicated, enduring patterns—*e.g.*, nest building. Animals low on the evolutionary scale characteristically exhibit innate patterns, depend little on learning, and tend to make inflexible responses to their environments; animals high on the scale also show some innate patterns, but their greater ability to learn produces a much more flexible repertoire of responses to stimulation (see HUMAN BEHAVIOUR, INNATE FACTORS IN).

Learning. Behavioral changes wrought by prior experience are used to define learning (excluding temporary changes such as in fatigue and enduring alterations as those resulting from growth). Learning may be classified as conditioned or complex.

In classical conditioning, a reflex that normally is activated by one stimulus is paired with a nonactivating stimulus that ultimately comes to initiate the response. The salivary reflex of an inexperienced dog, for example, normally is elicited by food on the tongue but not by the sight of food. If sight-of-food precedes food-on-tongue several times, however, the visual stimulus by itself begins to elicit the salivary response. In operant conditioning, responses seem to be modified by reward or punishment. **A** response that is rewarded tends to be repeated; when followed by punishment, the response tends not to be repeated. A dog, for example, normally is not likely to press a lever or bar that he sees. When, however, he presses the bar by chance and receives a food reward, subsequent sight of the bar tends to evoke the act of pressing. Many parental efforts to manipulate the behaviour of their children rest on operant conditioning (*e.g.*, spanking or rewards of candy).

Complex learning is a process in which diverse responses are integrated into a relatively smooth sequence, as in maze learning by laboratory animals and in perceptual motor learning and verbal memorizing by human beings. The features of complex learning are not adequately explained in terms of chains or combinations of classical or operant conditioned responses (see LEARNING THEORIES).

Thinking. Beyond sensing and perceiving, cognitive processes include such ideational activities as concept formation, inductive and deductive logical thinking, productive thinking (judging, comparing, and problem solving), novel thinking (originality and creativity), and chimerical thinking (phantasy and dreams; see THOUGHT PROCESSES, TYPES OF).

Intelligence. The term intelligence is used in attempts to evaluate and measure actual or potential ability to perform selected tasks by complex learning and thinking. Intelligence is not defined easily; operationally it is considered to comprise the ability to perform whatever tasks are arbitrarily included in intelligence tests. Some abilities ordinarily are poorly represented in such tests (*e.g.*, mechanical, artistic, or social intelligence). Intelligence-test content tends to reflect the theoretical approach of the person who constructed the test. Some theorists conceive of intelligence as an amalgamated, conglomerate entity and speak of intelligence as a generalized cognitive function. Others factor intelligence into "independent" dimensions, holding for a separation of several fundamental components of intelligence (see INTELLIGENCE, THEORIES OF).

Psycho-analytic theory

Personality. The dynamics of personality refer to the processes of adjustment by which each individual copes with the exigencies of environment. Psychoanalytic theorists propose a model of subjective activity at three levels of awareness: conscious, preconscious, and unconscious.

Metaphoric character actors play their drama in this psychoanalytic version of personality: the tempest-tossed ego is depicted as seeking to reconcile the requirements of the carnal id (physiological motives) and the puritanical superego (social motives) with practical considerations of the empirical world. In this framework, the ego often is said to fail because of behavioral conflicts and frustrations and to experience the terrible emotion of anxiety. It is held that the ego dispels anxiety by many defense mechanisms; for example, dreams, parapraxia (Freudian slips), humour, regression, withdrawal, and rationalization (see PERSONALITY, THEORIES OF).

Personality assessment is concerned with traits and clusters of behavioral traits. Words such as abject, Machiavellian, and lesbian describe personality traits peculiar to the individual. Personality traits often are clustered; *e.g.*, introversion–extroversion, dominance-submission. These traits and clusters may be appraised with techniques that include paper-and-pencil self-description inventories and projective tests (*e.g.*, those in which the individual describes what ambiguous inkblots remind him of, or in which he creates stories about test pictures).

Personality also is assessed through judgmental ratings, by direct interview, and via study of life histories. Despite the extraordinary number of words used to describe personality traits, investigators using factor analysis (a statistical technique designed to disclose independent, basic variables) have suggested that there are about 16 fundamental dimensions of personality (see PERSONALITY, MEASUREMENT OF).

Other topics. The study of group dynamics is concerned with the behaviour of social groups and the interactions of individuals within them. Special topics of investigation include social movements, group attitudes, leadership, networks of communication, cooperative problem solving, propaganda, and rumour (see PSYCHOLOGY, SOCIAL).

Behaviour pathology refers to such deviations as abnormal personality and disorders of speech, of hearing, and of reading (*e.g.*, see PSYCHOSES; PSYCHONEUROSES).

Major specialties. The special fields of applied psychology selectively draw on all the basic subject matter or topics of modern psychology. Thus, child psychology considers the topics that apply to children, genetic psychology concentrates on those concerning origin and development, physiological psychology stresses the neurological and anatomical bases of behaviour, and comparative psychology deals with such behaviour as it differs from one species of animal to another.

Some specialties of psychology emphasize particular basic topics. Thus, social psychology concentrates on secondary motives and group dynamics, and differential psychology compares measures of intelligence and personality as they vary with such characteristics as age, sex, and socio-economic status.

Other specialties apply the results of basic research to practical needs. Thus, industrial psychology applies them to employee selection and other problems of business, school psychology to questions surrounding educational instruction, and military psychology to such efforts as psychological warfare. Military purposes also are served by engineering psychology in the design of man-machine systems (*e.g.*, aircraft controls). Clinical psychology is concerned with personal adjustment and embraces such fields as counselling and psychotherapy, including behaviour modification (*e.g.*, see PSYCHIATRIC TREATMENT, CONCEPTS OF).

METHODS OF PSYCHOLOGICAL RESEARCH

Psychological research, as in other sciences, is experimental or at least observational and can be conducted in the laboratory or in the field. All subdisciplines rely on experimental techniques for their progress.

Modern psychological laboratories stock a variety of hardware: one-way mirrors, tape recorders, motion-picture cameras, electric timers, tachistoscopes (visual short-exposure devices), spectroscopes and colorimeters, sound-wave generators and amplifiers, mazes, electroencephalographs, surgeries for brain extirpation, anechoic (soundproof) rooms, and high-speed computers. And these represent only a fraction of such equipment.

Laboratory work with human beings is highly limited since it must be done with substantial ethical and legal constraint. Experimentation with other animals is less restricted, often at the sacrifice of the lives of healthy subjects (*e.g.*, monkeys or cats). Considerable empirical work is done without laboratory restraints to ensure normal spontaneity. Typical observations of this sort include those of human behaviour at traffic lights and of the political activity of social groups. Habitat observations of animals reveal territorial behaviour and migratory habits.

Suitable only for human beings, an extensively used method employs questionnaires on such topics as food preferences, racial attitudes, childhood experiences, and political sympathies.

Statistical methods

Psychologists analyze their quantitative data by statistical techniques. Since most data of psychological inquiry reflect the contributions of many factors, the effects of any logically identifiable simple variable are not always apparent. For example, measured intelligence reflects both genetic and environmental variables. The independent contribution of each variable and the effects of its interaction with others may be estimated through such statistical methods as analysis of variance and by advanced correlation matrix strategies (*e.g.*, factor analysis).

Psychologists tend to be categorized according to their preferences in research methods or in clinical techniques. Extreme behaviorists, for example, argue that the sole subject matter of psychology is observable, operationally defined behaviour; all else, they insist, falls outside the domain of science. Others argue that introspective verbal reports of subjective experience are the legitimate subject of scientific psychology and that experimental data constitute the heart of the discipline. The cleavage is manifest in major differences among clinical practitioners. To those who practice so-called behaviour modification, a phobia is considered cured when the symptoms are extinguished through operant conditioning; *i.e.*, when such observable behaviour as a fearful response to heights has been modified by rewards and punishments. To other psychotherapists the phobia is considered cured when the sufferer has gained introspective insight into his unconscious conflicts.

ACADEMIC AND PROFESSIONAL ASPECTS OF PSYCHOLOGY

People who earn their livings as academic psychologists or as practitioners are to be found worldwide, many with close ties to psychiatric medicine and education. Exchange of worldwide information on psychology is effected by such organizations as the United Nations, the International Union of Psychological Sciences, and the International Association of Applied Psychology, as well as by publications (*e.g.*, International Journal of Psychology).

Professional organizations for psychologists are active in many countries, including Japan, Mexico, West Germany, Thailand, Romania, The Netherlands, Belgium, and Czechoslovakia. The growing number of journals includes Magyar *pszichológiai* szemle (Budapest, Hungary), *Nordisk psykologi* (Copenhagen, Denmark), *Przegląd* Psychologiczny (Wrocław, Poland), Psychologia *Africana* (Johannesburg, South Africa), Rivista di *psicologia* (Florence, Italy), and Indian Psychological Review (Varanasi, India). In France there are about nine major journals of psychology, almost all receiving government support. Cuba has developed a five-year university curriculum in psychology (general, developmental, clinical, social, and mathematical). In the Philippines psychologists are concentrating on descriptive studies, especially in social psychology. Meanwhile, psychology in the U.S. has become so influential internationally that some authorities describe it as approaching the status of a "colonial science."

Of approximately 35,000 psychologists in the U.S. (early 1970s) about 40% worked for colleges and universities; 20% worked for federal, state, or municipal governments (in such institutions as hospitals, military establishments, and prisons); about 20% in business and industry; 10% in public and private schools; about 10% were self-employed in private practice. Of all U.S. psychologists, about 37% were primarily interested in clinical psychology; 14% in educational and school psychology; 14% in experimental, physiological, and developmental psychology; 11% in industrial and personnel psychology; 11% in counselling psychology: 9% in social and personality psychology; and about 4% in general and engineering psychology. Their incomes averaged roughly $16,000 per year. Industrial psychologists earned the most (an average of $19,600 per year, with a range beyond $35,000) ; school psychologists were paid the least (averaging $13,000 per year and ranging to $19,000).

Psychologists were unevenly distributed in the U.S.: in numbers per 100,000 of population, there were about 99.5 in the District of Columbia (largely reflecting employment), 29.6 in New York State, 21.6 in Vermont, 17.2 in Michigan, 9.4 in North Carolina, and 5.4 in South Carolina; the average for the U.S. was 17.1.

The bulk of U.S. psychologists belong to the American Psychological Association (APA), with central offices in Washington, D.C. Members of the APA, are to be found in such countries as Egypt, Israel, Nigeria, Turkey, Venezuela, Canada, and Ethiopia. There are about 29 divisions of the APA, including: General Psychology, Teaching of Psychology, Experimental Psychology, Evaluation and Measurement, Physiological and Comparative Psychology, Developmental Psychology, Personality and Social Psychology, Society for the Psychological Study of Social Issues, Psychology and the Arts, Clinical Psychology, Consulting Psychology, Industrial Psychology, Educational Psychology, School Psychology, Counselling Psychology, Psychologists in Public Service, Military Psychology, Maturity and Old Age, Society of Engineering Psychologists, Psychological Aspects of Disability, Consumer Psychology, Philosophical Psychology, Experimental Analysis of Behaviour, History of Psychology, Community Psychology, Psychopharmacology, and Psychotherapy, Hypnosis, and State Affairs.

Ethical standards

The APA holds annual meetings for exchange of research information and fur solving professional problems. It published approximately 13 journals in 1970 including: ***American Psychologist*** (stressing broad professional aspects), ***Contemporary Psychology*** (containing critical reviews of books and films), ***Psychological Abstracts*** (nonevaluative condensations of the world's literature of interest to psychologists), and the ***Psychological Bulletin*** (offering reviews of research literature). The APA maintains liaison with many state and regional societies and with foreign organizations. It demands a strict professional code of behaviour of its members; under its ***Ethical Standards of Psychologists,*** members are expected to maintain high standards of competence, to show sensible regard for moral expectations of the community, to use modesty and caution in public statements, and to withold information obtained in confidence.

The competence of professional psychologists is self-regulated by the APA in cooperation with local associations and through certification and licensing by many states. One of its agencies, the American Board of Examiners in Professional Psychology, administers examinations in clinical, counselling, and industrial psychology; it awards diplomas to those who exhibit outstanding competence. Such diplomas (owned by about 7 percent of all U.S. psychologists) are particularly esteemed by private practitioners. Psychologists in private practice also have to meet statutory certification or licensing requirements in approximately 30 states and in four Canadian provinces. Such certification requires written and oral examination, evidence of moral character, and satisfactory education at accredited institutions.

Almost all professional psychologists in the U.S. held advanced degrees in the 1970s: 65% had doctorates; 33%, the master's; and 2% held only bachelor's degrees. Approximately 210 universities in the U.S. and Canada offer courses leading to advanced degrees in psychology. Advanced training includes concentrations in psychology, clinical, general-experimental, physiological, personality-social, industrial, and counselling. The APA provides teams of experts for evaluating doctoral programs, approximately 70 institutions having been approved for clinical training.

While the Ph.D. is the usual doctoral degree in psychology in the U.S., a new degree, the Psy.D. (doctor of psychology), is being granted by the University of Illinois and was under active consideration by several other institutions. The Psy.D. requires an internship, but no research dissertation; it was designed for practicing psychologists working in clinical settings.

Most academic instruction in psychology is conducted in colleges and universities, but some courses are taught in U.S. high schools. The general course in psychology has the largest enrollment (exclusive of freshman English) of any single course—about 500,000 students each year. Specialized courses include history of psychology, medical psychology, legal psychology, design of experiments, hypnosis, psychology of race, and psychology of religion. Almost all U.S. colleges and universities grant bachelor's degrees with a major or minor in psychology.

The teaching and practice of psychology in the United Kingdom, France, West Germany, New Zealand, and Australia shows virtually the same pattern as the U.S. and Canada. In the Soviet Union psychologists are much less concerned with psychometric investigations of individual differences, giving little attention to intelligence tests and personality inventories. Psychological research in the U.S.S.R. is under the direction of the Central Institute of Psychology, with departments of General Psychology, Child Psychology, and Pedagogical Psychology. The various laboratories of the institute consider psychophysiological problems, the behaviour of laboratory animals, industrial work, and educational problems. Soviet care of psychiatric patients has been impressive to Canadian, British, and U.S. observers.

BIBLIOGRAPHY. A. ANASTASI, ***Fields of Applied Psychology*** (1964), a detailed discussion of the numerous applications of psychology; BEHAVIORAL AND SOCIAL SCIENCES SURVEY COMMITTEE, ***The Behavioral and Social Sciences: Outlook and Needs*** (1969), a survey of the relation of psychology to other disciplines; J. COHEN, "Eyewitness Series in Psychology" (1969–71), a collection of 20 paperback separates, historically oriented and richly illustrated, covering important areas of psychology; N. FRIEDMAN, ***The Social Nature of Psychological Research*** (1967), descriptions of psychological experiments that are unique to behavioral science; B. LUBIN and E.E. LEVITT, (eds.), ***The Clinical Psychologist: Background, Roles and Functions*** (1967), an analysis of the clinical psychologist in modern society; W.~SAHAKIAN (ed.), ***History of Psychology*** (1968), carefully selected excerpts from the literature upon which contemporary psychology is based; J.B. SIDOWSKI (ed.), ***Experimental Methods and Instrumentation in Psychology*** (1966), descriptions of laboratory techniques used by psychologists; E.B. WEBB (ed.), ***The Profession of Psychology*** (1962), discussion of the antecedents of contemporary psychology, its role as an academic and applied discipline, and its relations with other professions and the public.

(J.C.)

Psychology, History of

From its beginnings in supernatural beliefs, magic, and taboo, psychology has matured to a science of such broad proportions that professional psychologists today must specialize on narrow fragments of the broader discipline. Specialties are identified by such adjectival terms as physiological, comparative, developmental, personality, social, clinical, consulting, industrial, educational, counselling, military, engineering or human factors, consumer, philosophical, experimental, psychotherapy, psychopathology, and psychopharmacology. Hundreds of professional journals are devoted to psychology, and their numbers are augmented by more than a thousand journals for such allied fields as psychiatry. The world's first psychological laboratory appears to have been established by Wilhelm Wundt in 1879 at Leipzig University, Germany, marking what has been called the birth of psychology as a science. Two years later Wundt founded the first journal of experimental psychology. G. Stanley Hall (1844–1924) is thought to have founded the first laboratory of psychology in North America (1883); the first journal in English, still being published, the ***American Journal of Psychology***

First psychology laboratory

(1887); and (with other psychologists meeting at Clark University in Worcester, Massachusetts, on July 8, 1892), the American Psychological Association, Hall serving as first president. The first International Congress of Psychology met in Paris, France, in 1889 with psychiatrist Jean-Martin Charcot (1825–93) presiding over an attendance of 203. In the 1970s the American Psychological Association alone comprised 30,592 members; according to the International Directory of Psychologisfs (1966), of 20,000 questionnaires sent to psychologists outside the U.S., only 8,000 responded (the People's Republic of China declining; see also PSYCHOLOGY).

Contemporary psychology developed from a confluence of diverse streams: these include religions of the Orient and the classical Greek and Roman philosophies; Jewish, Christian, and Islāmic medieval thinking; philosophical psychologies of the Renaissance and modern periods; psychiatric and biological influences of every age; British physicists of the 17th to 19th centuries; 19th-century German and Russian physiologists, British-American evolutionists and statisticians; continental European psychiatry of the first half of the 20th century; German and North American experimentalists and theorists of the last and present centuries.

Philosophical-religious roots

NON-WESTERN INFLUENCES

Psychology had inchoate beginnings in animism, the view that objects such as trees have an indwelling vital principle, or "soul," and in hylozoism (or hylopsychism), the belief that matter has life or sensation. An epic poem of ancient India, the Mahdbhdrata, disclosed five senses, and an intermediary faculty (*manas*) for receiving sense impressions that were transmitted to *buddhi* (intuitive discernment, intellect) before reaching "soul" or self (*ātman*). A group of Theravada Buddhist realists (Vaibhāṣikas) held that "mind observes, manas deliberates, and consciousness (*vijñāna*) discriminates. In China, Buddhist philosophers of the school known as Ch'an (Zen in Japan) emphasized meditation as a way to enlightenment. Morita therapy, a form of Japanese psychotherapy drawing upon Zen, developed (1930s) by Shoma Morita, rests on efforts to diminish excessive self-consciousness. The Chinese humanism of the philosopher Confucius (551–479 BC) held for the perfectibility of man, emphasizing moderation, balance, and harmony. A Chinese philosopher, Mencius (371–289? BC), espoused a form of Confucianism that assumed original goodness in human nature. According to the ancient Chinese book Lao-tzu and its doctrine of Taoism, a desirable life was one of tranquility, spontaneity, and humility or weakness.

Morita therapy

By 525 BC, when the Egyptian physician Imhotep (c. 2850 BC) was pronounced god of medicine, sleep was employed in psychotherapy, as were dancing, drawing, music, and similar occupational measures. Later, the cult of Asclepius, Greek god of medicine, used sleep, dreams, and suggestion in treatment (see DREAMS). Ancient Mesopotamians seemed aware of psychosomatic medicine, relying on the suggestive power of incantations by priests. Babylonians, and Israelites and Christians at the time of Jesus, invoked demons as the cause of psychiatric disorder, holding that they could be driven from the sufferer in a ritual called exorcism.

THE CLASSICAL WORLD

Among the ancients, Greek philosophers, led by Aristotle (384–322 BC), were pre-eminent in advancing psychology. Heracleitus of Ephesus (c. 540–c. 480 BC) reasoned that "soul" in man was fire and that inner serenity was attained through reason. Anaxagoras of Clazomenae invested all of the functions of fire in nous ("mind"). A precursor of modern behaviorists, Empedocles of Agrigentum (c. 490-430 BC), identified thinking with perceiving and believed that it is dependent on bodily change. He offered what amounted to the modern doctrine that no matter how it is stimulated each sense organ responds specifically; for example, one has visual experiences whether his eyes are stimulated by light or by pressure on the closed lids. Refining Empedocles' theory of specific energy of sense organs, Democritus of Abdera (460?–357 BC) explained sensory experiences by asserting that atoms are emitted from objects and transported through air to produce images by stimulating sense organs; this early variation of physiological psychology reduced psychological activity to the motion of fire atoms.

During what is called the period of the Greek Enlightenment, the Sophist philosopher Protagoras of Abdera (c. 481–411 BC) enunciated a doctrine of sensationalism that attributed all psychological activity to sensory function alone. Sensationalism later was adopted by French philosopher Etienne Bonnot de Condillac (1715–80), who tried to imagine what human personality would be like if limited to the sense of smell.

The first Western thinker to suggest the existence of unconscious mental activities seems to have been Socrates (469–399 BC), who, in anticipation of modern psychoanalysis, called for self-analysis. Holding that knowledge can remain dormant in the "soul," he employed a so-called maieutic technique that (like contemporary psychoanalysis) was supposed to bring to consciousness that which had been forgotten. Taking the lead from Socrates, Plato (428–348/347 BC) discussed interpretation of dreams and unconscious motivation, as well as association of "ideas." Plato's was a biological psychology that localized "mind" in the brain, "will" in the heart, and appetites and desires in the liver or stomach. His pupil Aristotle (384–322 BC) has been called the founder of functional psychology and apparently originated the concept of emotional catharsis, a kind of "purging" of pent-up emotion through such acts as weeping. Attributing a double aspect to human "soul," Aristotle cited the first as passive "mind," analogous to a blank writing tablet (tabula rasa) on which sensory experience was imposed; the second he held to be actively rational and motivational. Aristotle elaborated on the traditional five senses, also considering memory, dreams, and sensory afterimages (*e.g.*, the effect of staring at a bright light). He claimed that there also is a "common sense" for unifying sensory experience and serving as a basis for memory.

Unconscious activities

The so-called Stoic philosophers held to a dualistic notion that flesh and "soul" are in opposition. Seneca (c. 4 BC–AD 65) propounded this doctrine, and Marcus Aurelius Antoninus (121–180) sharpened the distinction. The Stoic theory of perception compared sensations to impressions on wax (Cleanthes, 331/330–232/231 BC) and viewed the "soul" as originally devoid of impression (again, tabula rasa). In addition, "soul" was said to embrace five senses and numerous other faculties. Stoic psychology, practically applied by Epictetus (c. 55–c. 135), emphasized the role of "the will," one's attitude, indifference to pain, calling tranquillity the ideal of psychological adjustment and passion a disease. Lucretius (c. 95–55 BC) reduced "mind" or "spirit" to physical terms, explaining sensation as did his materialistic predecessor, Democritus (460?–357 BC), to hold emotion to emanate from animus, or physical "mind." Lucretius followed the lead of Epicurus (341–279 BC), according to whose atomic theory life, consciousness, and experience were equally as real as the atomic movements from which they sprang.

EARLY CHRISTIAN–MEDIEVAL PERIOD

The Hebrew term for "soul" (nefesh, that which breathes) was used by Moses (*c.* 13th century BC), signifying an "animated being" and applicable equally to nonhuman beings. The Hebrews used the term to apply to the entire personality but reserved the concept *ruaḥ* ("spirit") to denote a principle of life, "mind," and occasionally "heart." Nefesh was often used as if it were the seat of appetite, emotion, and passion and, conjoined with "heart," was held to encompass intellect, will, and feeling. New Testament usage of *psychē* ("soul") was comparable to nefesh. Jesus' complete dualistic demarcation between flesh and "spirit" was quite evident, and St. Paul (died AD 64) exceeded the dualism of Jesus to invent a triune man, regarding "spirit" (pneuma) as a divinely inspired life principle, "soul" (psychd) as man's life in which "spirit" manifests itself, and body (*sōma*) as the physical mechanism animated by "soul." In New Testament usage, man variously was regarded as having "free" will (*thelēsis*); as a rational being with faculties and "mind" (*nous*); as a contemplating, understanding, intelligent being with "mind"(*phronēma*); as a sublime being or "spirit" (*pneuma*); and as a being with predispositions and attitudes, or "heart" (*kardia*).

Plotinus and the "higher soul"

The duality of man as "soul" (defined as a vital force of the physical body), and man as a higher "spiritual" self (pneuma) was enunciated by Philo of Alexandria (c. 30/*c.* 20 BC–after AD 40). Plutarch (c. AD 48–c. 119) also distinguished rational "spirit" (*nous*) from psych.? ("soul"), conceived as the seat of man's emotional, sensuous, and animate nature. St. Irenaeus (c. 120/140–200/203) likewise separated a "psychic" breath of life (a perishable characteristic of man tied to body) from a kind of eternal, animating "spirit." Stressing "will" or "freedom" of choice, Origen (*c.* 185–254) differentiated between so-called rational and animal souls. Credited as founder of Neoplatonism, Plotinus (205–270) seems first to have developed an empirical (objective) psychology, though it was introspectively (subjectively) oriented. He held that conscious activity (cognition) by "higher soul" depended on data from the senses. Physical sensations, passions, and feelings were seen as passive states of "lower soul" and conscions perception or reflection as functions of "higher soul," and Plotinus was led to the experience of self-consciousness. The culmination of Patristic (by the Fathers of the Christian Church) and of Neoplatonic psychology was found in the psychology of St. Augusine (354–430), who constructed his system on Origen's hypothesis of will and Plotinus' theory of self-consciousness. Augustine saw the fundamental aspects of psychological experience as being memory, understanding, and will. His belief that experience represents external physical (phenomenological) truth was an early basis for a branch of philosophy now called phenomenology.

Seizing on Aristotle's threefold concept of "soul" (vegetative, sensitive, intellective), the Arabic philosopher Avicenna (980–1037) constructed a psychology entailing external and internal senses, the former utilized to perceive external objects. He traced psychiatric disorder to disorders of the brain. Another Aristotelian Muslim, Averroës (1126–98), postulated "soul" as a corporeal form connected with body, its supreme powers said to consist of "passive intellect" or imagination. The Jewish philosopher Maimonides (1135–1204) followed the same line, especially in regarding "agent intellect" (invented by Aristotle) as the source of intellectual knowledge; and a Christian philosopher, St. Thomas Aquinas (1225–74), in following the Aristotelian tradition, accepted a threefold division of "soul" and a belief in human freedom, while emphasizing instinctual aspects of psychological function. Whereas Aquinas placed intellect above will, John Duns Scotus (c. 1265–1308) was more voluntarist than intellectualist, assigning priority to will. Splitting conscious "soul" into the intellective and the sensitive, English schoolman William of Ockham (*c.* 1285–1349) ascribed numerous individual faculties or powers to "soul." Another British Scholastic (in the Aristotelian tradition), Roger Bacon (*c.* 1220–c. 1292), distinguished between external and internal experience, the former said to be related to sense experience, the latter to internal activities.

RENAISSANCE TO THE MODERN PERIOD

A Spanish humanist and precursor of modern psychology, Juan Luis Vives (1492–1540) inclined toward an empirical approach to psychology and invoked direct experience as a basis for knowing "things themselves," as is done by contemporary phenomenologists. The French skeptic Michel de Montaigne (1533–92) shared Vive's views of an empirical psychology. During this period the term psychology already was being used by the German theologian Philipp Melanchthon (1497–1560), and Niccolò Machiavelli (1469–1527) was declaring the domination of man by his passions.

On the continent of Europe, so-called rationalists sought to resolve what they called the "mind-body problem," which originated as a strict dualism in the system of French philosopher René Descartes (1596–1650), who wrote that the interaction of mind and body occurred in the brain's pineal gland. This view terminated in a psychophysical parallelism (the notion that mind and body do not interact but exist "side by side") in treatments by Dutch philosopher Benedict de Spinoza (1632–77) and German philosopher Gottfried Wilhelm Leibniz (1646–1716). Some give Descartes the distinction of founding physiological psychology, for it was he who explained the behaviour of nonhuman animals entirely on the basis of mechanistic functions in the nervous system, denying "souls" to such creatures. He also advanced a theory that accounted for visual perception of distance, shape, and size in terms of secondary sensory cues. A French materialist, Julien Offrey de La Mettrie (1709–51), taking his lead from Descartes, purported to explain the psychology of man in terms of the functions of a machine. Spinoza, asserting that the "order and connection of ideas is the same as the order and connection of things," laid the foundation of modern holistic psychologies; *i.e.*, those that consider man as a unified whole rather than as an arrangement of separate physical parts. Also attributable to him was the view that a rational cause accounts for every human action, a view that Sigmund Freud later also offered. Leibniz resolved the "mind-body problem" by asserting both to be in pre-established harmony and was credited with the notion of "petites perceptions" (unconscious representations), a theory of unconscious function.

Man as a machine

While rationalists on the continent occupied themselves with the "mind-body" dispute and concluded that "ideas" are inborn, British empiricist philosophers, prompted by John Locke (1632–1704), postulated "mind" as a "white paper" (again tabula rasa) on which all sensations were imprinted. Locke agreed with Aristotle that there was nothing in "mind" that was not first in the senses. Leibniz appended to this assertion "except the mind itself," a concept that Immanuel Kant later used to construct his a priori theory of time and space as pure forms of sensibility. According to Locke, man's knowledge and the development of his personality were accounted for by "ideas" experienced together with their associations. He held that "ideas" sprang from two sources, sensation and reflection, the latter being an inner sense. Although Locke coined the phrase association of ideas, British association psychology really dates back to Locke's compatriot, Thomas Hobbes (1588–1679), who claimed that sense impressions supplied consciousness with its content and that by their associations "thought" was produced and memory or forgetfulness was explained in their gradual decay. The wide use of the term association and the founding of associationism has been imputed to David Hartley (1705–57), though it seems to have stemmed from Aristotle. In preference to association, Thomas Brown (1778–1820) used the term suggestion and developed a neurological theory of memory, association, and learning. An Irish empiricist in the British tradition, George Berkeley (1685–1753), constructed a theory of visual perception of distance in which cues came from the surroundings (context) in which objects are seen. Berkeley's idealism as reflected in the assertion "to be is to be perceived," terminated in psychologism, a view that equated "reality" with experience. A Scot, David Hume (1711–76). carried the concept of association into a denial of the principle of causality, theorizing that so-called cause-and-effect relations represent mere sequence rather than necessarily consequential relationships. Hume also reduced human "soul" to "a bundle of perceptions." Taking his cue from Hume, James Mill (1773–1836) constructed a psychology of association based solely on sensory proximity in time and space, allowing no creative function to "mind." In an effort to correct what he considered his father's deficiencies, John Stuart Mill (1806–73) ascribed an active role to "mind" and accorded new qualities to complex "ideas" instead of regarding them as merely a summation of simple "ideas." While the father represenled the simple "mental compounding of ideas," the son held that "ideas" actually could be changed through "mental chemistry."

Mental chemistry

Contributions of the German idealists to psychology during the 18th and 19th centuries were essentially philosophical. Christian Wolff (1679–1754) distinguished between empirical and rational psychology as early as 1732. Immanuel Kant (1724–1804) contributed the notion of an a priori intuition of time and space, and Johann Gottlieb Fichte (1762–1814) agreed, while Johann Friedrich Herbart (1776–1841), credited by some with being author of the first textbook of psychology, opened the way to mathematical approaches to psychology. Rudolf Hermann Lotze (1817–81) endorsed the inborn-intuition-of-space hypothesis and advanced a theory of local signs (or local sensory cues to space). Arthur Schopenhauer's (1788–1860) "psychology of will" anticipated Freudian instinct theory, plus later movements in psychology th [illegible] rratic ism and pessimism; but Friedrich Ni [illegible] (1844–1900) optimistically turned his attention to a psychology of power, the base on which the psychology of Alfred Adler (1870–1937) later was founded. Nietzsche's pronouncements on repression, conscience, and defense mechanisms found their way into Freudian theory. German philosophical psychologists of the 20th century, though influenced by these idealistic philosophers, moved into what are now called phenomenology and existential psychology. The act psychology of Franz Brentano (1838–1917) set the stage for phenomenological psychology; he sought to ascertain the precise contents of consciousness about a given object by an analysis of specific human acts; *i.e.*, what people said about their immediate conscious experience. His student Edmund Husserl (1859–1938) virtually defined phenomenology as this kind of descriptive analysis of subjective processes or acts. Max Scheler's (1874–1928) application of phenomenology to values, especially to moral and religious experience, was to become the foundation of Viktor E. Frankl's (1905–) logotherapy, a form of psychotherapy based on meaning. Scheler conceived of phenomenology as the a priori knowledge of untutored experience. Within this phenomenological movement, Martin Heidegger (1889–) set the stage for existential psychiatry by his notion of dasein (man as a human being, operating within this world; *i.e.*, being-in-the-world). The existential psychology of the French philosopher Jean-Paul Sartre (1905–) is in the phenomenological tradition, but like Heidegger's thinking, it was heavily influenced by Søren Kierkegaard (1813–55), a Dane who sired such existential concepts as the phenomenology of anxiety and an emphasis on choice. While phenomenology was effectively applied to psychiatry by Frankl's logotherapy (or will to meaning) and by the U.S. psychologist Carl R. Rogers (1902–) to personality theory and psychotherapy, it was the Swiss thinkers Ludwig Binswanger (1881–1966) and Medard Boss (1903–) who were most successful in the application of existentialism to psychiatry. German existentialist Karl Jaspers (1883–1969), author of General Psychopathology (1913), was more active in applying existentialism to philosophy.

Origins in biology and psychiatry

ADVANCES IN BIOLOGY AND PHYSIOLOGY

The Pythagorean physician Alcmaeon of Croton (flourished 6th century BC) had identified thinking or consciousness as the distinguishing feature of man and localized these functions in the brain; he traced perceiving to specific sense organs and emotions to the heart. Britain saw Thomas Young's (1773–1829) publication of a three-component theory of colour and an undulatory (wave) theory of light. The former theory, verified by Hermann von Helmholtz (1821–94), came to be known as the Young-Helmholtz three-colour theory of vision. In 1811 Sir Charles Bell (1774–1842) noted distinct functions of sensory and motor nerves, independently suggested by the Frenchman François Magendie (1783–1855) a decade later; it is currently called the Bell-Magendie law of spinal nerve roots and forecasts the doctrine of specific nerve energy. Bell is also credited with discovery of a sixth sense, muscular sensation, although Aristotle seems to have been aware of the same function, having called it kinesthetic. Another Frenchman, Edme Mariotte (1620–84), earlier (1668) discovered the blind spot in the retina of the eye. Sir Isaac Newton (1642–1727) demonstrated that the experience of white is the composite awareness of all colours in light. In 1838 another Britisher, Sir Charles Wheatstone (1802–75), contributed the concept that the perception of depth depends on disparate images in each eye. In 1861 in France, Paul Broca (1824–80) identified the physical basis of articulate speech in a localized portion of human brain tissue, said to be the first direct evidence that "mind does not function as a single, unitary, homogeneous thing. By 1870 the Germans Gustav Fritsch (1838–1927) and Eduard Hitzig (1838–1907), by electrical stimulation of laboratory animals, had localized centres for voluntary movement in the brain. In 1889 a Spaniard, Santiago Ramón y Cajal (1852–1934), identified neurons as the functional cells of the brain. A British physiologist, Sir Charles Sherrington (1857–1952), coined the term synapse to describe the gaps between neurons and is known for his dictum that "behaviour is rooted in integration" of the nervous system. Others contributing to neurophysiology were Britain's John Hughlings Jackson (1835–1911), distinguished for work relating epilepsy and the brain, and John Langley (1852–1925), who studied the role of the autonomic nervous system in automatically regulating such functions as breathing and emotional expression.

Specific energy of nerves

The British also were stimulated by Charles Darwin (1809–82) and his theory of the evolutionary origin of species and his doctrine of the survival of the fittest. Darwin held that emotions in man were inherited in an evolutionary sense, reflecting emotional behaviour that served the survival of lower animal species. Darwin's friend George Romanes (1848–94) ushered in and coined the term comparative psychology with his book Animal Intelligence (1882). In this evolutionary tradition C. Lloyd Morgan (1852–1936) offered a psychological version of the law of parsimony in 1894, known as Morgan's cannon, by which he urged explanations of behaviour in the simplest, most parsimonious terms feasible. During the same year the evolutionist Herbert Spencer (1820–1903) defined life as a "continuous adjustment of internal relations to external relations." By 1908 William McDougall (1871–1938) had explored social psychology from the standpoint of evolution and of comparative psychology. He and Edward A. Ross (1866–1951) seem to have been the first to publish books with "Social Psychology" in their titles. An American, Christine Ladd-Franklin (1847–1930), offered an evolutionary account of the development of human colour vision.

Evolution

The founding of experimental psychology and experimental aesthetics is associated with Gustav Theodor Fechner (1801–87), a German who wrote *Elemente* der Psychophysik (Elements of Psychophysics) in 1860. The book marked the founding of psychophysics, experimental psychology, and the formulation of Fechner's law or the Weber-Fechner law: "The magnitude of the sensation . . . is proportional to the logarithm of the fundamental stimulus value." According to Ernst Heinrich Weber (1795–1878), "equal relative increments of stimuli are proportional to equal increments of sensation." Major accomplishments by Fechner include what are called psychophysical methods, techniques for measuring the intensity of subjective experience~.

The doctrine of the specific energy of nerves was enunciated in detail in 1838 by Johannes Müller (1801–58), a German who maintained that awareness is of one's nerves rather than of external objects and that "the same internal cause excites in the different senses different sensations; —in each sense the sensations peculiar to it. . . . External agencies can give rise to no kind of sensation which cannot also be produced by internal causes."

Hermann von Helmholtz (1821–94), a German, accelerated progress in experimental psychology with major advances in sense physiology, including an influential theory of hearing, the Young-Helmholtz three-colour theory of vision, an empirical theory of perception, a theory of unconscious inference, and the application of the theory of specific energies to experiences of the distinctive qualities of sense (*e.g.*, colour, loudness). In opposition to Helmholtz, Ewald Hering (1834–1918) espoused a theory that the power of visual space perception is inborn and offered a six-colour theory of vision. In 1904 Max von Frey (1852–1932) offered a theory of four skin senses: warmth, cold, pressure, and pain, challenging the traditional notion of a single sense of touch. A German, Adolph Jost, is known for Jost's law (1897); *i.e.*, for two associations learned equally well, the older one will be better retained.

DEVELOPMENT OF PSYCHIATRY

The beginnings of psychiatry may be traced to the Code of Hammurabi (c. 1950 BC), which recommended opium and olive oil be used as cures for psychiatric disorders held to be produced by demonic possession. The Greek physician and "father of medicine," Hippocrates (flourished c. 400 BC), developed a theory of personality based on bodily fluids (humours), reasoning that psychiatric disorder was attributable to natural causes rather than to demons. During the period 25 BC to AD 50, Aulus Cornelius Celsus published De rnedicina, a work in which the term insanity seems first to have appeared. During the 2nd century AD the Greek physician Galen of Pergamon (c. AD 130–c. 200) developed a humoral theory of psychopathology. The Arabian physician Avicenna later traced psychopathology to disorders of the brain. In the 12th century, an Italian, Roger Frugardi, recommended a form of brain surgery (trephining) in dealing with psychiatric problems, a practice that can be traced at least to ancient Greece and Rome, perhaps even to prehistoric men, By 1891 the Swiss psychiatrist Gottlieb Burckhardt performed a related kind of psychosurgery, removing a portion of the b: ain surface to make patients docile. A Portuguese psychiatrist, António Egas Moniz (1874–1955), helped develop the operation, and in the U.S. Walter Freeman (1895–   ) with his associate James W. Watts (1904–   ) performed psychosurgery by entering the skull through the eye sockets. Other techniques of neuropsychiatric treatment included giving sufferers of brain syphilis cases of malaria (1917); the use of insulin to induce shock in treating schizophrenia, electroshock therapy (used particularly in depression), and the use of chemical (metrazol) shock therapy.

Demonic possession

The 15th century saw the establishment of the first psychiatric asylum in Valencia, Spain, in 1410. A book called Malleus *maleficarum (The* Witches' Hammer), written in that century by two German Dominicans, Johann Sprenger and Heinrich Kraemer, treated witchcraft, psychopathology, and sexuality. The following century saw the Belgian physician Johann Weyer (1515–88), the "father of modern psychiatry," become one of the first to condemn belief in witchcraft as superstition; it also saw the establishment of Bethlehem Royal Hospital (nicknamed Bedlam) in London for disturbed people. The beginnings of electrotherapy were made by an English physician, William Gilbert (1540–1603), and of "animal magnetism" (mesmerism or hypnosis) by William Maxwell (1581–1641) a century before the Viennese physician Franz Mesmer (1734–1815). Autosuggestion ("I am becoming better and better") was developed by Émile Coué (1857–1926) about 1910.

The birth of modern psychiatry may be traced to the appointment in 1792 at Paris of Phillippe Pinel (1745–1826) as physician in chief of Bicêtre Hospital for the Insane, where he proceeded to remove chains from the inmates. An emotional basis for psychiatric disturbance was endorsed by Pinel's student Jean Esquirol (1772–1840), the reputed founder of French psychiatry. Several years later the first U.S. textbook of psychiatry was authored by a professor of chemistry, Benjamin Rush (1745–1813). What is now called schizophrenia, Bénédict-Augusten Morel (1809–73) labelled (1860) *démence précoce* (dementia praecox). He believed it to be a premature form of dementia. Ewald Hecker (1843–1909) identified a form of the disturbance as "hebephrenia," and Karl Kahlbaum (1828–99) called another "catatonia." They opened the way for Emil Kraepelin (1856–1926) to formulate an elaborate system of psychiatric classification that still is in use. The term schizophrenia itself was introduced by a Swiss psychiatrist, Eugen Bleuler (1857–1939).

During the 19th century, psychiatrists (especially Pinel and Kraepelin) tended to view all psychiatric disturbances as being related to physical disease, but a shift to psychological (functional) explanations was precipitated when hypnosis gained recognition. The term hypnosis, introduced in Britain by James Braid (1795–1860), was a topic of controversy in psychiatric circles, gaining reception only gradually and with difficulty. Othei Britishers who employed hypnotism with remarkable therapeutic success were John Elliotson (1791–1868), who used it to diminish pain during surgery, and James Esdaile (1808–59), who performed more than 250 painless operations on hypnotized people. In the U.S. Phineas Parkhurst Quimby (1802–66) used hypnosis to cure the founder of Christian Science, Mary Baker Eddy (1821–1910), of hysterical paralysis in 1862. Hypnotism was vigorously debated in France where it was employed in psychiatric treatment. Jean-Martin Charcot (1825–93) at the University of Paris and Ambroise-Auguste Liébeault (1823–1903) and Hippolyte Bernheim (1840–1919) argued about whether hypnosis was a normal state characterized by suggestion or a pathological symptom exhibited only by hysterics. The Nancy school subscribed to the former position, and at Paris the latter was endorsed. Supporting Charcot, Pierre Janet (1859–1947) theorized that the hysteric individual suffers from psychological weakness (what he called "psychasthenia").

It may be said that psychoanalysis was born when Sigmund Freud (1856–1939) learned about posthypnotic suggestion from Charcot and about the "talking cure" (free association), transference, and catharsis from Joseph Breuer (1842–1925). The birth was marked by Freud's publication of Studies in Hysteria in 1895. With Freudian psychoanalysis the transition from pkysiological to psychogenic (functional) explanations was enhanced. Although evidence for unconscious activities extended from the writings of Plato to those of Leibniz, Schopenhauer, and Karl von Hartmann (1842–1906), Freud made special contributions, resulting in what has been called depth psychology.

Depth psychology

De-emphasizing Freud's stress on sexual motivation, Carl Jung (1875–1961), a Swiss, founded a school of "analytical psychology," highlighting what he called "the will to live" (rather than sex), yet with a more-than-Freudian stress on unconscious function, dividing "mind" into what he called the "personal" and "collective" unconscious. He also coined and developed a technique (the word-association test) for uncovering "complexes"—related words that evoked signs of disturbance. Jung was also the inventor of the terms introvert and extrovert. Freud's Viennese associate, Alfred Adler (1870–1937), founder of the school of "individual psychologv," emphasized a "drive for superiority" or a reaction to the inferiority "complex," rather than sex. German-born Erich Fromm (1900–   ) shifted the attention of psychoanalysts from Darwinian notions of man as one of many animals to man as a unique creature. The cultural school of psychoanalysis drew influence with the development in the U.S. of Harry Stack Sullivan's (1892–1949) "interpersonal theory of psychiatry."

Development of psychology as a scientific discipline

GERMANY AND AUSTRIA

Recognized by many as the founder of scientific psychology, Wilhelm Wundt (1832–1920) saw the object of psychology as the introspective analysis of the contents of immediate experience, with elements called sensations and feelings. He regarded "mind" as activity and psychological causality, the sum of inner experiences. His work included studies of association, psychophysics, reaction time, and folk psychology, representing an elementistic and associationistic system. Two of his students, Oswald Kiilpe (1862–1915) and Edward Bradford Titchener (1867–1927), developed further the "structural psychology" (or content psychology) of their mentor. Husserl and Brentano influenced Kiilpe sufficiently to include function in his content psychology. Since Hermann Ebbinghaus (1850–1909) had successfully investigated the "higher mental processes" of memory by experimental

Structural psychology

methods, it was asked: Should it not be possible to do the same with thinking? Hence the development by Külpe and his colleagues at the University of Würzburg of a new method, "systematic experimental introspection," and a new finding, "imageless thought," suggested that thinking can proceed without sensory images. Ebbinghaus' experimental measurements of memory yielded evidence that forgetting is related to the passage of time (curve of retention) and endorsed the value of repetition in learning. Members of the Wiirzburg school stressed higher-mental process or function; Karl Marbe (1869–1953) studied conscious attitudes and judgment; Henry J. Watt (1879–1925) experimented with thinking itself, finding the free flow of thinking to depend upon one's preparation to undertake a "mental" task; Narziss Ach (1871–1946) investigated awareness, willing, and thinking through systematic experimental introspection, concluding that awareness (an imageless component) is the determiner of thinking; Karl Buhler (1879–1963) studied imageless thoughts, concluding that, though at times accompanied by images, they were devoid of sensory qualities.

In his Analysis of Sensations (1886), Ernst Mach (1838–1916) noted sensations of "space form" and "time form" as being independent of sensory elements; *e.g.*, a circle conceptually remains a circle despite change in size or colour. In 1890 at Vienna, Christian von Ehrenfels (1859–1932), introducing the concept of "form quality" (Gestaltqualitat), claimed that appreciation of form in space and time is a quality that arises from the perceiver's own activity rather than being directly given to the senses. Gestalt psychology as a movement is attributable to Max Wertheimer (1880–1943) and two associates from the Frankfurt Psychological Institute, Kurt Koffka (1886–1941) and Wolfgang Kohler (1887–1967). Their experimental findings proved antagonistic to Wundt's notions of elements of consciousness. When Wertheimer reported the phi phenomenon (*e.g.*, apparent visual movement as in the stationary lights on a theatre marquee), it marked the birth of Gestalt psychology. The Gestalt view is that perceptual qualities exist in the whole that are absent in the individual parts; *e.g.*, no single light bulb moves, but each light seems to move when seen as part of a whole. Other contributions by the gestaltists pertained to insightful learning, brain structure and function, creative thinking, discrimination learning, the physical basis of memory, and laws of gestalt (or perceptual organization). Efforts to apply gestalt principles to child psychology, learning, and personality were made by Kurt Lewin (1890–1947), who invented a "topological psychology" and "field theory" of personality. In the field of psychopathology, Kurt Goldstein (1878–1965) tried to apply gestalt notions in what he called a holistic or organismic approach aimed at helping people to come to terms with themselves and their environment. So-called gestalt psychotherapy is most notably associated with Frederick Perls (1893–1970). The Zeigarnik effect (interrupted tasks are more readily remembered than those that are completed) was contributed by Bluma Zeigarnik (1900–). With the exception of the Russian-born Zeigarnik, the founding members of the gestalt movement were German-born migrants to the U.S. during the rise of Nazism.

Gestalt psychology

ELSEWHERE IN EUROPE

In French-speaking countries, therapeutic success in using suggestibility and hypnosis to treat hysterics led Gabriel Tarde (1843–1904), Gustave Le Bon (1841–1931), and the Italian Scipio Sighele (1868–1913) to apply the notion of suggestibility to social psychology. In 1890 Tarde pointed to imitation and invention as being responsible for the formation, development, and nature of society. A few years later, Le Bon saw crowds and mobs as susceptible to the suggestions of leaders in a manner comparable to an individual's response to a hypnotist. In France and Switzerland, interest in psychology also had turned to the study of children. By 1908 Alfred Binet (1857–1911) had, with Theodore Simon (1873–1962), developed a test of general intelligence; Lewis Terman (1877–1956) at Stanford University revised it in 1916 to produce the Stanford–Binet Intelligence Scale. The concept of I.Q. (intelligence quotient) was contributed in 1912 by the German psychologist William Stern (1871–1938). A group of U.S. psychologists, headed by Robert M. Yerkes (1876–1956), administered about 2,000,000 (Army Alpha) tests to literate soldiers and another 100,000 (Army Beta) tests to illiterates during World War I and discovered the average mental age of the servicemen to be about 13.

Psychology in Switzerland centred at the University of Geneva where Theodore Flournoy (1854–1920), after studying with Wundt, established a laboratory in 1892. With his cousin and former student, Édouard Clapari.de (1873–1940), he founded the journal Archives de *Psy*chologie in 1901. On succeeding to Flournoy's professorship in 1920, Clapari.de founded the Institut Rousseau for child study. Successor to Claparède as professor at the University of Geneva and director of the Institut Rousseau, Jean Piaget (1896–) investigated the development in children of moral, abstract, logical, and concrete modes of thinking.

The Russian school of psychology, permeated with objectivity and physiology, effectively began with work by Ivan M. Sechenov (1829–1905), founder of Russian physiology and scientific psychology, who discovered evidence of localized inhibitory function in the brain (Sechenov's centre) in 1863. According to Sechenov, "all acts of conscious or unconscions life are reflexes." This school's most distinguished member, Ivan Petrovich Pavlov (1849–1936), is principally known for demonstrating conditioned reflexes; he showed that neurotic behaviour may be learned and provided the basis on which much of modern learning theory rests. Vladimir M. Bekhterev (1867–1927) founded Russia's first laboratory of experimental psychology in Kazan. The "reflexological period" of Soviet psychology lasted from 1917 to 1923, followed by the "reactology" of Konstantin Kornilov (1879–1957) until 1931; interested in child psychology, Kornilov sought to structure psychology in terms of Marxist dialectical materialism. A student of his, Aleksandr Luria (1902–), concerned with "defectology" (disorders of the handicapped), made efforts to investigate pathophysiology (psychiatric disorder) through rigorous scientific methods. Many other Soviet researchers were interested in child study, including L.V. Zanov (1901–), Leo Vigotsky (1896–1934), and Aleksey Leontiev (1903–), the latter two being involved in developmental psychology. Since 1931 Soviets have shown interest in a Marxist version of cognitive psychology; *i.e.*, the study of man as controlled by his goals, will, purposes, needs, thinking, and the like. On the basis of a resolution jointly issued in 1950 by the Soviet Academy of Sciences and the Soviet Academy of Medical Sciences, psychology in that country has been reconstructed along Pavlovian lines, characterized by a major emphasis on the neurophysiology of behaviour. The decade of the '50s saw the establishment in the U.S.S.R. of the Society of Psychologists, with A.A. Smirnov (1894–) as president.

Russian psychology

The beginnings of serious study in psychometrics and biometrics were in Great Britain, personified by Sir Francis Galton (1822–1911), Karl Pearson (1857–1936), and others at University College, London. Credited with being the first to investigate the psychology of individual differences and to develop the technique of statistical correlation, Galton was as well an active anthropologist, contributing to such fields as eugenics, the inheritance of behavioral traits, and fingerprint identification. Pearson developed a technique for computing correlation coefficients known as the Pearson r. Two mathematician-astronomers, a German, Carl Friedrich Gauss (1777–1855), and a Belgian, Adolphe Quetelet (1796–1874), also contributed to statistics, the former providing equations for the normal probability (Gaussian) curve and the latter providing applications of probability to social and life sciences. Charles Spearman (1863–1945), a British psychologist who immigrated to the U.S., advanced the method of correlation in identifying so-called factors of intelligence. At University College, London, James Sully (1842–1923), after writing the first English-language text for psychology, Outlines of Psychology (1884), established a laboratory of psychology in 1897. British experi-

Instinct

mental psychologist W.H.R. Rivers (1864–1922) directed that country's first psychological laboratory (1897) at Cambridge and headed another at University College, London. Cambridge University's psychologist, James Ward (1843–1925), in the tradition of Brentano and act psychology (see above), wrote a most influential article "Psychology" in the 9th edition (1886; and in revised form in the 11th) of *Encyclopædia Britannica* Ward's influence was carried on by his student George Frederick Stout (1860–1944), whose textbook written from the standpoint of act psychology, *A Manual of Psychology* (1899), became a standard work for a quarter of a century. First to define psychology as the study of behaviour (in 1905), William McDougall (1871–1938) founded in England his own school of "hormic" or "purposive" psychology based on the notion of instinct, McDougall later coming to Harvard College in the U.S.

Correlation

THE UNITED STATES

Recognized as the foremost U.S. psychologist of his time and a pioneer in the new scientific psychology, William James (1842–1910) taught physiological psychology with informal demonstration laboratories as early as 1875. He is known best for his *Ptinciples* of *Psychology* (1890). which offered views on personality types habits, stream of consciousness, and the James–Lange theory of emotions (see EMOTION). While James was at Harvard College, G Stanley Hall (1844–1924), who is believed to have earned the first U.S. Ph.D. in the "new psychology" under James, was at Clark University. Hall helped found the first U.S. psychological laboratory (1881), the *Amei ican Journal of Psychology* (1887), *Journal of Genetic Psychology* (1891), *Journal of Religious Psychology* (1904), and *Journal of Applied* P~chology(191.5)he was first president (and a founder) of the American Psychological Association. Hail, a genetic (evolutionary) psychologist deeply interested in human development and known best for his *Adolescence* (1904), is credited with writing the first text on the psychology of aging. *Senescence* (1922). Under the influence of Wundt, George Trumbull Ladd (1842–1921), a functionalist, published *Elements of Physiological Psychology* (1887), a standard text for many decades.

The next generation of "new" psychologists in the U.S. included James McKeen Cattell (1860–1944), James Mark Baldwin (1861–1934), Joseph Jastrow (1863–1944), British-born Edward Bradford Titchener, Eduard Wheeler Scripture (1864–1945), and German-born Hugo Munsterberg (1863–1916). Jastrow, a student of Hall's at Johns Hopkins University, seems to have been first in the world to receive a Ph.D. in psychology (1886). Another student of Wundt's, psychologist-philosopher Baldwin, founded a Canadian laboratory at Toronto (1889) and another at Princeton University (1893) and helped found the *Psychological Review* in-1894. He edited the first *Dictionaly of Philosoplzy and Psychology* (1901–05). Miinsterberg, another student of Wundt's, was brought to Harvard by James to teach experimental psychology, but his interests in law and business diverted him to become the first applied psychologist. Before settling at Columbia University, where he founded a laboratory, Cattell studied with Hall and Wundt and lectured at Cambridge University, where he became associated with Galton. In addition to being a founder (1921) of the Psychological Corporation for the promotion of professional psycholog ical services to industry and of such publications as *American Men of Science, The Directory of American Scholars,* and *Popular Science Monthly* (1900), Cattell became professor in the first independent department of psychology at the University of Pennsylvania in 1877; previously psychology had been the responsibility of philosophy departments. Interested in "mental tests" (a term he coined), he also investigated reaction time, association, perception, reading, psychophysics, and individual differences The leader of the structuralist school of psychology, Titchener promoted Wundt's tradition in the U.S., taking a post at Cornell University in 1892. Holding consciousness to be the only legitimate subject for psychology, he defended introspective psychology, concerned with the content of consciousness and the structure of subjective life. Titchener's influence, however, perished with his death.

Applied psychology

Functionalism as a school or movement sprang from the efforts of John Dewey (1859–1952) and of James Rowland Angell (1869–1949) at the University of Chicago and reached its height in the early 20th century. Dewey attacked the reflex-arc concept as an effort to reduce behaviour to "mere" neural events. By 1907 Dewey's student at the University of Michigan, Angell, propounded the functionalist's position in an influential article in 1907, "'The Province of Functional Psvchology." Functionalists focussed on the operations or functions of conscious activity (*e.g.*, thinking, learning), while structuralists studied so-called elements (*e.g.*, "ideas." "sensations") of consciousness. The successor of Angell as department chairman at Chicago and spokesman for functionalism was his student Harvey A. Carr (1873–1954), who introduced motivational topics (*e.g.*, striving) into functionalism. By the 1930s, interest in the functional-structural controversy had subsided, and a functionalist Robert S. Woodworth (1869–1962), spent 40 years at Columbia University developing "dynamic psychology." defined as the "study of cause and effect. motives and processes, or in the questions 'Why?' and 'How?' focussed on human activites and ach~eveinents.'Woodworth's work was a repudiation of Titchener and a reaction against McDougall and Watson. This "dean of American psychology" contributed the S–O–R formula (stimulus–organism–response), asserting that the older S–R formula (stimulus–response) failed to consider the responding organism itself: he observed that a given stimulus can produce a variety of responses, depending on the state of the organism. Edward L. Thorndike (1874–1949), Woodworth's colleague at Columbia. was a functionaiist interested in learning theory. He developed a psychology of learning called "connectionism," grounded on British associationism and on his own work with intelligent behaviour among laboratory acimals. Thorndike can be said to have established educational psychology as an autonomous discipline. A Johns Hopkins professor who had studied at Chicago under Angell, John B. Watson (1878–1958) launched U.S. "behaviourism" with his 1913 article, "Psychology as the Behaviorist Views It." Watson discarded consciousness. rejected introspection as a legitimate method of observation, replaced "mental" events with the study of behaviour, and regarded the only valid psychology as being physiological. Though Watson's behaviourism had major impact in the early 1920s in the U.S., it failed to gain wide acceptance elsewhere. Later forms of "neobehaviourism," especially the systems of Clark L. Hull (1884–1952) and B.F. Skinner (1904–), enjoyed wider popularity.

Educational psychology

Contemporary trends

DECLINING INFLUFNCE OF SCHOOLS OF PSYCHOLOGY

By World War II, schools of psychology (with few exceptions) had faded. Contributions to the common pool of psychological understanding were coming from many sources, no one school offering a sufficiently comprehensive account of the available evidence. As schools receded, their function passed to what may be termed miniature systems, models, limited theories, the three terms often being used interchangeabiy. Contemporary psychology is characterized by interest in theory, research, and experimentation. University courses specifically concentrate on personality theory, learning theory, theories of psychotherapy, and so on. While psychology in the 19th century gravitated toward Germany, by the 20th century, two-thirds of the world's psychologists were living in the U.S., with Canada and Britain playing important roles.

Contemporary learning theories still embrace older notions (functionalism, reflexology, gestalt, connectionism. and psychoanalysis; see LEARNING THEORIES). Personality theories commanding the attention of the psychological community include several based on learning. theories (see PERSONALITY, THEORIES OF),

GROWING STRESS OF QUANTIFICATION

The application of mathematics to psychology dates at least to Herbart in the early 19th century. In 1940 Clark

L. Hull applied mathematics to behaviour theory in a mathematico-deductive theory of rote learning, and there has since been a growing tendency to develop quantitative hypotheses accounting for human behaviour, beginning in 1950 with William K. Estes' (1919–) development of a statistical theory of learning based on stimulus sampling. The following year Robert R. Bush and Frederick Mosteller offered a mathematical or stochastic model for learning predicated on probability theory. Four years later, Frank Restle constructed a similar model for discrimination learning, and another, by Richard C. Atkinson, appeared in 1958. By 1963 R. Duncan Luce, Robert R. Bush, and Eugene H. Galanter had collaborated in assembling a three-volume handbook of mathematical psychology. In the same year, David Zeaman and Betty J. House developed a mathematical theory of attention based on studies of retarded individuals.

COOPERATION WITH ENGINEERING SCIENCE

Man–machine systems

Since the early 1940s human engineering (human factors psychology) has been applied to man–machine systems, the goal being to design equipment compatible with human structure and function. In 1948, with the publication of Norbert Wiener's ***Cybernetics: or Control and Communication in the Animal and the Machine*** and Claude E. Shannon's "Mathematical Theory of Communication," engineers and computers came to the service of psychology. Cybernetics, a term used by the Frenchman André-Marie Ampère (1775–1836) in 1834, was defined by Wiener as "the entire field of control and communication theory, whether in the machine or the animal." In addition to Wiener's views regarding the brain and computers and Shannon's information theory, computer technology grew, the electronic digital computer first being developed in 1946. The computer opened the way to machine simulation of complex psychological processes such as problem solving and perceiving. Information-processing (machine) theories of behaviour have been developed and applied to the interpretation of such human activities as learning, among the most influential being those of Herbert A. Simon (1916–) and his colleague A. Newell (1927–).

BURGEONING OF BIOPSYCHOLOGY

Behaviour genetics ("psychogenetics") treats behaviour as affected by heredity. In psychology there has been a growing interest in human genetic factors as they affect retardation and psychiatric disorder. Phenylketonuria (an inborn metabolic defect that can produce retardation) was discovered in 1934 in Oslo, Norway, by A. Fölling, stimulating efforts to advance biochemical genetics and to develop tests for carriers of the abnormal gene. In 1959 the French researchers Jérôme Lejeune, Gautier, and Turpin discovered the presence of 47 chromosomes (instead of the normal 46) in defective children with Down's syndrome (mongolism). Studying another form of retardation (Klinefelter's syndrome) in Britain, Patricia A. Jacobs and John A. Strong in 1959 reported an abnormal chromosomal count of 47. During the same year, Charles E. Ford and his associates discovered that in another variety of retardation (Turner's syndrome), bone-marrow cells contained only 45 chromosomes in the female sufferer studied.

Drugs

The so-called psychopharmacological revolution, beginning in the 1950s, ushered in the use of tranquillizers, antidepressives, nonbarbiturate sedatives, and nonhypnotic muscle relaxants. The term psychosomimetic refers to a drugged state mimicking a psychosis; hallucinogenic, psychedelic, and psycholvtic are more recent terms describing the effects of such-drugs as LSD and mescaline. Chlorpromazine (a phenothiazine chemical) and Rauwolfia alkaloid (reserpine) were among the first tranquillizers to come into general use in the early 1950s; the former, a French discovery in 1952, was used by Jean Delay of the University of Paris in treating psychotics. Reserpine, developed by Swiss researchers in 1952, has been used as a treatment in psychiatric disorders as well as in reducing high blood pressure. Drugs called monoamine oxidase inhibitors were introduced in 1951 for the treatment of severe depression and were followed by a number of other chemical compounds. Meprobamate, a minor tranquillizer used to treat anxiety, appeared in the early 1950s, followed by others such as chlordiazepoxide. The discovery of the effects of LSD (lysergic acid diethylamide) by Albert Hofmann (1906–) in 1943 in Switzerland occurred when he inadvertently swallowed the merest bit of it, with resultant hallucinations and confusion.

RISE OF PROFESSIONALISM

The multifarious nature of psychology in the world today has resulted from the rise of professional psychology since World War II. Academic psychologists, who once predominated, currently constitute a small minority. Most psychologists are distributed over many specialties, serving as industrial and personnel psychologists, school and guidance psychologists, clinical psychologists and psychotherapists, engineering and space psychologists, military psychologists, and social psychologists. They are employed in government agencies, business and industry, hospitals and clinics, private practice, prisons, and schools (see PSYCHOLOGY). By the 1970s, psychology had become a well-established discipline, respectably ranked among such other professions as the law and medicine.

BIBLIOGRAPHY. E.G. BORING, *A History of Experimental Psychology*, 2nd ed. (1950), and G. MURPHY, *Historical Introduction to Modern Psychology*, rev. ed. (1972), are two excellent histories to the mid-20th century. Current textbooks include: R. THOMSON, *The Pelican History of Psychology* (1968); D.P. SCHULTZ, *A History of Modern Psychology* (1969), a concise treatment limited to the period subsequent to Descartes; H. MISIAK and V.M. STAUDT, *History of Psychology: An Overview* (1966), another readable text similarly limited; R.I. WATSON, *The Great Psychologists: From Aristotle to Freud*, 3rd ed. (1971); and J.R. KANTOR, *The Scientific Evolution of Psychology*, 2 vol. (1963–69). W.S. SAHAKIAN (ed.), *History of Psychology: A Source Book in Systematic Psychology* (1968), outlines the historical landmarks. R.J. HERRNSTEIN and E.G. BORING (eds.), *A Source Book in the History of Psychology* (1965), is topically oriented but limited to experimental and quantitative psychology. F. ALEXANDER and S.T. SELESNICK, *The History of Psychiatry* (1966), and G. ZILBOORG and G.W. HENRY, *A History of Medical Psychology* (1941, reprinted 1967), are single-volume treatments of the history of psychiatry currently available in paperback. Other histories, limited in scope but well written, include: L.S. HEARNSHAW, *A Short History of British Psychology* (1964); A.A. ROBACK, *A History of American Psychology*, rev. ed. (1964); and J.C. FLUGEL, *A Hundred Years of Psychology, 1833–1963*, 3rd ed. rev. by D.J. WEST (1964).

(W.S.S.)

Psychology, Physiological

Introduction and historical background

Physiological psychology is the study of the physical basis of behaviour. Primarily it is concerned with how the brain and the rest of the nervous system function in activities (*e.g.*, thinking and perceiving) recognized as characteristic of man and other vertebrate animals. Since the nervous system critically depends on additional organs of the body for its normal function, physiological psychology is also concerned with mechanisms of metabolism, the production of hormones by endocrine (ductless) glands, and other regulatory effects on the internal bodily environment. Furthermore, the structural and functional characteristics of the nervous system, as well as of other organs, are mediated by mechanisms of heredity and are affected by diet, drugs, and disease. Thus, among all sufficiently evolved animals, behaviour may be influenced by all of these factors, since the nervous system for them is the major immediate (proximate) agent producing behaviour.

Although ancient philosophers poorly understood the nervous system, they usually wrote that psychological activities were in some way related to the body and its functions. Indeed, the historic roots of physiological psychology lie in classical mind–body notions of philosophy. Aristotle's view of this dispute, known as the double-aspect theory, defined mind and body as two aspects of the same thing, body serving as structure and mind being simply one of its functions. This surprisingly modern view failed to attract endorsement from many others, however. French philosopher Rent Descartes (1596–1650), for example, reserved soul as the realm of mind among human beings. He denied that mind existed for other animals and viewed their behaviour as being completely mechanistic, ascribing it entirely to the functioning brain and nerves. While he argued that basic reflexes in man were mechanistically determined, he depicted mind as a separate, spiritual (immaterial), uniquely human entity. To explain how two presumably distinct things called mind and body could influence each other, he proposed an interaction theory, holding that the physical site at which mind and body interacted was a structure in the brain called the pineal body (or gland).

The mind–body dispute

Even more independence of mind from body was envisaged by German philosopher Gottfried Wilhelm Leibniz (1646–1716) in his theory of psychophysical parallelism. Leibniz maintained that mental events and bodily changes occurred quite separately, holding them to be entirely independent but nevertheless paralleling each other perfectly, as though in accordance with a predetermined plan.

In modern history such behaviorists as U.S. psychologist John B. Watson (1878–1958) tried to carry the question one step beyond Descartes by saying that human beings were essentially the same as all other animals and by embracing another ancient philosophy that denied the existence of anything spiritual. They tried to strike all mentalistic terms—mind, consciousness, and feelings—from scientific use and directed attention only to observable behaviour, which they held to be the biologically determined outcome of the activity of the nervous system.

Although the term mind is more abstract than is readily observable overt behaviour, it remains a convenient word in ordinary conversation for designating the inner (private, subjective) aspects of human experience. Scientifically, however, mind cannot be used to refer to the nonphysical (since science is limited to phenomena; *i.e.*, to the physical), despite the great emphasis that philosophical and cultural heritage has placed upon the spiritual. Physiological psychology begins, then, with the basic understanding that if the word mind is to be used, it is to be conceived in such terms as the activities of the nervous system or other parts of the living body.

Basic properties of the nervous system. The anatomy and physiology of man's nervous system are given in detail in the article NERVOUS SYSTEM, HUMAN. Only a brief review will be given here.

Grossly, a normal vertebrate nervous system consists of the brain and spinal cord and their nerves. Stimuli from the environment prompt sensory nerves to deliver impulses to the spinal cord; motor nerves direct them from the cord to the muscles and glands. This provides the basis for such reflexes as the automatic withdrawal of a limb from a painful stimulus. In addition, incoming sensory impulses travel up to the brain, which in turn sends motor impulses down to the segments of the spinal cord. At various levels in the brain, incoming sensory nerves are linked by multitudinous connections to the outgoing motor nerves, making up a remarkably complex mechanism for integrating behavioral functions, similar in principle to the reflex mechanism.

The brain: structure and funcrions

To understand these connections and their importance in man's behaviour requires an appreciation of the gross structure of the human brain and its function (see Figures 1 and 2). The continuation of the spinal cord into the brain is called the medulla; in addition to major sensory and motor pathways and their integrating mechanisms, it contains many groupings of nerve cells that mediate such functions as breathing and the circulation of blood. Above the medulla is the midbrain, serving in many simple visual and auditory reflexes (*e.g.*, the pupillary reflex of the eye).

From K.L. Munn, Psychology (1951); Houghton Mifflin Company

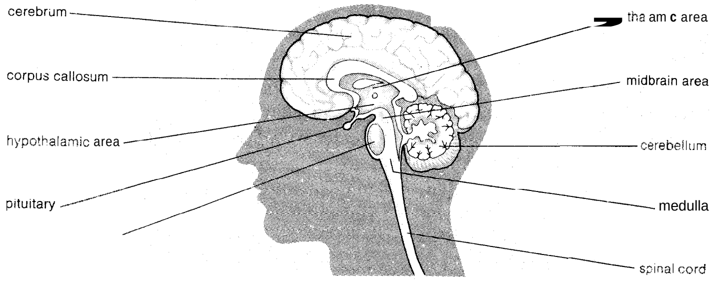


Figure 1: Internal structure of the human brain.

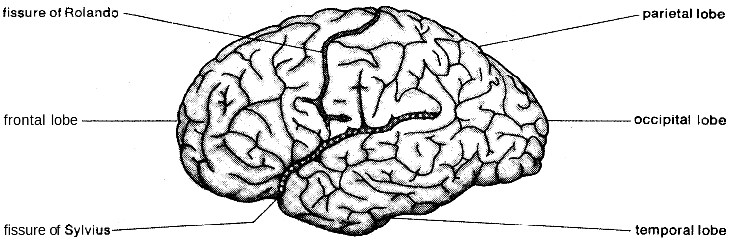


Figure 2: Major lobes and external surface of cortex in the human brain.

Behind the medulla and midbrain lies the cerebellum, acting primarily in the coordination of posture and in locomotion. The thalamus, situated above the midbrain, functions to integrate incoming sensory nerve impulses and relays sensory impulses to the cerebral cortex (the outer mantle of the brain). Below the ihalamus is a brain structure called the hypothalamus, concerned with control of the pituitary gland and with such complex functions as water balance in the body, thirst, sleep, food ingestion, reproduction, and emotional expression. Above the thalamus are the two cerebral hemispheres so prominent in man. The main mass of these hemispheres in humans consists of white matter entering and leaving outer layers of gray matter known as the cerebral cortex. Deeper parts of the cortex (the so-called old cortex) are parts of that structure that apparently evolved first and may be seen in many other animals as well as in man; it seems primarily concerned with vegetative (body regulating) and emotional functions. The new cortex, which evolved later, participates in sensory, motor, and such associative functions as learning. The new cortex includes occipital lobes (one at the back of each hemisphere) concerned with visual function, temporal lobes (at the lower sides of the brain) concerned with suditory and language functions, parietal lobes (behind the fissure of Rolando on each side) that integrate sensory information coming from the skin and muscles, and frontal lobes (in front of each fissure of Rolando) with motor areas that control discrete movements and postural adjustments; the remainder of each frontal lobe is presumed to serve complex associative functions still poorly understood.

In this greatly simplified picture of the human nervous system, sensory stimulation typically results in activation of integrative mechanisms at all levels, from the spinal cord to the cortex. The major sensory pathways described so far serve quite discrete functions; *e.g.*, some are specific for painful stimuli while others are specialized for tasting, still others selectively serve other parts of the body. This gives the nervous system considerable discriminative power. In addition, all sensory pathways send branches to a network of nerve cells (the reticular formation) lying in the centre of the hindbrain or brain stem (medulla and midbrain). The reticular formation relays nerve impulses to all parts of the brain and serves to activate or arouse the brain generally. It is held that this arousal function is crucial to wakefulness, attention, and emotion; reticular mechanisms may be thought of as preparing the brain for receiving impulses arising from stimuli.

Nerve cells

Much of the nervous system just described consists of billions of individual nerve cells (neurons), some of which are quite long, their tiny filaments extending all the way from the big toe to the spinal cord; others are not much longer than they are wide. These individual neurons connect with each other across tiny gaps between them (synapses). Connections may be such that the activity in one neuron will arouse a number of connecting cells, or activity in many cells may converge on a small number of others, or there may be reverberatory loops in which activity traverses a chain of neurons until the first cell is aroused again and again by the process it initiated.

The activity of neurons is detected as a nerve impulse: an electrochemical disturbance that may be propagated as rapidly as 120 metres (about 395 feet) per second. When these disturbances are recorded electrically with suitable amplification devices, the results show that if a nerve cell is to be activated at all, it must be stimulated at least at a minimal intensity (threshold energy). Once a nerve cell is activated, it discharges its electrical impulse fully; it responds in an all-or-none fashion.

Major integrative mechanisms of the nervous system depend on excitation (the arousal of activity) and on inhibition, in which the activity of a group of cells is arrested. Excitation and inhibition can be illustrated in spinal reflexes; touching a hot object with one's hand, for example, leads to excitation of nerves serving flexor muscles (that pull the limb away) and to simultaneous inhibition of nerves leading to opposing extensor muscles.

Sensing and perceiving. One of the oldest goals of physiological psychology is to explain how information (*e.g.,* a visual stimulus) is received from outside the body or from within the body itself (*e.g.*, as in a stomach ache). The aim is to specify physical bases for sensing and perceiving.

An early result of the endeavour was the doctrine of specific nerve energies offered by German physiologist Johannes Peter Miiller in 1826. While the anatomical and physiological details he specified were inaccurate, the basic notion still seems valid. The doctrine states that what is sensed (one's immediate awareness of stimulation) is determined by structures and events in the nervous system rather than by the quality of the stimulus itself. Thus a finger pressed on a closed eyelid mechanically produces visual experiences in the absence of the proper visual stimulus, light. Müller argued that people would see thunder and hear lightning if the nerves from the eye and the ear could be interchanged. Many phenomena support Muller's doctrine, and its general validity can be demonstrated most easily: an electrical stimulus will produce experiences appropriate to the sensory structure it stimulates, whether it be the eye, the skin, the tongue, or some point within the brain to which impulses from these organs go. Thus the experience one reports or the discriminative response he makes is specifically related to the part of the nervous system that is activated.

Hearing

To illustrate more fully: in hearing, the experience of pitch is determined largely by the frequency at which a sound source (*e.g.,* a tuning fork) vibrates—the faster the rate of vibration, the higher the pitch. Vibrations are transmitted through the air to the eardrum. The eardrum passes this vibration through tiny bones in the middle ear to vibrate the membrane of the oval window of the cochlea, a bony structure resembling a snail shell in the inner ear (see EAR AND HEARING, HUMAN for illustrations of these structures). This vibrates a fluid in the cochlea that, in turn, sets up a wave of vibration in the basilar membrane, which extends throughout the coiled length of the cochlea. The wave has its maximum effect on the basilar membrane at different places, depending on the frequency of vibration introduced. Low frequencies (relatively slow vibrations) affect the basilar membrane near the apex of the cochlea, while higher frequencies have their effect near the base. The basilar membrane is lined with the hair cells of the organ of Corti (crudely equivalent to the soft body of a snail coiled up in its shell), and as the membrane is moved, the shape of the hairs is distorted. Like crystals in an old-fashioned radio set that are distorted mechanically, the hairs initiate tiny electrical currents to stimulate the fibres of the auditory nerve .that arise from the vicinity of the hairs. In this way, different fibres of the auditory nerve are activated by vibrations produced by sound waves of different frequencies.

More than that, the different auditory nerve fibres are distinct (insulated) from each other in their course into the brain, at each of the neural relay centres through which they pass, and at the auditory region of the temporal lobe of the brain's cerebral cortex where they end. Hence investigators are able to record electrical activity from different parts of the auditory cortex, representing the different tonal frequencies that stimulate different parts of the basilar membrane in the inner ear. It is even feasible (in a cat, for example) to stimulate different parts of the basilar membrane directly with electrical current and to find the same arrangement of locations activated in the cerebral cortex. It would seem that the experience of pitch depends upon the place in the auditory nervous system that is activated by the stimulus.

Aside from pitch, other attributes of auditory experience are mediated by somewhat different mechanisms. Loudness depends on the intensity (*i.e.,* amplitude) of vibration, and it is held that the experience of loudness is a function of the density (or number per second) of nerve impulses conducted over the nerve fibres serving the specific frequencies associated with pitch experience. Density can be increased by discharging individual fibres more rapidly and by activating additional fibres of high threshold (*i.e.*, fibres that require a great deal of energy before they will discharge).

The temporal (time) pattern of pitches and loudnesses of sounds presumably is given by the temporal pattern of arousal of nerve fibres activated by a frequency- or intensity-patterned external stimulus. Finally, one's ability to localize sources of sound in space depends mainly on slight differences from one ear to another in the sound wave's time of arrival and on the intensity of vibration from a single source. These differences appear in electrical recordings from the auditory cortex to appropriately reflect differences in time of arrival and amplitude of nerve impulses on the two sides of the brain.

The same general principles apply to the other senses. Thus there are separate fibres and pathways in the spinal cord for carrying nerve impulses produced by stimuli that give rise to experiences of pain, temperature, and pressure in the skin. Different parts of the tongue are activated by sweet, sour, bitter, and salty stimuli; when the tongue is stimulated more intensely, the density of the patterns of nerve impulses increases proportionally. In vision different parts of the retina, the receptor surface of the eye, are sensitive to specific wavelengths of light and induce the experience of different colours. There appear to be different combinations of receptor cells in the retina that are maximally sensitive to different wavelengths in the visible spectrum (*i.e.*, the full range of the rainbow). The study of vision illustrates another property of sensory systems, for certain neurons may be found in the visual system that qualify as feature detectors. That is to say, they respond to some complex feature of the visual environment such as a line that tilts 45° to the left (but not the right or vertical or horizontal), or such as an edge or curve. Thus, some of our complex perceptions may be determined by the action of such feature detectors.

Vision

Physiological approaches to relatively complex perceptual problems can be illustrated with such visual examples as the awareness of real and apparent movement and ability to perceive depth and the location of objects in space. The visual localization of anything in view, for example, depends on the part of the retina stimulated and the manner in which that localization is represented in the brain. Philosophical interest in this problem arose when it was learned that the lens of the eye inverts the image of objects on the retina. It seemed difficult to understand how it was that people did not see the world upside down. Investigations were inspired into the role of learning early in life as related to the inversion of the visual field.

If the eye of a salamander is surgically rotated 180°, the animal will snap upward at a lure held below it and to the left at a lure held to its right. Even if the salamander's optic nerve is cut and allowed to grow back, the animal will not reorganize its visual behaviour in accordance with its experience and still will behave 180" out of phase upon recovery. The visual field of man can be rotated experimentally with special lenses. Although there are individual difference~some people seem to adapt to such visually reversed experiences and after a time will say that things look upright again. Thus, although visual localization is largely a function of the central (brain) connections of the retina, it seems also to be modified by experience (see LEARNING, PERCEPTUAL; PERCEPTION OF MOVEMENT).

The perceptual role of the brain has been studied in many experiments with laboratory animals and is shown through complex effects of human brain injury on visual processes. A man with damage to the region of the visual cortex may fail to recognize familiar objects for what they ale, although he may clearly be able to see them, to describe their appearance, or trace them with a finger. In other cases, such a brain-injured man may seem totally to ignore all events in a specific part of his visual field, although it can be shown by careful tests that he can indeed see in that part of the field; sometimes he may experience gross distortions of objects in the affected part of the field. The human brain (particularly the cerebral cortex) seems concerned with interpreting (perceiving) rensory information.

Motivated behaviour and the emotions. Problems of historic interest to physiological psychology concern the physical bases of feelings, urges, and emotions—a most difficult area of scientific inquiry (see EMOTION; MOTIVATION).

Drive (sometimes called instinct) has been defined as the arousal of the organism, observable as an increase in general activity or as the appearance of highly specific behaviour. Thus a hungry animal tends to become generally restless and hyperactive. Typically, however, any drive has a highly specific goal (such as food or a mate), or it might be expressed bodily as in sleep or flight or rage. Once the activity has been executed, the organism tends to become satiated; drive is reduced, hyperactivity ceases, and behaviour no longer is directed strongly toward the goal. The case of a three-year-old boy with an abnormal craving for salt is illustrative. This boy's history showed that his general appetite was poor and that he was restless and upset in his first years when fed an ordinary diet. He strongly preferred salty foods such as bacon or smoked fish, and these seemed to appease him a great deal; but this great craving did not become apparent until he discovered the saltshaker at the age of 18 months and ate salt by the spoonful. Belatedly taken to a hospital for observation, the child was mistakenly placed on a standard diet with only a normal amount of salt, and within seven days he died. Autopsy revealed extensive damage to his adrenal glands, leading to a disorder in which the body cannot retain salt normally. Apparently the boy had a deficiency that strongly drove him toward the specific goal of salt; when he gorged on salt, the drive was reduced; it was only through repeated cycles of drive and satiation that the boy managed to stay alive as long as he did.

Hypothalamic mechanisms

There seem to be interrelated brain functions centring in the hypothalamus: an excitatory process for the arousal of drive and an inhibitory mechanism for its reduction or for satiation. For example, experimental destruction on both sides of the brain of small areas in part of the hypothalamus of a rat. cat, or monkey results in a doubling or tripling of food intake, leading to great obesity; on the other hand, similar destruction of nearby regions of the hypothalamus results in refusal to eat and starvation in the presence of customary food supplies. That some parts of the hypothalamus are inhibitory is shown by experiments in which electrodes are surgically implanted into those regions of the brain. Electrical stimulation through the electrodes in a waking, hungry animal results in inhibition or depression of feeding. Conversely, the excitatory nature of other hypothalamic regions is shown when their electrical stimulation increases feeding activity.

Similar control mechanisms can be found in the hypothalamus in the cases of thirst, sexual behaviour, and sleep. It is clear that the hypothalamus also is directly influenced by sex hormones; mating behaviour may be aroused in animals by introducing small quantities of hormones directly into the brain. The cerebral cortex also contributes to the arousal of sexual behaviour, particularly among males; extensive damage to the cortex may eliminate the ability to show sexual arousal. That the animal is still capable of mating, however, quite clearly emerges when a massive dose of sex hormones distributed by the bloodstream throughout the body restores the mating response lost through injury to the brain cortex. Sexual behaviour does not seem to depend on any particular sensory stimulation, including impulses arising from the genitals (external sex organs); cutting the nerves from the genital region or surgical elimination of any one of the other sensory systems (*e.g.,* the visual apparatus) does not by itself block sexual behaviour. Among naïve rats, it takes elimination of no more than two sensory avenues at once to preclude sexual arousal; among more experienced animals, however, sexual behaviour may survive the elimination of three sensory systems, illustrating the role of learning and experience in the physiological control of sexual behaviour. It would seem that the mechanism of sexual behaviour must have gone through marked changes in evolution. In comparing animals from rat to man, it is found that the role played by sex hormones diminishes as dependence increases on sensory stimuli, on learning, and on the cerebral cortex.

Sexual behaviour

Much of the same physiological mechanisms that serve sexual function also operate in emotional expression. Cats surgically deprived of their cerebral cortex are easily aroused to apparent rage by almost any mild stimulus, but their response is short lived and poorly directed. If only the new cortex is removed, cats tend to become placid; if the old cortex is removed alone, they almost invariably become fierce. Monkeys, on the other hand, become strangely placid in the absence of selected portions of the old cortex, particularly those portions called the hippocampus and amygdala (see EMOTION for anatomic illustrations). In addition to placidity, the monkeys display exaggerated sexuality and markedly increased and indiscriminate oral activity, even to the point that some have been observed to put a snake or a lighted match into the mouth. Damage to different parts of the hypothalamus also can produce profound changes in emotional expression. Appropriate electrical stimulation of the hypothalamus in people will lead to such emotional responses as apparently unmotivated weeping and laughter.

Electrical stimulation of the hypothalamus and related brain structures under somewhat-different circumstances further illustrates physiological mechanisms that underlie motivation and emotion. For example, an animal may be allowed to press a lever that serves to electrically stimulate its own brain through appropriately implanted electrodes. Such animals have been observed to operate the lever as often as 5,000 times in one hour; under some conditions, pressing the lever seems to provide greater reward than do sexual activity or food; and it has been shown that the animals will readily walk across a painfully electrified grid repeatedly, apparently just to press the lever. When electrodes are placed in other parts of the brain, animals behave as if this kind of self-stimulation produces intense punishment rather than reward. Such experiments demonstrate the probable existence of selective brain mechanisms that serve all aspects of emotion and motivation, including the experiences called pleasure and anguish.

In general, such animals as men and rats seem to be aroused to strongly motivated behaviour and emotion as a result of physiological activity in an excitatory neurological mechanism centring in the hypothalamus. Contributing to this are major influences from sensory stimuli, often as modified by learning, and such internal factors as hormones, blood temperature, and levels of salt and other dietary substances.

Learning and intelligence. While many remarkable adaptations are accomplished through instinctive (or built-in) physiological mechanisms, a major factor in the adaptation of mammals (especially man) is an ability to learn. While they remain poorly understood, the mechanisms of learning and remembering seem to depend on relatively enduring changes in the nervous system. Physiological psychologists are especially concerned with discovering which structures in the body mediate learning, and they seek to understand the nature of neurological changes produced by learning.

In all kinds of learning the organism (from one-celled amoeba to man) makes new responses to some stimulus or general situation. Thus a dog can learn to salivate consistently in response to a bell, cats will learn to press a pedal that opens a door to a cage in which they are locked, and the English-speaking student of German learns to say "almond tree" when presented with the German term Mandelbaum. Aspects of the process of learning and remembering are described by so-called laws of association, some formulated by Aristotle himself: the laws of contiguity (in time and place) and repetition. (For example, one learns to associate ham with eggs; both tend to be contiguous in that they are apt to occur together at the same time and place.) Other such laws formulated later include those of effect (or reinforcement) and interference (leading to forgetting or extinction). In training dogs, for example, it has been found effective to present a bell together with meat (a normal stimulus for salivation); with this pairing, if repeated many times, the bell alone will begin to elicit salivation. The meat functions as a reinforcer for learning (in the sense that salivation tends to lessen when the bell is presented without the meat and to increase when the meat is included). The waning of salivation in response to the bell alone is called extinction and is similar to forgetting. Both extinction and forgetting appear to result from responses that interfere with those that have been learned; these interfering responses are favoured when reinforcement is omitted. The evidence is that the effects of most learning are stored over the lifetime of the individual and that forgetting ordinarily is a temporary failure to retrieve stored information through interference or competitive learning.

Some have held that even simple learning such as conditioning depends on the cerebral cortex, but it has been amply shown that animals still may be conditioned after the whole cortex has been removed from the brain by experimental surgery. In a complex learning situation, like a maze, however, an animal like a rat is heavily dependent upon its cerebral cortex for learning and remembering. In his classical studies, U.S. psychologist Karl S. Lashley showed that the larger the amount of cortex destroyed, the larger the impairment (principle of mass action). Since the same degree of impairment occurred regardless of the locus of the lesion, Lashley concluded that all parts of the cortex are equipotential for learning.

It once was held that the frontal lobes of the brain were critical in memory processes involved in problem solving. Indeed, monkeys without their frontal lobes tend to fail what is called the delayed-reaction test; for example, they are shown food placed under one of two identical cups and then made to wait several seconds to a minute or so before being allowed to uncover the food if they are able. Yet, some function other than memory seems to be involved in this defect, since monkeys without frontal lobes turn out to be quite distractible. If their attention is fixed firmly on the baited cup before the delay, monkeys with frontal lobes removed do succeed; if distractions are reduced by putting the animals into the dark during the delay period or by giving them sedatives, they also tend to perform successfully. The frontal cortex probably contributes more in the way of perceptual-attentive function than in storing the specific effects of learning.

The concept of mass action seems to have arisen from studies designed only to assess the effects of experimentally damaging brain tissue. By contrast, when the temporal lobes (at the sides of the cortex) of human beings are exposed for surgery while the individual remains awake on the operating table, it is found that specifically localized electrical stimulation can evoke in him dreamlike or hallucinatory experiences of familiar episodes from the past. (Mass action does not seem to apply.) When the temporal lobes are damaged in otherwise normal people, curious memory defects appear. While these people are able to recall past events very well, they typically are unable to remember what they have learned or experienced a few hours earlier. Study shows that they remember a definite event for about 15 minutes afterward; beyond that period, memory seems to fade and is lost. It is as though two memory processes have been separated by the temporal lobe damage: a temporary (short-term) memory process of very short retention seems to be spared by the injury, while more enduring storage of new information (long-term memory function) is impaired.

Short-term and long-term memory function

Support for a distinction between long-term and short-term memory comes from experiments with nonhuman animals, too. In one study, an octopus exposed to a crab that was dangled in the far end of its tank every two hours, six times a day, quickly learned to venture out of its shelter and to eat the bait. After a white card (that signalled an impending electric shock) was lowered from time to time with the crab, the octopus learned not to take the food when this cue was presented. Following removal of the vertical lobe of its brain, the octopus no longer could make the discrimination learned earlier; it kept emerging when the warning card was dangled. Even repeated retraining attempts every two hours with the card and electric shock proved ineffective. (This seems to contradict the notion of full equipotentiality.) But if training trials were closer together, the octopus was able to perform much better; within 15 minutes to an hour after the octopus was shocked on approaching the crab and the card, it stayed away from them. The animal's ability to retain information over the long term seemed to be impaired, even though short-term memory remained.

Rats subjected to electric shocks that produced convulsions after each daily learning trial also gave evidence of at least two kinds of memory. There were few signs of learning when shocks were administered within an hour after any trial, but if shocks came more than an hour later, the usual evidence of learning was observed. It seems to take some time for the effects of a learning experience to consolidate in the brain with some stability. For about the first hour afterward the storing of information can be disrupted by convulsive shock; after that the function is not easily disturbed.

Apparently the effects of learning first are retained in the brain by some reversible process, after which some more permanent neural change takes place. It has been suggested, therefore, that learning is mediated neurologically by at least two types of process as time passes. Perhaps the short-term function, temporary and reversible, reflects a physiological mechanism (*e.g.*, synaptic electrical or chemical change) that maintains a reverberatory loop (described above in discussing nerve cells) over a limited period to keep the memory trace alive. The ensuing, more permanent (long-term) storage may depend on an anatomical process in which nerve endings grow larger or increase in number so that synapses have enriched connections; perhaps new synapses are formed. Another speculative possibility is that long-term learning depends on changes in the chemical structure of neurons through alterations of peptides or complex protein substances such as ribonucleic acid (RNA) contained in the cells. Support for the inference comes from evidence that hereditary information is carried by chromosomes, coded in the structural arrangements of a related, protein-like molecule called deoxyribonucleic acid (DNA). While it has not been shown that learning and memory are similarly coded in RNA, the observation that memory in mice can be obliterated by injecting into the brain a drug (puromycin) that inhibits protein synthesis tends to endorse such a conclusion.

Since, however, other drugs (acetoxycycloheximide) can block protein synthesis in mice without impairing memory, it appears that some other aspect of puromycin is important in blocking memory. This is all the more impelling, for it turns out that the effect of puromycin can be reversed by subsequently injecting saline into the same regions of the brain as received puromycin. That is, memory lost through puromycin injection into the brain may be recovered by saline injection into the brain. The evidence is that a puromycin-peptide stays in the brain for long periods after intracranial injection and serves to inhibit the expression of memory rather than destroy it.

Since the neurophysiological reactions necessary for learning and memory seem to occur at the synapse, the chemistry of memory is apt to be the chemistry of synapses. It is here that we should look for the effects of drugs like puromycin.

Akin to the problem of learning are such relatively complex activities as reasoning, problem solving, and intelligent and linguistic behaviour. These have been studied

Reasoning, intelligence, and language

primarily in cases of brain damage, particularly in man, but to some extent also in other animals. Large amounts of human brain tissue may be destroyed without measurable impairment of these complex processes. For example, although removal of one entire cerebral hemisphere from a human being results in specific sensory and motor defects within the limits imposed by such impairments, intelligence may be effectively unaltered. On the other hand, damage to specific, restricted regions of the brain (particularly the cortex) may have devastating effects on intelligent behaviour. Following injury to the lateral surface of the frontal cerebral cortex, human-intelligence testing and delayed-reaction studies with monkeys may reveal severe impairment. In man a region fairly specific for speech functions is found on the side of the cortex (in almost all cases only on the left hemisphere). Damage there results in varying degrees of aphasia, an inability to name objects properly or to point to objects named by an examiner, even when the brain-injured person can see the object and may be able to say what it is used for. When parts of the temporal, parietal, or occipital cortex are damaged, there may be other language difficulties such as trouble in reading (dyslexia) or writing (dysgraphia).

Lesions in limited portions of the occipital cortex produce visual agnosia: once-familiar objects are no longer recognized when they are seen. With parietal-lobe lesions a man may have difficulty in recognizing parts of the environment, including his own body (to produce so-called distortions of the body image). In people with these injuries such skilled acts as putting on clothes cannot be coordinated, even though individual movements can be made normally; the disorder is called dyspraxia. In some cases of frontal-lobe damage a person may not be able to go through the motions of drinking from an empty glass, although he can easily oblige a request to drink from a glass with water in it. While such disorders are incompletely understood, it is clear that relatively small lesions of the brain can lead to major psychological impairment.

Personality and its disorders. While normal personality, as well as such disorders as neurosis and psychosis are to some degree the functional products of wholesome or stressful experiences, they also reflect physiological processes, some of which may be determined by heredity. In the 5th century BC, the Greek physician Hippocrates wrote that human temperament is largely determined by body fluids (humours). While his view was inaccurate in detail, many everyday examples emphasize the potent role of body chemistry in personality; consider the effects of alcohol, marihuana, and tranquillizers on behaviour.

These drugs affect the nervous system, as do vitamins in foods and hormones and enzymes produced in the body. The production of chemicals in the body that influence the function of the nervous system often is traceable to hereditary factors. For example, a striking variety of retardation in intelligence known as phenylketonuria (or phenylpyruvic oligophrenia) is transmitted through a gene that produces a defective enzyme in the body. The resultant inability of the sufferer properly to use specific kinds of protein from his diet in some way impairs brain function. A number of studies of the major psychoses implicate heredity as an important factor. When the psychosis called schizophrenia, for example, is diagnosed in one identical twin, it is most apt to occur in the other, even if the two have been reared apart; fraternal twins (those with no more similar heredity than ordinary brothers or sisters) are much less likely to exhibit symptoms of schizophrenia in common. Quite clearly heredity factors predispose the individual to some forms of psychosis; life experience, stress, disease, or other factors apparently contribute to precipitate the symptoms. It is suggested that the genetic predisposition is represented as a biochemical defect that affects the nervous system; the expectation that psychiatric symptoms may be controlled with substances that inhibit or excite brain chemistry is increasingly being fulfilled.

Genetic factors are less clear-cut in neuroses, and the disturbance usually is not so severe nor so incapacitating as in psychoses. Considerable experimental evidence, however, indicates that psychological stress and anxiety are characterized by widespread correlates in bodily activity; neurosis thus may be understood to involve altered physiological function. Indeed, presumably "mental" symptoms such as anxiety are physiologically controlled with tranquillizers. Experiments have shown that stress activates the pituitary gland, in part through the regulatory activity of the brain's hypothalamus. The pituitary, in turn, releases hormones that activate other endocrine structures of the body, particularly the adrenal cortex (see ENDOCRINE SYSTEM, HUMAN). Once stressed, this system may become increasingly sensitive, resulting in an oversecretion of the hormone adrenaline that seems to generate such psychosomatic symptoms as high blood pressure. Chronic overactivity of adrenal tissue may lead to exhaustion of its ability to secrete, perhaps contributing to psychosomatic symptoms of rheumatism and arthritis. It appears that prolonged psychological stress may affect the physiology of the organism and sensitize it to further emotional disturbances.

BIBLIOGRAPHY. J.A. and D. DEUTSCH, *Physiological Psychology* (1966); and PETER M. MILNER, *Physiologicnl Psychology* (1970), two up-to-date textbooks that give a thorough grounding in modern concepts of physiological psychology; V.G. DETHIER and E. STELLAR, *Animal Behavior,* 3rd. ed. (1970), a brief and concise modern treatment of the evolutionary and neurological basis of behaviour in invertebrates as well as vertebrates; W.H. THORPE, *Learning and Instinct in Animals* (1956); and N. TINBERGEN, *The Study of Instinct* (1951), extensive monographs representing the ethological approach to the study of physiological psychology; *Psychobiology: The Biological Bases of Behavior* (1967), an excellent selection of modern papers from *Scientific American;* J. FIELD (gen. ed.), *Handbook of Physiology,* sect. 1, *Neurophysiology,* 3 vol. (1959–68), an authoritative and extensive handbook covering the broad reaches of physiological psychology; E. STELLAR and J.M. SPRAGUE (eds.), *Progress in Physiological Psychology* (1966–), a continuing handbook of physiological psychology, up-to-date and authoritative, covering the full range of the field in the series.

(El.S.)

Psychology, Social

Social psychology is the scientific study of the behaviour of individuals in their social and cultural setting. Though the term may be taken to include the social activity of laboratory animals or those in the wild (see SOCIAL BEHAVIOUR, ANIMAL), the emphasis here will be on human social behaviour. The discipline of social psychology overlaps with other fields of study such as PERSONALITY, THEORIES OF, as well as with other disciplines such as COMMUNICATION; LINGUISTICS; SOCIOLOGY; and ANTHROPOLOGY.

Once a relatively speculative, intuitive enterprise, social psychology has become an active form of empirical investigation, the volume of research literature having risen rapidly after about 1925. Social psychologists now have a substantial volume of observation data covering a range of topics; the evidence remains loosely coordinated, however, and the field is beset by many different theories and conceptual schemes.

Early impetus in research came from the United States, and much work in other countries has followed U.S. tradition, though independent research efforts are being made elsewhere in the world. Social psychology is being actively pursued in the United Kingdom, Canada, Australia, Germany, The Netherlands, France, Belgium, Scandinavia, Japan, and the Soviet Union. Most social psychologists are members of university departments of psychology; others are in departments of sociology or work in such applied settings as industry and government. In the 1970s there were several thousand people employed as social psychologists in the U.S., about 200 in the U.K., and smaller numbers in other countries.

Scope of social psychology

Much research in social psychology has consisted of laboratory experiments on social behaviour, but this approach has been criticized in recent years as being too stultifying, artificial, and unrealistic. Much of the conceptual background of research in social psychology derives from other fields of psychology. While learning theory and psychoanalysis were once most influential, cognitive and linguistic approaches to research have become more popular; sociological contributions have also been influential.

Social psychologists are employed, or used as consultants, in setting up the social organization of businesses and psychiatric communities; some work to reduce racial conflict, to design mass communications (*e.g.*, advertising), and to advise on child rearing. They have helped in the treatment of mental patients and in the rehabilitation of convicts. Fundamental research in social psychology has been brought to the attention of the public through popular books and in the periodical press.

AKEAS OF STUDY

Research methods. Laboratory experiments, often using volunteer students as subjects, omit many features of daily social life. Such experiments also have been criticized as being subject to bias, since the experimenters themselves may influence the results. Research workers who are concerned with more realistic settings than with rigour tend to leave the laboratory to perform field studies, as do those who come from sociological traditions. Field research, however, also can be experimental, and the effectiveness of each approach may be enhanced by the use of the methods of the other.

Many colleges and universities have a social-psychology laboratory equipped with observation rooms permitting one-way vision of subjects. Sound and video recorders and other devices record ongoing social interaction; computing equipment and other paraphernalia may be employed for specific studies.

Social behaviour is understood to be the product of innate biological factors resulting from evolution and of cultural factors that have emerged in the course of history. Early writers (*e.g.*, William McDougall, a psychologist) emphasized instinctive roots of social behaviour (see INSTINCT). Later research and writing that tended to stress learning theory emphasized the influence of environmental factors in social behaviour. In the 1960s and 1970s field studies of nonhuman primates (such as baboons) drew attention to a number of similarities with human social behaviour, while research in cultural anthropology has shown that many features of human social behaviour are the same regardless of the culture studied. Human social behaviour now seems to have a biological basis and to reflect the operation of evolution as in the case of patterns of emotional expression and other nonverbal communication, the structure of language, and aspects of group behaviour.

Much research has been done on socialization (the process of learning from a culture), and learning has been found to interact with innate factors. An innate capacity for language, for example, makes it possible to learn a local language. Culture consists of patterns of behaviour and ways of organizing experience; it develops over the course of history as leaders and innovators introduce new elements, only some of which are retained. Many aspects of social behaviour can be partly accounted for in terms of their history.

Social perception. In some laboratory experiments, subjects watch stills or moving pictures, listen to tape recordings, or directly observe or interact with another person. Subjects may be asked to reveal their social perception of such persons on rating scales, to give free descriptions of them, or to respond evaluatively in other ways. Although such studies can produce results that do not correspond to those in real-life settings, they can provide useful information on the perception of personality, social role, emotions, and interpersonal attitudes or responses during ongoing social interaction.

The effect of cultural stereotypes

Research has been directed to how social perception is affected by cultural stereotypes (*e.g.*, racial prejudice), by inferences from different verbal and nonverbal cues, by the pattern of perceptual activity during social interaction, and by the general personality structure of the perceiver. The work has found practical application in the assessment of employees and of candidates for positions.

There has also been research on the ways in which perception of objects and people is affected by social factors such as culture and group membership. It has been shown, for example, how coins, colours, and other physical cues are categorized differently by people as a result of their group membership and of the categories provided by language. Other studies have shown the effect of group pressures on perception.

Interaction processes. The different verbal and nonverbal signals used in conversation have been studied, and the functions of such factors as gaze, gesture, and tone of voice are analyzed in social-interaction studies. Social interaction is thus seen to consist of closely related sequences of nonverbal signals and verbal utterances. Gaze has been found to perform several important functions. Laboratory and field studies have examined helping behaviour, imitation, friendship formation, and social interaction in psychotherapy.

Among the theoretical models developed to describe the nature of social behaviour, the stimulus–response model (in which every social act is seen as a response to the preceding act of another individual) has been generally found helpful but incomplete. Linguistic models that view social behaviour as being governed by principles analogous to the rules of a game or specifically to the grammar of a language have also attracted adherents. Others see social behaviour as a kind of motor skill that is goal-directed and modified by feedback (or learning), while other models have been based on the theory of games, which emphasizes the pursuit and exchange of rewards and has led to experiments based on laboratory games.

Small social groups. All small social groups do not function according to the same principles, and, indeed, modes of social activity vary for particular kinds of groups; *e.g.*, for families, groups of friends, work groups, and committees.

Early research on social groups

Earlier research was concerned with whether small groups did better than individuals at various tasks (*e.g.*, factory work), while later research has been directed more toward the study of interaction patterns among members of such groups. In the method known as sociometry, members nominate others (*e.g.*, as best friends) to yield measures of preference and rejection in groups. Others have studied the effects of democratic and aathoritarian leadership in groups and have greatly extended this work in industrial settings. In research on how people respond to group norms (*e.g.*, of morality or of behaviour), most conformity has been found to the norms of reference groups; *i.e.*, to such groups as families or close friends that are most important for people. The emergence and functioning of informal group hierarchies, the playing of social roles (*e.g.*, leader, follower, scapegoat), and cohesiveness (the level of attraction of members to the group) have all been extensively studied. Experiments have been done on processes of group problem solving and decision making, the social conditions that produce the best results, and the tendency for groups to make risky decisions (see ATTITUDES; PERSUASION). Statistical field studies of industrial work groups have sought the conditions for greatest production effectiveness and job satisfaction.

Social organizations. Such organizations as businesses and armies have been studied by social surveys, statistical field studies, field experiments, and laboratory experiments on replicas of their social hierarchies and communication networks. Although they yield the most direct evidence, field experiments present difficulties, since the leaders and members of such organizations may effectively resist the intervention of experimenters. Clearly, efforts to try out democratic methods in a dictatorship are likely to be severely punished. Investigators can study the effects of role conflict resulting from conflicting demands (*e.g.*, those from above and below) and topics such as communication patterns in social organizations. Researchers also have studied the sources of power and how it can be used and resisted. They consider the effectiveness of different organizational structures, studying variations in size, span of control, and the amount of power delegation and consultation. In factories, social psychologists study the effects of technology and the design of alternative work-flow systems. They investigate methods of bringing about organizational change; *e.g.*, in the direction of improving the social skills of people and introducing industrial democracy.

Changes in approach since **1900**

Ways of looking at working organizations have changed considerably since **1900.** Classical organization theory was criticized for its emphasis on social hierarchy, economic motivation, division of labour, and rigid and impersonal social relations. Later investigators emphasized the importance of flexibly organized groups, leadership skills, and job satisfaction based on less tangible rewards than salary alone. There has been a rather uneasy balance in the industrial social psychologist's concern with production and concern with people.

Personality. It is evident that there are individual differences in social behaviour; thus, people traditionally have been distinguished in terms of such personality traits as extroversion or dominance (see PERSONALITY, THEORIES OF). Some personality tests are used to predict how an individual is likely to behave in laboratory discussion groups, but usually the predictive efficiency is very small (see PERSONALITY, MEASUREMENT OF). Whether or not an individual becomes a leader of a group, for example, is found to depend very little on what such personality tests measure and more on his skills in handling the group task compared with the skills of others. Indeed, the same person may be a leader in some groups and a follower in others. Similar considerations apply to other aspects of social behaviour, such as conformity, persuasibility, and dependency. Although people usually perceive others as being consistent in exhibiting personality traits, the evidence indicates that each individual may behave very differently, depending on the social circumstances.

Socialization. The process by which personality is formed as the result of social influences is called socialization. Early research methods employed case studies of individuals and of individual societies (*e.g.*, primitive tribes). Later research has made statistical comparisons of numbers of persons or of different societies; differences in child-rearing methods from one society to another, for example, have been shown to be related to the subsequent behaviour of the infants when they become adults. Such statistical approaches are limited, since they fail to discern whether both the personality of the child and the child-rearing methods used by the parents are the result of inherited factors or whether the parents are affected by the behaviour of their children.

Problems in the process of socialization that have been studied by experimental methods include the analysis of mother–child interaction in infancy; the effects of parental patterns of behaviour on the development of intelligence, moral behaviour, mental health, delinquency, self-image, and other aspects of the personality of the child; the effects of birth order (*e.g.*, being the firstborn or second-born child) on the individual; and changes of personality during adolescence. Investigators have also studied the origins and functioning of achievement motivation and other social drives (*e.g.*, as measured with personality tests; see MOTIVATION).

Several theories have stimulated research into socialization; Freudian theory led to some of the earliest studies on such activities as oral and anal behaviour (*e.g.*, the effect of the toilet training of children on obsessional and other "anal" behaviour). Learning theory led to the study of the effects of rewards and punishments on simple social behaviour and was extended to more complex processes such as imitation and morality (*e.g.*, the analysis of conscience).

The self. Such concepts as self-esteem, self-image, and ego-involvement have been regarded by some social psychologists as useful, while others have regarded them as superfluous. There is a considerable amount of research on such topics as embarrassment and behaviour in front of audiences, in which self-image and self-esteem have been assessed by various self-rating methods. The origin of awareness of self has been studied in relation to the reactions of others and to the child's comparisons of himself with other children. Particular attention has been paid to the so-called identity crisis that is observed at various stages of life (*e.g.*, in adolescence) as the person struggles to discern the social role that best fits his self-concept.

Attitudes and beliefs. Research into the origins, dynamics, and changes of attitudes and beliefs has been carried out by laboratory experiments (studying relatively minor effects), by social surveys and other statistical field studies, by psychometric studies, and occasionally by field experiments. The origins of these socially important predispositions have been sought in the study of parental attitudes, group norms, social influence and propaganda, and in various aspects of personality. The influence of personality has been studied by correlating measured attitudes with individual personality traits and by clinical studies of cognitive and motivational processes; so-called authoritarian behaviour, for example, has been found to be deeply embedded in the personality of the individual. Early research based on statistical analyses of social attitudes revealed correlations with such factors as radicalism-conservatism. Later research on consistency provided extensive laboratory evidence of consistency but little evidence of it in actual political behaviour (*e.g.*, in attitudes on different political issues).

Factors in the origin of predispositionsof individuals

Research on attitude change has studied the effects of the mass media, the optimum design of persuasive messages, the effects of motivational arousal, and the role of opinion leaders (*e.g.*, teachers and ministers). Research has been carried out into the origins, functioning, and change of particular attitudes (*e.g.*, racial, international, political, and religious), each of which is affected by special factors. Attitudes toward racial minority groups, for example, are affected by social conditions, such as the local housing, employment, and the political situation; political attitudes are affected by social class and age; and religious attitudes and beliefs strongly reflect such factors as inner personality conflict.

EDUCATION IN SOCIAL PSYCHOLOGY

Undergraduate courses. Most social psychologists first take a bachelor's degree in psychology or in psychology and sociology. As a branch of psychology, social psychology typically constitutes between **10** and 25 percent of the course of studies (see PSYCHOLOGY). Other course titles that are particularly relevant to social psychology include personality, developmental psychology, animal behaviour, cognition, and psycholinguistics. Practical classwork includes a number of experiments carried out under supervision; students may also carry out an individual project, usually experimental, and on a modest scale. A smaller number of social-psychology students work toward degrees in sociology. Such students receive training in methods of field research rather than in laboratory work.

Graduate courses. Commonly, the graduate student is directed toward a Ph.D. There is often course work of a more advanced character than that for undergraduates; emphasis may be placed on studying journal articles, research monographs, and advanced handbooks, for example, instead of the less specialized undergraduate textbooks. There is an emphasis on the mastery of research techniques and the advanced use of statistics. A primary part of the training (for the Ph.D. especially) typically consists of writing a thesis or dissertation based on a series of original experimental or other studies. This work is done under the guidance of an experienced social psychologist and usually with the additional advice and evaluation of other senior members of the department.

THE PRACTICE OF SOCIAL PSYCHOLOGY

Teaching. Many social psychologists earn their living by teaching in psychology departments of universities and colleges. A smaller number teach in departments of sociology, social science, education, and management. Within psychology departments there may be subgroups or divisions; one division may embrace social psychologists (as well as those in related fields such as the study of personality), and there may be other groups of psychologists concerned with such fields as clinical, experimental, physiological, and comparative psychology. A number of universities have separate departments of social psychology. Social psychologists in psychology departments often maintain strong links with sociologists, as evidenced by the joint degrees offered by some schools.

Research. Most research is done by social psychologists working in university departments; *i.e.*, by university teachers and their graduate students, research assistants, and research workers on fellowships financed by the university or by outside grants from government research councils or from private foundations. There are also research institutes, often connected with universities, whose members do no teaching; well-known examples are the Institute for Social Research at the University of Michigan and the Tavistock Institute of Human Relations in London. These institutes are financed partly by grants from the government and foundations; partly they earn their way by acting as consultants for such clients as industrial firms.

Industry and commerce. A number of social psychologists work for research institutes or firms of industrial consultants who advise companies on organizational problems; *e.g.*, on the organizational or work-flow structure, incentive schemes, or personnel selection and training. Sometimes this service includes social-skills training (*e.g.*, in supervision and in personnel interviewing). One form of training offered (so-called T-group, or sensitivity, training) has aroused some controversy over its effectiveness. While many T-group trainees seem to improve their social skills, others have become emotionally disturbed. A number of social psychologists work in mass communication for advertising, market research, and similar organizations. They make use of research on attitude change and use techniques for ascertaining consumer "images" of products, services, and even people (*e.g.*, politicians) being presented to the public.

Sensitivity, or T-group, training

Government and military. An increasing number of governments throughout the world are employing social psychologists. They work in employment (selection, vocational guidance, training), social surveying, mass persuasion, criminology, mental health, and in the selection and training of civil servants. The military of many countries use social psychologists for officer selection and training and for research on leadership styles, reactions to the stress of combat, and the organization of crews of vessels and aircraft.

SPECIALTIES AND SUBSPECIALTIES

Propaganda and public opinion. Social psychologists are concerned with such aspects of public opinion (social survey) research as the design of standardized interviews and questionnaires. Forms of questions have been devised to compensate for errors that arise from the efforts to respond in a socially approved manner; some are designed to detect lying. Mass communications have been devised on the basis of research into persuasion (*q.v.*). Use is also still made of Freudian symbolism and theory.

Social prychiatry. Research into the causes of mental disorders has shown the importance of social factors in the family and elsewhere. Mental patients often show deficiencies in social performance that may be the cause of other symptoms. Many social psychologists hold that social factors may also apply to such disorders as schizophrenia, which also seem to have hereditary and chemical bases. There has been a corresponding growth in the use of various kinds of social therapy in psychiatry (*e.g.*, group therapy, therapeutic communities, and social-skills training; see also PSYCHIATRIC TREATMENT, CONCEPTS OF).

Industrial social psychology. Considerable research has been devoted to industrial productivity, absenteeism, labour turnover, accidents, and job satisfaction. Factors that have been found to be important include the style of supervision and management, the size and composition of working groups, the technology and the work-flow systems, the span of control, and other features of the organizational structure. Research results point strongly toward the advantages of a less rigid hierarchical structure of authority, with more delegation of authority and consultation, training in supervisory skills, small and cooperative work teams, and interesting and varied work.

Social-skills training. A major application of research in social interaction and group behaviour is in training in social skills, as in the T-groups, or sensitivity training, noted above. Role playing with video-tape playback and training in the imitation of other persons who serve as behavioral models are used in teaching people new skills. Actual training on the job has the advantage that there is no gap between the training and the work itself. All of these methods have been shown to be effective, depending on the job and the teacher. Social-skills training has been given successfully to industrial managers and supervisors, social workers and clergymen, interviewers, public speakers, mental patients, and juvenile delinquents.

Other fields. A great deal of research has been done on factors underlying racial prejudice, but the understanding thus obtained has not had much effect upon the social problems involved. Similarly, the causes of delinquency and crime have been extensively studied, but it is not feasible to manipulate the factors influencing crime, such as genetic factors, methods of upbringing, and inequalities of opportunity. Social psychology has made some contribution to education; sociometry is quite widely practiced as a means of grouping children, and evidence is growing about the optimum styles of teacher behaviour (see MENTAL HEALTH AND HYGIENE).

PROFESSIONAL ORGANIZATIONS

International organizations. Papers and symposia on social psychology are presented at meetings of the International Congress of Psychology and the International Congress of Applied Psychology, both of which meet in different countries every three years. The European Association for Experimental Social Psychology holds conferences every two or three years and arranges a number of smaller working parties and summer schools. Social psychologists from eastern Europe and Israel attend some of these meetings.

National organizations. A number of countries have social psychology sections or divisions of their national psychological associations. In the U.S. there is also the Society of Experimental Social Psychology and the Society for the Psychological Study of Social Issues.

Journals. Major English-language journals covering the field include the *Journal of Experimental Social Psychology* (U.S.), the *Journal of Personality and Social Psychology* (U.S.), *Sociometry* (U.S.), *Human Relations* (Great Britain), the *British Journal of Social and Clinical Psychology* (Great Britain), and the *European Journal of Social Psychology.*

BIBLIOGRAPHY. Excellent general textbooks include EDWIN P. HOLLANDER, *Principles and Methods of Social Psychology*, 2nd ed. (1971); and ROGER W. BROWN, *Social Psychology* (1965). A comprehensive account of research is G. LINDZEY and E. ARONSON (eds.), *Handbook of Social Psychology*, 2nd ed., 5 vol. (1968–69). A useful account of theories in social psychology is MARVIN E. SHAW and PHILIP R. CONSTANZA, *Theories of Social Psychology* (1970). Social psychology approached through detailed analysis of social interaction is described in MICHAEL ARGYLE, *Social Interaction* (1969). A. PAUL HARE, EDGAR F. BORGATTA, and ROBERT F. BALES (eds.), *Small Groups* (1965), is a most useful collection of papers. Research on social psychology in industry is described in BERNARD M. BASS, *Organizational Psychology* (1965). Social behaviour in relation to personality is dealt with in EDGAR F. BORGATTA and WILLIAM W. LAMBERT (eds.), *Handbook of Personality Theory and Research* (1968). *A* so-called symbolic interactionist approach is represented by a large volume of readings: GREGORY P. STONE and HARVEY A. FARBERMAN, *Social Psychology Through Symbolic Interaction* (19'70); and by ERVING GOFFMAN, *Relations in Public* (1971).

(M.Ar.)

Psychoneuroses

Psychoneuroses are patterns of behaviour classed as a major category of psychiatric disturbance; often preferably called neuroses or neurotic reactions. In general, neurotic behaviour is characterized first by its apparently defensive intent (of which the sufferer seems unaware) and by its commonly self-defeating consequences. The diagnosis of neurosis is made on the basis of any of a variety of symptoms, including exaggerated character traits, changes in mood, preoccupations, or untoward fears. Usually, neurotic people suffer from the loss of ordinary feelings of equanimity, happiness, satisfaction, and personal-social effectiveness. Neurotic needs nevertheless can contribute to achievement, and anxiety may serve socially constructive purposes. Indeed, such neurotic symptoms as depressive reactions or compulsive behaviour often are associated with superior intelligence.

Neurotic manifestations are widespread if not universal; the concern they arouse depends on their degree and quality rather than on their mere presence or absence. It is distinctive and critical that neurotic symptoms, such as an exaggerated fear of heights, cannot be banished by the sufferer's immediate effort.

General considerations

The term neurosis came into use at about mid-19th century when psychiatrists commonly held that the symptoms had their origins in neural (nervous-system) disturbances.

Neurosis and psychoneurosis

The prefix psycho- was added toward the end of that century to suggest that some neurotic symptoms had nonphysical, psychic ("mental"), or emotional origins. Over several decades it became popularly accepted that all neurotic patterns were based on psychic phenomena, and efforts to find gross physical (organic) bases in the nervous system were gradually abandoned. Distinctions between neurosis and psychoneuroses blurred, and the terms are now used interchangeably by most psychiatrists.

As compared with psychoses (*q.v.*), neuroses are characterized by minimal loss of contact with popularly accepted views of reality. The neurotic person generally recognizes that his feelings and reactions are inappropriate. He may not be able to recognize, however, that such physiological symptoms of neurosis as skin rashes have so-called mental bases; indeed, he may reject that possibility and seek to prove that they have organic origins.

Neurosis is regarded by most theorists as a form of maladaptation that reflects unconscious conflict (*e.g.*, between one's needs and the restrictions of society). Ordinarily less severe than psychosis, neurotic behaviour can lead to psychotic reactions; with sufficient improvement, neurotic symptoms may return.

GENERAL THEORIES OF THE ORIGINS OF NEUROSES

Anxiety and conflict

Nearly all modern authorities relate neurotic symptoms to anxiety, to attempts to avoid anxiety, and to internal (intrapsychic) conflict. Fear is a response to external danger (*e.g.*, a poisonous spider) and roughly reflects the degree of threat one perceives. While anxiety is a similar apprehensive feeling, it occurs in response to unknown, or at least unclear, hazard; *e.g.*, one is uneasy without knowing why. Anxiety is an experience common to mankind; being a terribly uncomfortable feeling, anxiety is generally avoided in every way possible, often through the development of neurotic manifestations.

While conflict may be conscious or unconscious, most theories of neurosis stress the latter. Conflict is said to occur between instinctual forces (which make up what is called id by psychoanalysts) and the inhibiting pressures of conscience (superego in psychoanalytic language) and the constraints of the environment.

According to psychoanalytic theory, the resolution of conflicts is sought through pressures brought to bear on the id and superego by the reality-oriented part of the personality called the ego, which tends to reconcile and compromise the opposing forces. Ego is said unconsciously to utilize so-called defense mechanisms in attempts to cope with conflict. Well-known ego-defense mechanisms have become part of ordinary language and include rationalization and compensation. The latter, for instance, is the process through which, many believe, one unwittingly attempts to make up for his perceived inadequacy by developing other attributes. Thus, in an aggressive person, aggressiveness may be a compensation for feelings of inferiority about his height or attractiveness or intelligence. There is a consensus that when ego defenses prove to be insufficient, so-called neurotic symptoms evolve as a further defensive endeavour by the person in conflict. Such dynamic or analytic views have been espoused by many psychiatric workers in the United States, with varying but lesser support in the United Kingdom, western Europe, and in most other parts of the world (where physiologic interpretations also enjoy support). In the Soviet Union the principles of learning or conditioning are widely invoked to account for psychiatric disorders including the neuroses. These stress the role of the nervous system in reflex conditioning. In this context, it is denied that neuroses constitute any form of "mental" illness; neurotic persons are not held to be "sick." Instead, they are said to have "learned" self-defeating habits that may be broken (rather than "cured") through retraining. Such theories explain the development of neurosis as direct responses to stress or as the result of a sensitization process in dealing with life problems. In general, these two theories, the psychoanalytic and the reflex-learning, are the dominant views in this field, with the former considerably more widely held in Western countries.

Neuroses as learned patterns of behaviour

PREVALENCE

Symptoms of neuroses appear to be universal or nearly so; most people at some time are likely to develop one form of neurotic manifestation or another. Symptoms warranting medical-psychiatric treatment, however, can be expected sooner or later in one out of every three or four persons.

Specific ethnic or national backgrounds do not seem to be strongly predisposing factors; neither is one's sex. A person's age, however, does seem to be related to neurotic tendencies since during one's life cycle there are specific times of added stress. Some of these in childhood include the death of a parent or other close associate, changing residence or school, puberty and the onset of menstruation, and such frightening events as sexual assaults. Emotional hazards later in life may stem from marriage, divorce, menopause, vocational and military experience, and senescence. The intensity of stress or conflict generally is increased at such times, and one's vulnerability to neurosis becomes accordingly greater.

GENERAL APPROACHES TO TREATMENT

In psychoanalytically oriented psychotherapeutic efforts to treat neuroses, the sufferer is expected to gain understanding (insight) into the psychic origins of his symptoms through talking and thinking about himself with the help of a therapist. Sufficient insight is expected to lead to enduring relief, when the patient finds that his symptoms are no longer necessary to defend against conflict and that he can surrender them in favour of more constructive activity. Psychotherapy more superficially also may include suggestion, advice, or hypnosis. A growingly popular approach to neurotic disorder is through retraining or reconditioning (sometimes called behaviour modification). A considerable number of physicians use drugs and other physical methods (occasionally electroshock) in managing a wide variety of neurotic symptoms, including anxiety and depression. Less technical efforts may be inspirational or religious, variously based on admonishment, encouragement, or even punishment.

Neurotic reactions

ANXIETY REACTIONS

Anxiety reactions are estimated to comprise 12 to 15 percent of the psychoneuroses seen in medical practice and may be conveniently distinguished as acute anxiety attack, anxiety-tension state, and anxiety neurosis. These distinctions are made to describe the extent, intensity, and duration (chronicity) of the sufferer's distress. In advanced countries, about 1 in every 400 people suffers significant disability from anxiety reactions.

Acute anxiety attack

The acute anxiety attack is a dramatic and intense episode in which the victim may show panic and be temporarily incapacitated. Typically appearing abruptly and without obvious warning, the attack is marked by signs of extreme apprehension or dread and by bodily symptoms such as sweating, pallor, and changes in heart rate. It is usually short-lived, the most intense manifestations continuing less than an hour. An acute anxiety attack may be an isolated event or, for particularly vulnerable peopie, a series of them can occur at varying intervals. Because of its dramatic nature and the briefly severe disability exhibited, it is not infrequently mistaken for a serious organic disorder such as a heart attack. Physical bases for such episodes may be presumed and actively sought, although psychiatric workers may hold the precipitating events to be psychic and unconscious and to be susceptible only to psychotherapy. The victim comes to fear such episodes and may even avoid places where he has experienced prior attacks. Acute anxiety attacks of some severity may be experienced by 1 to 3 percent or more of the general population at some time in their lives. While dread of subsequent attacks can be substantial, the predicted outcome (prognosis) is generally good; attacks will abate with no formal treatment.

By comparison, the anxiety-tension state is a less intense (subacute), more continuous anxiety reaction, although still considered to be limited in duration. Signs of anxiety, tension, and accompanying bodily changes seem to be precipitated by situational (environmental) stress and to be eased by reduction in stress; for example, anxiety-tension states are likely to be observed among soldiers in combat. When it is possible to remove the individual from battle or other environmental stress, the symptoms will not subside all at once; improvement may occur over several weeks or longer. As long as tension and anxiety continue, bodily functions may be disturbed (*e* g., digestive upsets, increased perspiration, headaches, pounding heart). 'There are wide differences among individuals; the same environmental stress that brings on subacute anxiety in one person may leave another relatively undisturbed and disable still another.

Anxiety neurosis

The anxiety neurosis is distinguished from other anxiety reactions by its relatively chronic nature. Otherwise. the sufferer reports the usual subjective experiences of anxiety and exhibits an array of typical bodily signs. Such secondary physical effects tend also to be subacute and prolonged (*e.g.*, chronic indigestion or high blood pressure). These can seem so dominant that the sufferer and the diagnostician may fail to recognize them as neurotic signs. The neurotic person himself, often unwittingly, directs attention away from any underlying anxiety.

Indeed, the clinical manifestations of anxiety neurosis are almost legion. Psychomotor symptoms include lestlessness, muscular tension, and a tendency to jump or be startled at the slightest unexpected stimulus. Perceptual-rational functions, particularly the ability to concentrate and to solve problems in logic, tend to be impaired. Physiologic signs include palpitation of the heart, weakness, headaches, and sexual impotence or frigidity; there may be evidence of emotional changes such as heightened irritability and decreased ability to show love or affection. Alterations in appetite can lead to obesity or underweight; insomnia is not infrequent. The person may describe what he calls feelings of fear, dread, panic, nervousness, apprehension, anxiety. or uneasiness; nevertheless, he cannot explain the real bases for these.

Anxiety reactions may be interpreted as arising through mechanisms of learning and reinforcement. In the more elaborate psychoanalytic interpretation of the anxiety reactions, psychic defenses are said to be mobilized against the experience of anxiety but to be poorly organized and relatively ineffective. It is suggested that the resulting symptoms may represent unconscious compromise, that while symbolism influences symptom "selection," its role is less prominent than in other types of neurosis. In general it is theorized that when some unwanted, forgotten (repressed) experience "buried in the unconscious" threatens to emerge into awareness, feelings of anxiety signal the hazard.

The prognosis in anxiety reaction is moderate to good; controlled studies reveal that roughly 70 percent of sufferers can improve even without treatment. When tranquillizers are used, more than 95 percent of these people secure some relief. The effectiveness of psychoanalysis or other forms of insight psychotherapy is endorsed by many authorities and challenged by others.

DISSOCIATIVE REACTIONS

Dissociative reactions comprise a large group of disorders that are generally classified with the neuroses. Dissociation generally may be understood as a process in which a limited aspect of human function becomes split off (dissociated) from the mainstream of activity. The concept of dissociation includes several reactions in which a portion of the personality seems to operate more or less independently; the consequences may include amnesia (in which certain aspects of the function of memory are said to have become dissociated).

The process of dissociation is widespread, and evidence for it seems to be uncovered in studying many neurotic people. Instances in which a dissociative process constitutes the most prominent feature of a neurosis are relatively rare, however; 2 to 4 percent of all cases of neuroses are estimated to be of this kind.

Sleep-walking

Normal dreaming can be considered dissociative since it occurs outside of ordinary consciousness, beyond voluntary control, and since it comprises activities that clearly are split off from one's usual awareness. Dreams can be disturbing or can indicate psychiatric disturbance. Somnambulism (sleepwalking) can be similarly viewed since it involves movements that are dissociated from the mainstream of consciousness.

Depersonalization (in some classifications considered to be a separate category of neurosis) is experienced as feelings of dissociation or unreality; the victim may report that his own experiences of the environment around him seem unreal. Such feelings are not uncommon and have long been recognized, but the term depersonalization came into widespread use only in the latter part of the 19th century. A related experience is that of déjà vu, in which one inaccurately feels that he has earlier had the same experience. One theory is that depersonalization reflects one's desire to escape from reality. This symptom complex may be induced with such drugs as mescaline and LSD.

Fainting is a temporary loss of consciousness; in some cases it map be interpreted as an unwitting neurotic dissociation from otherwise intolerable circumstances. Fainting of course has its physical basis, as in reduced blood circulation in the brain. There are different origins for amnesia; some forms can be induced by suggestion under hypnosis and others from gross injury such as brain concussion. Psychologic-emotional amnesia reflects the involuntary banishment of otherwise intolerably painful memories or experiences (see MEMORY, ABNORMALITIES OF).

Fugue

The fugue state is a major dissociative activity characterized by loss of memory and by flight from one's usual environment. The fugue may have physical origins (*e.g.*, in epilepsy) or may represent a drastic and involuntary neurotic escape from intolerable conflict. The victim may wander about, perhaps in a strange city or in other unfamiliar surroundings, sometimes in an aimless or confused fashion for days, weeks, or longer. On occasion he may carry on fairly well, unsuspected by others in his new role or identity. On recovery, memory for events of the fugue period is generally poor.

Rare instances of alternating or multiple personality occur in which one portion of the "mind" or "psyche" becomes dissociated and takes over to function independently, producing what seems to be a behaviorally different person. This can continue for months or years; instances of this extremely rare disorder have led to such literary classics as ***The*** Strange Case of Dr. *Jekyll* and *Mr. Hyde* (1886). Automatic writing, in which one's hand seems to write independently of his thoughts, can be induced through suggestion; in hypnosis of sufficiently suggestible people, any part of the "self" may be made to dissociate, take over, and act under the direction of the hypnotist.

PHOBIC REACTIONS

The word phobia comes from the Greek phobos, meaning dread, panic, or fear; among neuroses, phobias are specific fears out of proportion to the apparent stimulus. According to psychoanalytic theory, the painful feelings have become displaced from their internal and unconscious bases to become attached to a specific external situation or object. At any rate, a phobia is an obsessively persistent kind of unrealistic fear that seems inappropriate and unreasoning. It has become traditional to give phobic reactions names that indicate the object or situation feared. Theoretically the list of phobias could be endless; since some people fear specific numbers (*e.g.*, 3 or 7), the potential for specific number phobias (numerophobias) might be infinite. Some commonly encountered phobic reactions are listed in the accompanying Table, according to their technical names, objects, and derivations.

Common Phobic Reactions

Acrophobia	height (Greek *akra*, heights or summits)
Agoraphobia	open spaces (Greek *agora*, market place, the place of assembly)
Ailurophobia	cats (Greek *ailouros*, cat)
Anthophobia	flowers (Greek *anthos*, flower)
Anthropophobia	people (Greek *anthropos*, man, generically)
Aquaphobia	water (Latin *aqua*, water)
Astraphobia	lightning (Greek *asteropē*, lightning)
Bacteriophobia	bacteria (Greek *bactēria*, small rod)
Brontophobia	thunder (Greek *brontē*, thunder)
Claustrophobia	closed spaces (Latin *claustrum*, bar, bolt or lock)
Cynophobia	dogs (Greek *kynas*, dog)
Demonophobia	demons (Latin *daemon*, demon)
Equinophobia	horses (Latin *equinus*, horse, adj.)
Herpetophobia	lizards or reptiles (Greek *herpetos*, a creeping or crawling thing)
Mysophobia	dirt, germs, contamination (Greek *mysos*, uncleanliness of body or mind; abomination or defilement)
Numerophobia	a number or numbers (Latin *numeri*, numbers)
Nycotophobia	darkness or night (Greek *nyx*, night)
Ophidiophobia	snakes (Greek *ophidion*, snake or serpent)
Pyrophobia	fire (Greek *pyr*, fire)
Spatiophobia	self-confining phobia; phobically imposed area or spatial restrictions; often gradually, occasionally rapidly, progressive (English space)
Zoophobia	animals (Greek *zoos*, animal)

A psychiatric diagnosis of phobic reaction is likely to be made when one shows specific handicapping or self-defeating fear that is out of keeping with the danger as assessed by the diagnostician. While phobic attacks are not uncommon in childhood, many of these tend to subside spontaneously. Phobias can develop at any age, and the ease with which they respond to therapeutic efforts varies widely with their complexity and with the method of treatment used.

There is some support for the hypothesis that phobic behaviour can be learned. Encouraging reports of therapeutic success in eliminating phobias have come from practitioners and researchers who seek to help the sufferer unlearn his apparently illogical fears. It is not necessarily valuable for a claustrophobic, for example, to recall after months or years of effort, in analysis, that he once accidentally was locked in a closet when he was two years old and for him to say that he understands this traumatic event to have produced his phobia. Usually the phobia is lost following significant recall—abstraction along with real insight. In some cases, the chances of relieving his symptoms within a few weeks seem to be substantially enhanced if methods of behaviour modification (retraining, reconditioning) are used. In this particular case, the technique may be to have him relax comfortably in a large room at first and then to move him through a series of smaller and smaller rooms until he has learned to associate his feelings of relaxation with the same kinds of small spaces that once evoked feelings of panic.

Psychoanalytic theorists consider the phobic object as an external representation of unconscious bases for fear and that one's "selection" of a phobic object is an unwitting process determined through symbolism or circumstantial influences. Thus, a child who has attacks of fear when he sees white horses may eventually learn through analysis that the horses symbolize his father, for whom he holds an unconscious (but unacceptable) fear.

OBSESSIVE-COMPULSIVE REACTIONS

Obsessive-compulsive reactions are a group of neuroses marked by the repetitive intrusion of ideas, impulses, or behaviour that the sufferer finds consciously unwelcome. The subjective activity is called obsessive; the objective behaviour is compulsive. Freudian theorists write that these intrusions evolve unwittingly and with defensive intent in one's attempts to allay or to prevent anxiety, to which the obsessive person is very vulnerable. Psychoanalytically, the aim of the obsessive intrusion also can be viewed as that of maintaining or reinforcing repression. By contrast, behaviour theorists maintain that the reactions are undesirable habits that developed from unfortunate learning experience.

Kleptomania

Often included with the obsessive-compulsive reactions are socially undesirable manifestations referred to as impulsions, including exhibitionism (extreme attention seeking), kleptomania (compulsive stealing, for other reasons than the intrinsic value of that which is taken), pyromania (the compulsive setting of fires on neurotic bases), and voyeurism (peeping-tom behaviour).

Examples of obsessive disorders are common, comprising about 10 percent of the neuroses encountered by therapists. Some of these are so severe as substantially to cripple the person, but less serious obsessive-compulsive symptoms are widely observable throughout the world (*e.g.*, minor mannerisms such as ear pulling, eyebrow raising, and finicky cleanliness or dress).

Obsessively oriented people are more likely to be reserved, finicky, and conforming, exhibiting inhibition, indecision, control of hostility, and exaggerated feelings of guilt and responsibility; sexual maladjustments are frequent.

Obsessive activity can run the gamut from intrusive and unwelcome thinking about harming someone (*e.g.*, urges to injure a child or spouse) to preoccupation with musical themes or prayers or recurrent ideation (*e.g.*, profanity) that the neurotic individual finds to be personally repugnant. Compulsive behaviour can include repetitive hand washing and so-called compulsive security rituals ("making-sure" routines) such as repeated list making, door locking, switch checking, making amends, or counting supplies of money. What are called attitude-symptoms include perfectionism, obsessive interest at cleanliness, supermoralism, obsessively pervasive criticism of others, social withdrawal, and general inhibition of emotional expression. In seeking to protect himself from threat, the neurotic may become obsessed with keeping people at arm's length. Such a person is likely to have a low level of self-esteem and may manifest a preoccupation with the smallest details of his work.

Compulsive rituals

Obsessive predispositions of adults seem to originate in infancy and childhood and to stem from the vicissitudes of personal relationships with others. Frequently there is evidence of early parental overemphasis on control of expressions of anger and hostility, the child's behaviour being manipulated through his fears of rejection. Compulsive rituals among such individuals tend to become more complex than their antecedents (*e.g.*, the usual childhood rites of avoiding pavement cracks or touching every third fence post). The subsequent behaviour may include hair-brushings far beyond the number needed, repetitive telephoning, and compulsive talking that effectively keeps others from saying any words that might provoke anxiety; it is as though the obsessive-compulsive person can never really feel secure.

When obsessive-compulsive symptoms are generalized and complex, response to any form of psychotherapy (*e.g.*, behaviour modification or psychoanalytic insight) can be slow and difficult. While the persistence, concern with detail, and conscientiousness of such persons are assets in any therapeutic endeavour, their inhibitions, restrictions of emotionality, and continuing security rituals delay and slow the process.

DEPRESSIVE REACTIONS

Depression as a major symptom of neurosis refers to lowered spirits, undue sadness, dejection, or melancholy.

The depressed mood appears out of proportion to any loss or injury as objectively evaluated (see EMOTION).

Suicide

Frequently observed in some degree in many psychotic disorders, depression also can appear in neurotic syndromes as the most prominent symptom. There is a relatively high frequency of neurotic depression among the more sophisticated, educated, mature, and intellectually favocred groups of people, the disturbance tending to be most common in middle age. One consequence is that the rate of suicide in Western countries is greatest between the ages of 35 and 75 years, with a peak at about 55. Depressive reactions are estimated to comprise about 14 to 18 percent of instances of neuroses observed in psychiatric practice

Symptoms include lack of ambition, withdrawal, decrease in interests, restricted social activity and interests, a pessimistic or even hopeless outlook, and bodily accompaniments of low spirits; the latter can include fatigue, aches and pains, constipation, and appetite and weight loss. Coldness of the hands and feet is not infrequent, as are difficulties with sleep (especially dawn insomnia, so termed because of a difficulty in -remaining asleep as dawn approaches). All types of emotional, intellectual, motor and physiologic activity tend to be slowed as the severity of depressive reaction increases.

Some theorists interpret neurotic depression as arising from an inabiiity consciously to cope with hostility, sometimes describing it as a "frozen state of rage." The dynamics of such depression are held not to be simple or superficial matters, and their working out during insight therapy is seen to require a great deal of patience and determination on the part of sufferer and therapist alike.

Some forms of depressive reaction are seemingly paradoxical; one type can follow the achievement of a long-sought goal and has accordingly been named the depression of success; the so-called promotion depression occasionally follows a long-awaited advancement. There are also situational or reactive depressions that follow clear-cut disappointment or loss. Anxious depression apparently arises when the defensively evolved and unwitting symptom of melancholy is not sufficient protection against anxiety.

intensive psychotherapy in neurotic depressions often is followed by improvement, although the process is slow. Suicide is a hazard in severe depressive reactions. More rapid response is sometimes produced with mood-elevating drugs and in some instances of psychotic depressions with electroshock therapy.

CHARACTER REACTIONS

Character neurosis is a reaction in which normal personality traits known as character defenses become exaggerated to the point of personal handicap and self-defeat. Their evolvement is an unwitting defensive process each person undergoes in his personality growth. When such normal character traits become overdeveloped, they may not be readily apparent but are considered roughly equivalent to the overt symptoms (*e.g.*, obsessive or depressive) evolved in other types of neurosis (sometimes termed symptom neuroses because the symptoms are the clinically prominent manifestations).

Manifestations of the overdevelopment of any character trait are often largely subjective and may become apparent only through therapeutic study. Major categories of character neurosis (constellations of exaggerated personality traits) include the obsessive personality, the conversion personality, and the depressive personality.

The obsessive personality is characterized by such traits as orderliness, restricted emotionality, resistance to change, meticulousness, ovirconscientiousness, tendency to worry over responsibilities, self-doubts and feelings of inadequacy, overconcern with detail, intellectualization, pessimism, and sexual inhibitions. His intelligence is likely to be high, and he well may have superior capabilities for a particular business, artistic, or professional pursuit.

Needs for attention

The conversion personality is marked by strong, shifting emotional activity, susceptibility to suggestion, tendencies toward dependency and helplessness, and strong needs for attention. Mature sexual adjustment rarely is achieved, but such a person is likely to be responsive, appealing, warm, imaginative, and flexible. While his mood may swing rapidly, he is in good spirits frequently; he is likely to be more buoyant than are persons with other types of character neurosis.

The depressive personality is marked by overseriousness. by a restricted sense of humour, and often by low spirits. He tends to be studious, overconscientious, and has unusual vulnerability to disappointment, frustration. or loss. He tends to be highly dependable, responds poorly to hostility, but has an excellent capacity for work, application, and long-term endeavour. Readily assuming responsibility, he tends to be overly straitlaced, and his levels of intelligence, ability, and potential for achievement typically are high.

Efforts to treat character reactions psychoanalytically may involve so-called character analysis, in which inquiry and study of the individual's personality traits or defenses are undertaken. The therapist and the neurotic person together seek to elucidate and resolve the sufferer's conflicts that contribute to his elaborate character defenses. This comprises a personal educational process of some depth that some authorities feel can be useful for nearly anyone. The personality traits of the individual are enlisted in treatment; his persistence, application, and conscientiousness are applied to his work and his collaboration in the therapeutic process.

Estimates are that character reactions comprise from 16 to 20 percent of all neuroses. Their recognition is hampered by the relative absence of obvious symptoms as encountered in symptom neuroses. Their pervasiveness and their profound effects on the life, satisfactions, and happiness of the individual concerned nevertheless make their recognition and therapy of considerable moment.

FATIGUE REACTIONS: NEURASTHENIA

Emotional fatigue refers to weariness produced by psychic factors rather than from such activities as muscular exertion. It arises from continued emotional stress and tension that stem from inner conflicts. Evidence for emotional fatigue is to be found in a discrepancy between how tired a person feels and how much effort he has been exerting. If one is in good health, has been getting enough rest, and still complains of marked weariness, this provides presumptive evidence of the fatigue reaction. Such neurotic reactions, in which the most prominent feature is emotional fatigue, once often were associated with the term neurasthenia.

Neurasthenia, as described in the latter half of the 19th century, was a widely inclusive syndrome: since then its popularity as a diagnostic term has fluctuated widely. For decades neurasthenia was a catchall for many neurotic and even psychotic manifestations. Nearly abandoned in most quarters, this term was again gaining a more specific definition and growing currency by the 1970s.

Features of neurasthenia include dependency, self-absorption, preoccupation with physical problems, a tendency toward regression, and attenuated interaction with others. Emotional features include difficulty in concentration, irritability and tension, problems with memory, detachment, feelings of inferiority, and lowered spirits. Impairment of sexual function is prominent. Physical features of neurasthenia are especially likely to include fatigue, feelings of weakness, headache, and insomnia; shifting bodily symptoms can be indigestion, heartburn, loss of appetite, rapid heart rate, and varicus aches.

Sexual origins

Neurasthenia once was attributed to sexual origins (*e.g.*, frustration and masturbation), views held by many psychiatrists into the 20th century. It is now wideiy agreed that these manifestations have many other bases as well. The consensus is that sexual difficulties often are secondary to otherwise hidden conflicts that also underlie such symptoms as weakness and fatigue. Mood-elevating drugs are widely employed in managing neurotic disorders of this kind. Psychotherapeutic measures include efforts to help the person cope with feelings of hostility, to deal with unconscious dependency needs, and to selectively maintain repression of conflicts.

While chemical therapies are useful, the effectiveness of psychotherapy in neurasthenia or fatigue reaction depends on hcw chronic the symptoms have been and on the person's motivation and ability to cooperate. Emotional fatigue is frequently encountered in ordinary medical practice and may yield to drug therapy; at any rate, any psychologic basis often goes unrecognized. Instances of major neurotic disordei in which emotional fatigue is the outstanding feature are far less freouent, an estimated 3 to 4 pcrcent of neuroses being called fatigue reaction; the figure is somewhat higher when use of the older and broader category of neurasthenia is included.

HYPOCHONDRIACAL REACTIONS

The term hypochondriasis is derived from a Greek word meaning under or below the rib cage, a frequent site for symptoms of the disorder. The reaction is marked by a persistent overconcern with health, in an obsessive kind of preoccupation with bodily functions. Uncomfortable feelings may appear anywhere in the body and may shift from one part to another. Observable signs of physiologic disorder or disease are insufficient to account for the person's reports of pain and discomfort or for the !eve! of his concern. An alternative term, hygeiaphrontis (hygeia, "health," and *phrontis*, "anxious concern about," both from the Greek), more accurately describes the symptoms.

Reactions that fit well into the category of hypochondriasis (or hygeiaphrontis) include about 3 to 5 percent of the major neurotic complaints encountered. Preoccupation with health can become so neurotically pervasive in hypochondriasis that the person suffers considerable pain or other discomforts. Unfortunately he not infrequently is certain that organic disease is present (although these is none), an often troubling barrier to initiating psychotherapy. A wide variety of discomforts can occur on a hygeiaphrontic basis, sometimes simply representing an increased awareness of normal function. At other times, functional change (*e.g.*, indigestion) car occur through stress or conflict, sometimes with implications of symbolism. Thus, "I am sick" can have figurative meanings; one may feel "sick" of his work situation, of his marriage, or of another relationship. At times the figurative csn carry over into the literal expression of illness. This is reflected in many other common figures of speech, such as "He gives me a pain in the neck," "That situation was a headache," or "I can't stomach that professor "

Hygeiaphrontic manifestations are rarely seen in conjunction with serious organic disease, but accompany many instances of emotional disturbance. The diagnosis is best made on positive evidence of psychic origins rather than by exclusion of organic illness. The prognosis for treatment is most variable; results are sometimes very good. In many other cases, despite the best efforts of experienced therapists, hypochondriacal symptoms may persist until the sufferer's last days.

CONVERSION REACTIONS: HYSTERIA

In conversion reactions or hysteria, elements of an inner, emotional conflict and its consequences are converted into external expression through bodily, physiologic mechanisms in the form of many eymptoms the most severe of which con include blrndness and paralysis. Since symptoms of this sort can serionsly hamper daily activities, conversion reactions were among the earliest neurotic disorders to be so recognized and enjoy a position of historical precedence.

Ancient Greek medical writings state that such symptoms derive from the wanderings of the uterus (womb) within the body; their word for this female organ (*hystera*) provides the root for the term hysteria. Widely held contemporary accounts of the origins of hysteria are apt to draw on psychoanalytic theory and on modern understanding of how brain function affects activities in other parts of the body. Psychoanalysts posit that an ego-defense mechanism (an unconscious process they call conversion) acts to produce the signs of hysteria as disguised expressions of intrapsychic conflicts that otherwise would give rise to anxiety. Consciously disowned ideas and associated emotions, plus the psychologic defenses against them, are said to be transmuted or converted so as to secure a measure of outward expression in bodily or sensory symptoms. Psychoanalytically, the conversion reaction is understood generally to symbolize factors involved in repression and to exact a certain toll of self-punishment (to atone for the repressed. unacceptable needs or drives). Theorists of a more heavily neurologic orientation are likely to replace such terms as intrapsychic conflict and repression with the more mechanistic language of biochemistry and physiology, attributing hysteria to physical changes within the brain. Neither the neurologic nor the psychoanalytic view excludes the other, however: indeed, they may prove to be different ways of talking about the same set of phenomena.

Conversion reactions may be divided into those that are somatic (bodily) and those that are functional or physiologic, the latter including or overlapping with psychosomatic and somatization reactions. With the two grouped together, conversion accounts for an estimated 15 percent or more of neurotic reactions encountered, physiologic conversions far outnumbering the somatic. Younger and less sophisticated people are most subject to conversion reactions; indeed, perhaps gross somatic conversions are less frequent than they once were because their underlying emotional bases have become more clearly understood by ordinary people.

Neurotic deafness

Symptoms in hysteria or conversion reactions can simulate those of almost any kind of organic illness. Conversion pain is not infrequent and can affect any part or region of the body; major symptoms also can include deafness, spasms, cramps, tics, losses of tactual sensation, and othei behavioral disturbances. Despite the severe disabilities and limitations that conversion symptoms can produce, the attitude of persons so afflicted can seem paradoxically indifferent. Almost characteristic of somatic conversion, this feature is called in French *la* belle *indifférence* ("the beautiful indifference"); it is as if the symptom is protecting the person from experiencing anxiety. He is quite indifferent to its extent, limitations, the sacrifices entailed, and handicap.

It is most distinctive of hysteria that such symptoms as blindness and paralysis can be at least temporarily abolished by suggestion while the sufferer is under hypnosis (*q.v.*). The prognosis with educationally oriented psychotherapy is generally good in young people whose conversion symptoms are less fixed and chronic and when they are not the sole neurotic manifestation. Encouraging reports have been published concerning the effectiveness of behaviour modification or re-education (*e.g.*, with rewards or punishments such as electric shocks). Additional more superficial treatment methods that have been used with varying efficacy include various types of direct suggestion without hypnosis, reassurance, threats or trrckery, and supportive measures.

RELATED DISORDERS

Neuroses following trauma (or harmful experience). Alternatively termed traumatic reactions or traumatic neuroses, these are disorders that show a distinct relationship to such trauma as a bodily injury or an emotionally stressful experience.

Responses to a given harmful experience vary widely from person to person; subsequent emotional effects are often unpredictable. Sometimes, limping may persist years after a broken ankle has mended. If the disability is functionally disproportionate to the structural impairment, it represents a distinguishing feature of traumatic neurosis. In a broad sense all emotional disorders might be regarded as ultimately traumatic in origin. Symptoms follow the pattern of other neurotic reactions, the most frequent being those of the anxiety, conversion, and hypochondriacal patterns. A distinguishing feature is the sharpness of the precipitating event (trauma), without which it is believed the neurotic reaction would not have occurred. It is widely held that pre-existing psychic factors influence the form and character of symptoms in the traumatic reactions as in any type of neurosis.

Typically in such cases an injury is followed within a period of days or months by the onset of a variety of symptoms that are out of proportion to what one might ordinarily expect from the injury alone. The neurotic victim himself is usually certain that a direct cause-and-effect relationship exists between the injury and his symptoms. Since such behaviour is even more frequently encountered among victims of industrial accidents, the potential economic gains for the victim may lead some to confuse neuroses with malingering (feigned illness).

The prognosis in traumatic reactions varies widely and is related to the chronicity, fixation, and duration of the symptoms, as well as to their meaning to the individual who suffers them. Psychotherapy is not invariably beneficial, especially when the sufferer fails to cooperate actively.

Soterial reactions. Essentially the converse of phobic reactions, in soterial behaviour an external object, situation, or person becomes the source of feelings of comfort, security, and well-being that seem objectively out of proportion. An adult, for example, may habitually go to sleep clutching the teddy bear he first owned as a child. The soterial object (from the Greek *soter,* "saviour") counters fear and anxiety by stimulating feelings of salvation, solace, preservation, and safety.

Since soterial objects help to abort experiences of dread or threat, they are rarely a basis for complaint; soterial persons seldom seek assistance and are apt to come to the therapist's attention only when their behaviour is a source of distress to others.

The "choice" of a soterial object is practically unlimited; examples include money, special foods, people, locales (*e.g.*, the physician's office), rituals and procedures, prayers, religious objects, talismans, and drugs. Compensation pensions and even symptoms of disability or illness can secure soterial significance. Deprivation of a soterial object provokes resistance and anxiety; the neurotic individual who uses milk as a security food, for example, may become frantic without it.

In the treatment of soterial reactions, symptom resolution is readily obtained through reconditioning and other forms of behaviour modification. Therapists of a psychoanalytic bent interpret remission of symptoms as a limited goal and seek to trace the unconscious antecedents of the symptoms.

Military reactions. While various types of neurotic reactions are evolved under the stresses of military conditions, the term military reactions delineates those that develop during operational (*e.g.*, combat) situations. Alternative designations include combat fatigue, combat exhaustion, effort syndrome, acute situational reaction, and operational fatigue. In World War I, shell shock was a prominent diagnosis, based on the assumption that the shock of exploding ammunition affected the nervous system.

Shell
shock

What one man tolerates well under military stress can prove overwhelming for another. In acute combat reactions, prominent symptoms include psychomotor expressions of anxiety (*e.g.*, tremors), sleep disturbances, restlessness, exaggerated tendency to startle, and such bodily manifestations as heart palpitation and increased perspiration. Conversion symptoms such as headache, fatigue, backaches, and functional cardiovascular and gastrointestinal symptoms are frequent. There may also be signs of depression, guilt feelings, withdrawal, apathy, and scattered amnesia. The battle dream is a frequent and outstanding feature; irritability is quite common and sometimes leads to destructive behaviour.

Manifestations of military reactions gradually subside following the victim's removal from stress. If stress has been particularly intense or prolonged, remission of symptoms can be long delayed. Factors that influence susceptibility to combat stress include the individual's prior adjustment and emotional stability, the adequacy of his military training and his leaders, the level of his morale, and the degree to which he and his comrades suffer exhaustion or defeat. Treatment in acute combat and military reactions above all demands removal from the stress of war. Beyond this, the provision of adequate rest, reassurance, food, sleep, shelter, supportive psychotherapy, and judicious sedation can be quite beneficial.

Additional terms and concepts. ***Psychophysiologic disorder.*** A more recent term, psychophysiologic disorder describes reactions marked by the morbid bodily expression of emotions. It is akin to psychosomatic illness, somatization reaction, and physiologic conversion, all of which label significant overlapping concepts that stress the wide potential for bodily and visceral expression of psychological disturbance.

Somatization reactions. This term also covers functional and physiological symptoms held to evolve as expressions of emotional conflict. The concept of somatization stresses the principle that any type of functional disorder of the body can be based upon or contributed to by emotional stress and tension. Subtle, progressive functional changes stemming from emotional factors can produce structural alternations such as skin rashes, as in hives. This inimical sequence can progress to organic pathology sufficient to require medical intervention (*e.g.*, skin lesions that may be managed with hormone salves or creams).

Psychosomatic illness. This related diagnostic label emphasizes the interaction of physical (somatic) and emotional (psychologic) aspects in generating serious physical illness; noteworthy examples include asthma and allergic disorders, gastric ulcers, forms of high blood pressure, some heart attacks, endocrine gland disorders, inflammation and ulceration of the colon, and arthritis.

Asthma

Impulsive reactions. This diagnostic term is akin to the character or personality disorders, patterns of responses that are in varying degree socially disapproved and that are accompanied by minimal evidence of anxiety. The more common impulsions include exhibitionism, kleptomania, pyromania, and voyeurism. At times impulsive gambling, drug addictions, and dipsomania (craving for alcohol) are so included.

Impulsions are characterized by repetitive efforts to carry out socially disapproved or unlawful actions and as such often are considered a subtype of the obsessive-compulsive reactions. More or less characteristic are evidences of minimal discomfort, unconscious satisfaction, resistance to therapy, concealed hostility, and defiance toward society.

Psycho-
pathic
personality

Psychopathic personality sometimes is used to refer to a type of antisocial reaction characterized by impulsive and irresponsible behaviour that is accompanied by unstable emotional and social adavtation. Psychoyathic individuals seem poorly to anticipate the social consequences of their acts, appear to lack conscience, are undeterred by the prospect of punishment, and seem to learn little through experience.

Transient reactions and special syndromes. ***Grief state.*** Intense reaction to loss, particularly following the death of someone close, is normally called grief. Exaggeration and prolongation of normal grief constitutes the grief state, which can lead to the more severe symptoms of emotional depression or depressive reaction.

Stress reaction. Individual responses to emotional or situational stress are influenced by prior adjustment. Stress reactions can include acute manifestations typical of almost any of the psychoneuroses; fatigue state and anxiety-tension state may be so generated. The stress reaction is also commonly encountered in military situations.

Anorexia nervosa. This is a severe neurotic reaction in which appetite and weight loss on emotional bases are the prominent features. The anorexia (loss of appetite) can result in such an active aversion to food that it may lead to death. Anorexia nervosa in the 1970s was likely to be regarded as a symptom of some other neurosis rather than as a distinct clinical entity.

Ganser syndrome. Regarded by some as a variety of conversion reaction, this disorder is one in which objects are misnamed, questions elicit senseless answers, and the sufferer displays an abrupt loss of general competence. Especially described among prison inmates, the term Ganser syndrome has been utilized to include a broad spectrum of neurotic and psychotic manifestations (see also PSYCHOSES). In view of its locale and symptoms, the disturbance also has been called prison psychosis and nonsense syndrome.

Nail-biting and thumb-sucking

Conflict indicators. A significant group of attitudinal and behavioral manifestations serve to indicate emotional tension and conflict. They include somnambulism, enuresis, nail-biting, stuttering, nightmares. tics, various mannerisms, persistent insomnia. thumb-sucking, teeth clenching. hair pulling, head banging, and ceriain rhythmic movements. While these attitudinal and behavioral manifestations are often developmental and transient, they may persist, recur under later adult stresses, and accompany or indicate the potential for the progression of further emotional difficulties.

GENERALLY USEFUL CONCEPTIONS ABOUT NEUROTIC BEHAVIOUR

Many concepts evolved in relation to the dynamics, cause, prevention, and therapy of psychoneuroses have applicability in the general study of human reactions and behaviour.

One such concept is that an individual's earlier experience tends to take emotional precedence over subsequent impressions. According to this impression-priority rule, impressions resulting from a first visit, initial meeting, original experience in a sport or activity, learning to drive a car, first day at a new school, or initial contact with one's employer all tend to have special significance. Subsequent events, attitudes, and reactions henceforth can be profoundly influenced.

A number of advantages are consciously or unconsciously derived through emotional defenses or neurotic symptoms. The so-called primary gain refers to what are held to be the deeply buried, hidden needs, basis, and purpose of the neurotic activity. These defensively intended gains are said to comprise its raison d'être; included is a need to avoid conflict and consequent anxiety. The primary gain is said to be basic to the genesis of every neurotic disorder. Secondary gains refer to situational advantages that derive from an established symptom or illness. Thus a soldier who develops hysterical blindness becomes eligible for the secondary gain of being removed from the scene of battle. The indicted criminal who suddenly finds himself unable to tie his own shoelaces (as in Ganser's syndrome) may reap the secondary gain of avoiding prosecution on grounds of legal insanity.

A kind of neurotic status quo sometimes is achieved when the symptoms efficiently serve their defensive intent in allaying anxiety and in holding repressed conflict in abeyance. When such a balanced neurotic position is secured, the psychopathology tends to become more or less stable. The individual's freedom from anxiety and his inability to recognize any self-defeat or handicap can comprise an almost iron-clad defense against so-called therapeutic intervention. Indeed, many social associations (for example, choice of a marital partner) seem to be based on complementary neurotic symptoms. Neurotic interdependency of people in large measure is responsible for the development of social relationships and their long-range vicissitudes (for example, chronic discontent in marriage).

BIBLIOGRAPHY. O. FENICHEL, *The Psychoanalytic Theory of the Neuroses* (1945), a basic and thorough exposition of theory from the classically Freudian psychoanalytic view, including an extensive bibliography; H.P. LAUGHLIN, *The Neuroses* (1967), a comprehensive text and reference that systematically covers psychoneuroses, with many case illustrations and tables, a useful glossary of psychiatric terms, and index, *The Ego and Its Defenses* (1970), a detailed study of major ego defenses and minor mechanisms that contribute to improved personal adjustment as well as to self-defeat and neuroses, with numerous case illustrations, a useful appendix, and index; E. WEISS and O.S. ENGLISH, *Psychosomatic Medicine*. 3rd ed. (1957), a thorough study of relationships betwcsn emotions and bodily functions, with interesting illustrations, bibliography, and index; H.J. EYSENCK and S. RACHMAN, *Causes and Cures of Neurosis* (1965), a consideration of neurotic Sehaviour from the standooint of learning principles and of behaviour therapy.

(H.P.L.)

Psychoses

Psychosis is the term commonly used to designate a severe or major psychiatric disorder. In practice, the concept is more difficult to define. since severity is not an inflexible sign and a relatively small number of cases diagnosed as psychosis may in fact be less serious to the sufferer or to society than are some of those assigned to other psychiatric categories. The term psychosis is at times indistinctly equated with insanity: the latter term, when used legally or in popular language, suggests a person who is so incompetent that he may require special control or supervision. Difficulty in definition also arises from the inclusion in the psychotic category of many disturbances of different origin, course. and symptoms.

Insanity

According to some authorities, the term psychosis indicates not only actual or potential severity but also connotes that the disturbed state is accepted by the sufferer as a normal way of living. Most people who are labelled psychotic seem not to know that there is anything wrong with them. No matter what unusual behaviour the psychotic exhibits (*e.g.*, severe withdrawal from ordinary life or major emotional changes, or defect in memory or perception), ihat behaviour represents the routine way of responding to the environment around him. By contrast, people who are placed in another psychiatric category, psychoneurosis (or neurosis), generally appear to know they suffer some disorder and typically strive to get rid of the symptoms. 'Thus a psychoneurotic who is panicked by cats, or who tries to wash his hands every five minutes, almost invariably seeks relief from the symptoms; he is said to have insight into his abnormal state (see PSYCHONEUROSES).

A psychotic European who says that he is the emperor of China may insist on being recognized as Chinese royalty even though there is no observable evidence for such a belief. That the psychotic seems to accept his abnormality as normal often sharply distinguishes him from most of those around him. Although philosophers argue about the nature of "reality" (see METAPHYSICS), observers soon discover that psychotic ways of perceiving what most people call reality follow distinct patterns that cannot be validated logicaily by others in the individual's culture. Such characteristics common to psychotic people are not enough to specify a single biological category, as in classifying animal and plant species. The term psychosis is effectively limited to practical use, suggesting that one is suffering major psychiatric disorder of which he has no insight and that he probably requires hospitalization.

Since no collective biological category for psychosis is defined, causes, symptoms, and treatment vary considerably, each type of disturbance requiring separate examination.

Psychoses once were called mental alienations; psychiatrists dealt almost exclusively with them and were therefore called alienists. Nowadays psychiatrists have expanded their concerns to many additional problems, including minor personality disorders, questions of child rearing, and mental hygiene. Psychoses nevertheless continue to attract substantial attention as serious problems that require so much care that the great majority of inmates of psychiatric hospitals are drawn from ranks of psychotics.

Benign and pernicious psychoses

Psychoses once were classified as benign if recovery was satisfactory and prompt, and as pernicious if symptoms were protracted or became worse. Benign cases often were labelled as manic-depressive, while those with poor signs of recovery were diagnosed as suffering "dementia praecox." The practice was not dissimilar to that prevailing in general medicine before bacteria were well understood: when a child with a sore throat recovered, he had "tonsillitis," and if not, he was said to have died of "diphtheria." While many other classifications for psychoses have been proposed, one that has been generally adopted distinguishes between so-called organic and functional psychoses.

A psychosis is called organic if it seems to result from some demonstrable physical abnormality, such as a brain tumour, or some chemical derangement in the body. When the individual behaves abnormally, and no chemical or anatomical basis for that behaviour is found, the psychosis is said to be functional; for example, many psychotic signs of emotional maladaptation appear at least partially to be provoked by how one perceives the stresses of daily life.

Major controversy centres on the concept of functional disorder. Many authorities caution that although no organic bases are observed, even with the most refined techniques available, it is unwise to conclude that none exists. New, more sensitive methods indeed might permit detailed exploration of the nervous system with enough precision to disclose physical abnormalities. Additional debate concerns biochemical alterations detected in the bodies of some sufferers of functional psychoses, the question being whether the chemicals produce the disordered behaviour or result from it.

In many European countries the term endogenous (produced within the individual) is preferred to the term functional. The term endogenous directs attention away from environmental factors in psychosis to emphasize abnormality that may exist within the sufferer himself and that predisposes him to develop signs of disorder. The American (U.S.) Psychiatric Association has officially distinguished psychoses associated with organic brain dysfunction from those not yet demonstrably based on physical disorder. This avoids taking a stand on whether functional psychoses have organic origins; thus psychiatrists who speak of functional disturbance do not necessarily deny organic or physical predisposition.

FUNCTIONAL PSYCHOSES

Functional psychoses have been defined as those in which an organic basis has not yet been generally established nor widely accepted.

Schizophrenia. Schizophrenia is the most frequently made diagnosis among the major psychoses, schizophrenics constituting 20 to 25 percent of first admissions to public psychiatric hospitals. Since the symptoms tend to appear relatively early in life and tend to result in long periods of hospitalization, schizophrenics make up about 55 percent of the disturbed population at such hospitals. The prevalence rate (the total number of schizophrenics living at a given time) can only be estimated and has been calculated variously from about 170 to 950 per 100,000 of all living people. No known society or culture seems to be immune; in so-called developed lands schizophrenia tends to be more frequent among the unmarried, immigrants, those of low social status, people who belong to ethnic minorities, and those living in large cities. Assuming the lowest prevalence value (170 per 100,000), and a world human population of 4,000,000,000, a modest estimate of the total number of schizophrenic people is 6,800,000.

Schizophrenia (Greek: "split mind"), a term coined by Swiss psychiatrist Eugen Bleuler (1857–1939) to replace the older designation dementia praecox, indicates a *group* of disturbances with symptoms occurring in different combinations and intensity but generally having in common disturbances of feeling and thinking—a kind of withdrawal ("splitting") from the outside world. Neither dementia praecox nor schizophrenia is satisfactory as a term to cover all syndromes (combinations of symptoms) usually included in the group. During the 20th century there have been alternating trends to consider the term schizophrenia as denoting a single "disease" process or as a number of similar patterns of psychological reaction to the stresses of life.

Symptoms

All major types of schizophrenia now generally differentiated exhibit signs of withdrawal from the environment. (1) The simple type is differentiated by insidious and gradual reduction in external relationships and interests. Emotional expression is shallow, thinking tends to be simple and to refer to concrete things rather than to abstractions. The sufferer's general level of activity diminishes; he shows progressively less and less use of his inner resources and retreats to less demanding or stereotyped forms of behaviour. For example, a skilled industrial designer who develops simple schizophrenia may retreat to routine factory work. (2) The hebephrenic type is characterized mainly by shallow, inappropriate emotional reactions, silly or bizarre behaviour, false (delusional) beliefs and perceptions in the absence of the usual stimuli (hallucinatory activity). (3) The catatonic type is characterized by striking changes in bodily movement; sufferers may become almost completely immobile, often assuming statuesque positions. Mutism, extreme compliance, and loss of almost all voluntary activity are also common. Such inactivity may be preceded or interrupted by episodes of excessive movement and excitement, generally unpredictable and apparently impulsive. (4) The paranoid type, usually appearing later in life than the others, is characterized by unrealistic, illogical, hallucinatory thinking and by bizarre delusions of being persecuted or of being a great person (*e.g.*, of being Jesus Christ or the "emperor of the Pacific Ocean").

Beyond the four major types, many authorities also describe: (5) schizo-affective schizophrenia, distinguished from other types by moods of elation and depression; (6) undifferentiated schizophrenia, characterized by disturbed, but not sufficiently distinctive, behaviour and thinking; (7) pseudoneurotic schizophrenia, with predominantly neurotic behaviour, psychotic episodes being brief.

Although not consistently found, hallucinatory signs are perhaps most conspicuous in the hebephrenic and paranoid types. The most common are auditory, the individual reporting "voices" that no one else seems to hear, but the "reality" of which he accepts. Visual hallucinating also is relatively common, especially during acute episodes; less often there are signs of hallucinatory smelling and tasting and of abnormal bodily perceptions (*e.g.*, a depersonalized feeling that one's arm does not belong to him). The symptoms of withdrawal that are common to all types may be minimal and difficult to detect but are reflected as withdrawal into concrete ways of thinking and impaired abstract reasoning.

These "textbook" types of schizophrenia are by no means mutually exclusive; mixed syndromes are found especially in early acute cases as well as in some later chronic phases. As the disorder continues, some sufferers show few of their original symptoms, these having yielded to a mixed set of others.

The course of schizophrenia is extremely variable; a 48-hour schizophrenia has been described, yet people who never recover sufficiently to leave the hospital also constitute a significant percentage. Most improve to make at least a social recovery that permits them to live more or less adequately in society.

The course of schizophrenia

The classical course of schizophrenia sometimes is said to begin with a prodromal phase, characterized by hypochondriacal symptoms, introversion, restlessness, depersonalization, and abnormal preoccupation. The change in personality can be slow or abrupt and is followed by overt signs of psychosis in which the intensity and quality of symptoms become manifestly abnormal. Symptoms next tend to diminish in intensity and may finally disappear (go into remission). In chronic cases the abnormal signs tend to become less acute, but with increasing duration there is gradual deterioration.

Theories about the origin of schizophrenia include hypotheses that are anatomical, biochemical, psychological, hereditary, and environmental. In some countries (*e.g.*, Great Britain and Germany) biological or organic theories have been prevalent; in others (Switzerland and the U.S.) psychological or functional theories have attracted more adherents, although the number of authorities in all lands who endorse at least a partial organic basis grows steadily larger. Many theories cannot be considered *exclusively* psychological or biological; however, one extreme or another tends to predominate in each.

Predominantly psychological theories. German psychiatrist Emil Kraepelin (1856–1926) seems to have been first to distinguish what is now called schizophrenia from other psychiatric disorders and to describe it in detail. He gave it the name dementia praecox, a term that had been used by others earlier and in a more restricted sense. Kraepelin held that its varied manifestations all result from the same underlying physical disorder, suggesting either degenerative brain disease or metabolic self-poisoning (auto-intoxication), and pessimistically pronounced that all cases end in deteriorated intelligence (dementia). Bleuler, however, early influenced by Freudian theory, emphasized psychological roots (impairment of thinking) as the primary source of schizophrenic symptoms, although he did not deny the possibility of a coexisting toxic basis as well.

A Swiss-American, Adolf Meyer (1866–1950), attributed schizophrenic reactions to maladaptive learning (habit disorganization), emphasizing the anatomic and chemical integrity of the sufferer while affirming that such functions as learning were not to be separated from physical structures, a view called psychobiological.

Psychoanalyst Sigmund Freud (1856–1939) showed more interest in psychoneuroses than in psychoses, but his theories and investigations have nevertheless been widely applied in efforts to account for schizophrenia. Schizophrenic symptoms have been interpreted psychoanalytically as symbolic manifestations similar in form, content, and motivation to ordinary dreaming (see DREAMS). The psychoanalytic view interprets regressive schizophrenic symptoms (*e.g.*, soiling oneself) as a retreat or withdrawal to less mature levels of development. Such restitutive symptoms as hallucinating, delusional activity, phantasizing, or peculiarity of language are viewed as attempts to replace the world from which the person has withdrawn.

Swiss psychiatrist Carl Gustav Jung (1875–1961) asserted that schizophrenia is caused by unusually intense regressive activity and by an abnormal resurgence of primitive tendencies, evolutionary throwbacks from what he called the person's "collective unconscious." He imagined the latter to be an inherited repository of primordial images (archetypes) that accumulate from the experiences of each individual's ancestors.

An American psychiatrist, Harry Stack Sullivan (1892–1949), blamed schizophrenia on the sufferer's faulty interaction with others, especially on poor parent-child relationships. One's anxiety and lack of self-esteem, originating in these early relationships, he held, lead to peculiar ways of coping with life (parataxic distortions), to lack of consensual validation (decreasing agreement with others in perceiving the world), and finally to what Sullivan called schizophrenic panic.

An Italian-American psychiatrist, Silvano Arieti (1914–), has attributed major, but not exclusive, importance to psychological factors. He believes that the psychological life history of the schizophrenic, from birth to the onset of the disorder, can be divided into four stages, of which only the last can be considered psychotic. In the first three stages the future patient responds to an unhealthy environment in specific ways that distort and magnify the environmental abnormality. Consequently, it will be more difficult for him to integrate psychologically. When, under the stress of psychological problems with which he cannot cope, the patient decompensates and becomes psychotic, he adopts a special way of thinking, which makes him delusional so that he deals with reality in an individualistic and inappropriate way. The patient follows a particular way of thinking, called paleologic, which is not reconcilable with Aristotelian logic. Arieti feels that the schizophrenic process also may be interpreted psychosomatically, inasmuch as intensely disturbing emotional factors may bring about the resurgence of obsolete functional and neuronal patterns.

Predominantly biological theories. Many authors have interpreted schizophrenia as a hereditary disease. A detailed discussion of these efforts is given in HUMAN BEHAVIOUR, INNATE FACTORS IN: Personality. In general, these genetic theories of schizophrenia are based on the observation that the disorder seems to run in families.

Heredity and metabolic disorder

Many other investigators agree that schizophrenia is more frequent in some families than in others but that what is transmitted by heredity is only a predisposition for the disorder. Whether or not actual symptoms develop is said to depend on environmental factors (presumably psychological) to change this potentiality into clinical actuality. It is evident that the available data do not permit an explanation of schizophrenia solely on the basis of the laws of heredity, as has been done for such inherited medical disorders as hemophilia and muscular dystrophy.

Biochemical interpretations of schizophrenia based on the notion of pathological metabolic activity have also been advanced. According to one view, the body produces a substance similar to mescaline, a chemical obtained from a cactus, which produces hallucinatory behaviour when swallowed. Some data have confirmed this hypothesis while many other studies have failed to do so. Other investigators have reported isolating from the blood of schizophrenics a substance called taraxein that is said to be responsible for the disorder. Others have theorized that a deficiency of serotonin, another chemical produced in the body, is responsible for the psychosis, although actual studies have yielded contradictory results.

Gross and microscopic examinations of the nervous systems of deceased schizophrenics have failed to reveal distinctive anatomical alterations. If such alterations exist, they must be so subtle that they escape detection with available methods.

Treatment. Earlier pessimism about the outlook for recovery in schizophrenic cases has changed markedly. Even orthodox psychoanalytic therapy, about which Freud himself was skeptical in the treatment of psychoses, has been modified for use with schizophrenics. These modifications include dispensing with the couch, direct intervention by the therapist in changing the symbolic language of the sufferer, and much less reliance on free association. Other individual psychological techniques in Europe and North America include the treatment of non-hospitalized (ambulatory) schizophrenics during periodic visits to the therapist's office.

By far the majority of sufferers are treated with physical methods. Drug therapy consists in the administration of large doses of tranquillizers, particularly one called chlorpromazine. Other techniques include electroconvulsive therapy in which electric current is passed through the individual's head; insulin therapy, in which enough of that hormone is administered to produce coma; and psychosurgery, in which connections between different parts of the brain are severed. The latter technique (sometimes called lobotomy) underwent a great decline of interest by the mid-1950s; psychosurgery now is applied quite rarely and only in the most recalcitrant cases of schizophrenia.

An increasing number of schizophrenics are treated outside hospitals, but most continue to be hospitalized, especially during acute episodes. Almost all psychiatric hospitals are equipped to provide physical treatments and occupational therapy, and an increasing percentage provide psychotherapy. In many cases, what seems most helpful is to provide the individual with an environment in which he is protected and in which, away from the demands of ordinary life, he can gradually work his way through his periods of confusion and disturbance.

Success of treatment appears to vary with the stage of the disorder at which it is started. If the symptoms have been present for a long time, successful results are relatively difficult to attain. Good outcome is related to: acute, sudden onset of a confused, excited kind of disturbance; absence of schizophrenic family history; presence of additional emotional symptoms usually found in manic-depressive psychosis (see below Affectivepsychoses); and evidence of some fairly clear psychological conflict that seems to have precipitated the disturbance. Statistical studies indicate that shock treatment (insulin and electroconvulsive) produces a greater number of improvements among schizophrenics than do other methods; however, shock seems to work even better for psychoses primarily characterized by emotional symptoms (*e.g.*, depression).

Prevention. Without clear-cut understanding of the origins of schizophrenia, specific and reliable preventive measures remain to be established. A number of authorities, influenced largely by psychoanalytic points of view, hold that one's relationship in the early years of life with his family, particularly with his mother, plays a considerable role in the development of schizophrenia. Their recommendations for prevention therefore centre largely in the provision of security for the child through warm, understanding relationships in the family.

There are also indications that urban areas, very large cities in particular, tend to promote the development of this disorder and industrial societies seem to spawn the disturbance more than do agricultural settings. On these grounds some have looked to wider dispersion of the population as hopefully reducing the prevalence of schizophrenia, although the disorder is found in all cultures and societies.

Affective psychoses. Affective (emotional) psychoses include what are called manic-depressive psychosis and psychotic depression; it is debated whether the latter is an incomplete form of the manic-depressive psychosis or an independent disorder. While schizophrenia was not recognized as an entity until the 19th century, forms of manic-depressive psychosis have been well described since ancient times. People who today would be classified as manic-depressive are discussed in the writings of the Greek physicians Hippocrates (flourished 400 BC) and Aretaeus of Cappadocia (2nd century AD). Kraepelin conceived of manic-depressive psychosis as one syndrome that includes many varieties.

Manic-depressive psychosis manifests itself with recurring, sometimes cyclic, attacks of depression and of manic (excited) behaviour, often elated. Attacks of depression are characterized by signs that the person has deep feelings of melancholia (sadness); onset may be acute and dramatic, or slow and insidious. The sufferer typically looks unhappy, says he feels hopeless and worthless, that he considers his life a torment. He seems to experience a desire to punish himself by self-destruction; evidence of suicidal thinking is observed in about 75 percent of these people, and actual suicide attempts are found to be made by at least 10 to 15 percent. In fact, suicide constitutes the greatest risk in manic-depressive psychosis, although some sufferers (especially females) kill their small children, whom they seem to consider parts of themselves, immediately prior to their suicidal attempts. Generally, however, the depressed person is dangerous to himself rather than to others.

Suicide

The individual apparently convinces himself that his depression and suffering are justified, saying that he is unhappy because he is sick, is destitute, has lost all his money, has sinned, is guilty, is worthless, is a failure, and so on. Hypochondriacal thinking, as an almost full-blown delusion of being ill, is not uncommon. Evidence of depression is also seen in retardation of motor activity; bodily movements are reduced in number and are carried out at a slow pace. The depressed person is neglectful of his appearance and manifests limited ability to work or to take care of household and family duties. Depression may become so profound as to reach the state of stupor; the individual seems unable to talk or to move and may be mistaken for an immobile catatonic schizophrenic. Insomnia, decrease in appetite, loss in weight, and marked decrease in sexual desire are common symptoms.

In classical cases of the manic-depressive psychosis, periods of depression are followed by manic attacks in which the person becomes excited; while the excitement may be unpleasant for the sufferer, the mood often is one of elation (euphoria). The manic individual typically shows unrestrained good humour; however, he may change to sarcasm, irritation, and hostility when he comes into conflict with the environment. He may consider himself rich, strong, and very healthy and make grandiose plans; he may dissipate large amounts of money in a few days. His thinking seems flighty, he cannot concentrate on any topic more than briefly, and his speech is rapid. In this flight of ideas the manic individual is ineffective in self-criticism and manifests poor judgment. Motor activity may be so excessive that the sufferer becomes exhausted; in other cases the attack is of mild intensity (hypomanic) and continues without fatigue.

During depression or mania the psychotic patient shows no insight into the abnormal nature of his mood. He seems to accept his behaviour as an appropriate and justifiable way of living and feeling. The most typical form of the psychosis is circular, being characterized by an alternation of manic and depressed episodes. The first attack may occur any time between adolescence and the age of 45. Depressive attacks are much more common than are manic, especially as the disorder progresses. There are, however, numerous variations; in one (agitated depression), motor restlessness, more typical of manic excitement, is superimposed on a set of markedly depressive symptoms.

The disorder is likely to be called psychotic depression when there is no history of repeated episodes of depression or of circular attacks of depression and mania (as in classical manic-depressive psychosis), and when the depression is intense and not perceived as abnormal by the sufferer (as it would be in neurotic depression). When environmental precipitating factors (*e.g.*, loss of a loved one) are easily recognized, the diagnosis of psychotic depression is more likely to be given as a separate clinical entity; nevertheless, some authorities still consider it a partial form of manic-depressive psychosis.

The psychological symptoms of all such psychoses have been classified by Arieti as either "self-blaming" or "claiming." In self-blaming depression the focus is on self-accusation, self-deprecation, and guilt feelings that lead one to feel he does not deserve help. This type of depression is traditionally described in textbooks of psychiatry. In claiming depression, a term that has begun to attract more diagnostic use, the symptoms of the individual seem to reflect a gigantic claim on the attention and sympathy of others. All the symptoms seem to have a message: "Help me; pity me; it is in your power to relieve me. If I suffer, it is because you don't relieve me of this suffering."

Some psychiatrists have reported a decline in the incidence of the disorder; for instance, in one study of hospitals over a period of 20 years (from 1928 to 1947), manic-depressive psychosis declined from 13.5 percent of all first admissions to 3.8 percent. After the mid-1960s, however, psychiatrists detected what seemed to be an increase in the number of depressed people though their symptoms are not typically manic-depressive.

Theories. Some authorities have stressed hereditary factors in the etiology (causation) of manic-depressive psychosis, one view attributing it to a single dominant gene. A German psychiatrist, Ernst Kretschmer (1888–1964), noticed a large number of short, heavy people (so-called pyknic body type) among those suffering from manic-depressive psychosis, and other investigators have found what seem to be distinctive fluctuations in the levels of certain chemicals (17-hydroxycorticosteroids) in the urine of depressed patients.

Psychoanalysts have compared psychiatric depression to normal grief, suggesting that it is normal to feel sad through concern about a lost dear one, while the depressed psychiatric patient is disturbed by guilt feelings. Unconscious hostility that he had for the lost person is now said to be directed toward himself; in a sense, the personality of the lost person is thought to have been introjected (incorporated) into the unconscious life of the depressed person.

Arieti believes that the depressed patient has sustained a trauma in early childhood, when from an environment characterized by great parental care and devotion he found himself carried into an atmosphere of great expectation. To recapture what he had lost, he becomes dependent and claiming, or he becomes duty bound, compliant, and hard working. Later in life, any loss is experienced as a reactivation of the early loss. Arieti has also tried to illustrate the importance of the cultural environment in facilitating those psychodynamic mechanisms that lead to manic-depressive psychosis. The psychosis occurs more frequently in those cultures that have been called inner-directed and tends to disappear where this culture is disappearing. This hypothesis receives some validation from research on Hutterites, an ethnic minority of German ancestry who settled in the Dakotas and Montana. They are held to constitute a typical inner-directed culture; among them manic-depressive psychosis occurs 4.33 times more frequently than schizophrenia, contrary to the comparative incidence of these two psychoses elsewhere.

Treatment. While psychoanalytic treatment has been attempted, electric-shock treatment seems to cure practically all episodes of depression but does not prevent their recurrence and is less effective in manic attacks. Drugs are commonly used to control attacks of depression, especially so-called antidepressants, like the drug imipramine. In manic attacks the use of compounds of lithium has given good results, but not invariably. The poor judgment of manics and the possibility of suicide among depressed people make hospitalization advisable in many cases.

Involutional psychoses. Involutional psychoses are generally classified as involutional melancholia, and as involutional paranoid state (in Europe often called paraphrenia). Several authorities have denied the existence of these conditions as separate clinical entities, being inclined to hold that involutional melancholia is a variety of affective psychosis and that involutional paranoid state is a variety of schizophrenia. Some official diagnostic publications nevertheless retain the category.

Involutional melancholia

These disorders are observed during the so-called involutional period of life; that is, between about 48 and 60 years of age when signs of aging become apparent. Observed three times more frequently in women than in men, involutional melancholia may occur at the time of menopause (cessation of menstruation) but in most cases appears from two to seven years after menopause. Among men, the symptoms generally appear even later; these include anxiety, insomnia, restlessness, and loss of appetite and of sexual desire. In many cases there is a marked loss of weight; the person worries about his health and consults a physician, hoping to find a medically treatable reason for his distress. Soon the mood of melancholia becomes predominant and covers all the other symptoms; agitation is more common than motor retardation, this serving as a distinguishing characteristic from other types of depression. Expressions of feelings of worthlessness, self-condemnation, and guilt are prominent in involutional melancholia.

Biological, especially glandular, factors have been considered responsible; yet, treatment with hormones generally seems ineffective in involutional melancholia. Electric-shock treatment and drugs (especially antidepressants) are most profitably used, as they are in other types of depression. Many psychiatrists see in the psychosis the culmination of a life of disappointment, an inability to face the future, and a sense of regret for that part of life unsuccessfully spent.

The involutional paranoid state is characterized by suspiciousness and hostility, changing into delusions of persecution and ideas of reference. Often it is difficult to differentiate this disturbance from paranoid schizophrenia. Mixed cases occur with symptoms of melancholia and paranoid thinking.

Paranoia. The existence of paranoia as a separate entity is doubted by many psychiatrists, who consider it a variety of schizophrenia. As described in old textbooks of psychiatry, paranoia is a condition characterized by apparently logical systems of delusions (false beliefs). The person nourishes beliefs that appear to be absurd to others; nevertheless, he gives convincing and apparently plausible arguments to support them.

The paranoiac does not resort to primitive ways of thinking to demonstrate the validity of his delusions (as schizophrenics do); rather, his thinking is generally logical once the false premises are accepted. Often, mere possibilities and coincidences are considered as conclusive evidence; the paranoiac feels persecuted or imagines plots and conspiracies organized against him or against parts of society with which he identifies. At times the person sees himself as a saviour of society whose mission it is to unmask the obscure plotting of clandestine agencies. The paranoiac becomes so sure of the validity of his beliefs that he may go to the extent of killing the alleged persecutors; more often he becomes the plaintiff in almost endless legal suits.

Incidence of paranoia

Paranoia is most likely to be diagnosed among unmarried men in their 40s and 50s who have made a poor sexual adjustment, and who always have tended to experience difficulties in relating to people. Paranoiacs as described in classic books of psychiatry are rarely encountered and seem most unlikely to respond to treatment. They may require forced hospitalization when authorities judge them to endanger the safety of others. Some authors have described what they call mild cases of paranoia that may be of short duration and that are said to respond to psychotherapy.

Rare functional psychoses. Other types of functional psychoses have been described, but their occurrence is rare; a few of them are mentioned below.

Ganscr syndrome. This occurs especially among prisoners who are facing trial, occasionally among people confronted with other unpleasant conditions. Together with partial amnesia, the sufferer shows an abrupt loss in his ability to reason, giving absurd answers or making unbelievable statements; for instance, he map say that 3 plus 3 is 7 or that a horse has five legs. The syndrome often is confused with voluntary malingering (faking) by prisoners who want to escape indictment. Some interpret the Ganser syndrome as an involuntary escape from rationality in an unconscious effort to avoid responsibility for past actions.

Capgras' syndrome. This diagnosis has received considerable attention by European psychiatrists since the original description in 1923. On meeting someone he knows very well, a sufferer will claim that the person is a double or an imposter; for example, he may claim that his mother is another person who has tried to assume her appearance to deceive him. The phenomenon is more characteristic and specific than the usual misidentifications occurring in schizophrenia, but some question whether this is a distinct clinical entity or a schizophrenic syndrome.

Latnh syndrome. An extremely rare psychosis in Western countries, latah syndrome is fairly common in the Fast East, especially in Malaya. The patient acquires the habit of repeating the words or sentences of other people, especially persons in authority. The patient, in pantomime fashion, repeats or imitates the gestures and acts of others. The same kind of behaviour is observed among schizophrenics.

Paranoid behaviour in pairs

Folie à deux. A rare syndrome characterized by the simultaneous occurrence of psychosis in two persons who have a close relationship (*e.g.*, live together) is called folie à deux. In such cases, one of the two (the inductor) is usually suffering from paranoia or some other paranoid disorder. The inductee (generally a woman or a passive male) seems to accept the delusional attitudes of the dominant person and often is his spouse or child. When the passive person is separated from the domination of the inductor, his psychosis is likely to disappear, while the inductor retains his abnormal symptoms.

ORGANIC PSYCHOSES

Psychoses for which a definite anatomical or chemical basis has been definitely established are called organic.

Senile psychosis. After about the age of 60, mild impairment in such functions as remembering and learning is so common as to be considered statistically normal. The term senile psychosis is used when such defects are so pronounced as to deprive the person of insight into his condition and to impair his ability to take care of himself. In a large-scale study, about 15 percent of all patients admitted to psychiatric hospitals were diagnosed as suffering from senile psychosis. Differences from one society to another in attitudes toward the shortcomings of old age understandably affect the number of admissions; for example, the number of hospitalized senile people is reduced to the extent that they are cared for at home.

Paranoid thinking, as of being robbed or victimized, often occurs in senile psychosis. Even more pronounced are defects in orientation as to time, place, and occasionally about personal identity. Memory may be impaired only in relation to recent happenings, but in advanced cases the loss may extend to long past events (see MEMORY, ABNORMALITIES OF). Ability to learn new skills, to adapt to new environments, to calculate, and to use logic may be severely impaired; and mood tends to change rapidly from laughter to tears and vice versa.

Among sufferers of senile psychosis the brain is markedly reduced in size; while normal adult brain weight is about 1,300 grams, that of senile psychotics has been observed to range from 1,100 to 900 grams. The reduction in volume is particularly pronounced in the front and middle parts of the brain surface (cortex). Microscopic studies of nerve cells reveal that the parts of the cortex believed to have evolved earliest are least affected. Often there is a marked reduction in the number of nerve cells in all the layers of the cortex and selective shrinkage or coalescence of parts of the cells that remain. The brains of senile psychotics also are likely to show many senile plaques—*i.e.*, small roundish areas of tissue degeneration scattered throughout the cortex, but particularly numerous in the frontal lobes and the hippocampal cortex.

Varieties of the disorder, generally labelled presenile psychoses, occur between about 40 and 60 years of age. Alzheimer's disease is fundamentally similar to senile psychosis but appears earlier in life and involves more severe behavioral and neural change. Pick's disease, again with similar symptoms, is characterized anatomically by circumscribed regions of atrophy (wasting away) in localized parts of the brain.

Psychoses with cerebral **arteriosclerosis.** In senile psychosis the nervous tissue is directly involved and degenerating, but in psychoses with cerebral arteriosclerosis brain deterioration results from changes in the circulatory system. Arteriosclerosis (hardening of the arteries) may be present throughout the body or may be restricted to the brain; in any case, the blood supply of the brain is disturbed. Large blood vessels become thickened and tortuous, their inner walls becoming coated with a fatty material that reduces the volume of blood they can carry; smaller blood vessels are even more likely to become occluded (plugged). As a consequence undernourished bits of brain tend to soften, nerve cells degenerate, and blood vessels in the head may burst and bleed. Psychologically the symptoms of cerebral arteriosclerosis resemble those of senile psychosis; disorientation, memory defects, and inability to learn and calculate are similar. Generally, however, the symptoms fluctuate, at least in the beginning of the illness, and the progression of the disorder is less uniform than in senile psychosis. A change of personality, with decay in judgment and loss of ethical behaviour, is observed in some cases; confusion with apprehension and panic, childish and impulsive behaviour also occur.

In both senile and arteriosclerotic psychosis, psychiatric management consists of offering the sufferer an environment adapted to his limitations. Efforts to use chemicals related to ribonucleic acid (RNA) in treating arteriosclerotic and senile patients have shown some success, but results have been inconsistent. The most commonly used medical treatment is symptomatic in nature, aiming at ameliorating the symptoms; for example, drugs that dilate blood vessels and similar techniques for nourishing oxygen-starved brain tissue are used.

General paresis. Although this disease once was thought by most authorities to be functional, in 1913 paresis was shown to be the result of syphilis of the brain. The illness typically develops several years after the original syphilitic infection (see REPRODUCTIVE-SYSTEM DISEASES). The infectious organism (***Treponema*** pallidurn) produces clearly discernible alterations in the microscopic structure and circulation of the cerebral cortex.

At the behavioral level the psychosis assumes different patterns, at times resembling that of schizophrenia, at others that of cerebral arteriosclerosis. Most typical, however, is the expansive paretic, a grandiose, euphoric psychotic who may claim to be a millionaire, a great inventor, or a genius. He lacks insight, his judgment is very poor, and his behaviour is impulsive and capricious. Once very common, the problem of general paresis has faded wherever medical care and public health education are adequate. When they were considered to be suffering from a functional psychosis, paretics were doomed to hopeless, wasted months or years of psychotherapy; standard therapy today consists of giving large quantities of penicillin or other antibiotics. Since early infections of syphilis are so readily treatable, the complication of general paresis has become rare in enlightened communities. Some psychiatrists have not seen even a single case for years.

Alcoholic psychoses. Alcoholic psychoses are psychiatric disorders produced by withdrawal from or by abuse of alcohol. Delirium tremens is the most common alcoholic psychosis; after prolonged and excessive use of alcohol, if a person abstains from drinking he tends to become restless, tremulous, fearful, unable to sleep or rest. Increasing tremulousness ("the shakes") is followed commonly by visual and auditory hallucinating. Most characteristic are visions of small animals, especially insects, crawling around the sufferer's body. Optical illusions, with external objects appearing to change size and form, are common. The major part of modern treatment consists of massive doses of vitamin B.

Delirium tremens

Alcoholic hallucinosis seems related to schizophrenia and to delirium tremens; alcoholics who abstain after heavy drinking tend to become hallucinated and deluded like schizophrenics. The trembling of delirium tremens is lacking, and visual hallucinating occurs more frequently than in schizophrenia.

Psychoses produced by continuous abuse of alcohol include Wernicke's encephalopathy, in which damage to structures deep within the brain is found. Confusion and excitement, followed by somnolence and stupor, are the main symptoms. In a similar disorder called chronic alcoholic deterioration, there is a gradual decline of sensory, rational, and ethical behaviour, the sufferer becoming inconsiderate, quarrelsome, crude, and often obscene.

Korsakoff's psychosis is another of the typical disturbances that develop after the prolonged use of alcohol. The individual suffers severe memory defect that he somehow may succeed in hiding by fabricating all kinds of fanciful tales. In about 50 percent of the cases, polyneuritis (painful nerve inflammation) accompanies the psychiatric syndrome.

Psychosis with Huntington's chorea. Often cited as the only clear example of a hereditary psychosis, Huntington's chorea is transmitted by a dominant gene, and most cases in North America can be traced to three men who went there in 1630 with John Winthrop, the first governor of the Massachusetts Bay Colony. Brain pathology consists of atrophy of the caudate and lenticular nuclei, parts deep within that organ. The illness is characterized by choreiform (twisting) movements of the neck, trunk and limbs, and grimacing of the face. The psychiatric picture consists of progressive intellectual deterioration, with occasional paranoid or depressive features; the illness in the 1970s was irreversible and incurable.

Psychosis with epilepsy. Epilepsy is a neurological disorder characterized by convulsions and by distinctive electroencephalographic changes. In a minority of cases epilepsy produces some symptoms that may be considered psychotic. Among these are so-called psychic seizures in which the epileptic feels forced to think or to act in ways that are in disagreement with his conscious determination. Hallucinations, delusions, and illusions occur; a characteristic symptom is the déjà vu feeling that what one is experiencing now has already occurred in his past life.

Another typical symptom associated with epilepsy is the so-called fugue; the sufferer may leave home, travel for several days or weeks without seeming to have full realization of what he is doing. He may suddenly come to himself and interrupt the fugue without appearing to remember what he has done. Although he apparently is not fully conscious during the fugue, he is able to abide by the rules of society; for instance, respecting traffic regulations.

In some cases of epilepsy (especially those involving temporal-lobe brain function) there are symptoms resembling schizophrenia, depression, or organic deterioration. The main treatment is directed toward the underlying epilepsy, while symptomatic therapy is applied to the concomitant psychiatric signs.

Other organic psychoses. Psychoses may occur in a large number of organic conditions; these are best studied and managed by neurologists. Among them are postencephalitic conditions, conditions following head injury, brain tumours, and various chronic neurologic disorders. A wide range of symptoms is observable, including memory defect, disorientation, poor judgment, inability to learn, and emotional instability. Depressive features and paranoid thinking may complicate the clinical picture. Toxic psychoses are the result of taking poisonous materials (*e.g.*, bromides, mescaline) into the body.

Disorders that some consider organic, that others call functional, and that still others deny as separate clinical entities are the so-called postpartum (after childbirth) psychoses. Some authorities, focussing on the symptoms, consider them as cases of schizophrenia or depression precipitated in the mother by the stresses of giving birth. Some attribute them to endocrine or other chemical changes engendered by labour and the termination of pregnancy. Others stress the psychological aspects of giving birth. The woman who is unprepared for motherhood is said to be presented with a challenge with which she cannot cope. She is seen to have doubts about fulfilling her maternal role and to have mixed feelings toward her child and her marriage. Whatever the causes, there is generally an interval of a day or two between the birth of the child and the appearance of postpartum psychosis.

Such differences in authoritative opinion eventually may be expected to yield to careful research as the subtler bases of human behaviour and experience become more fully understood.

BIBLIOGRAPHY. S. ARIETI (ed.), *American Handbook of Psychiatry,* 3 vol. (1959–66), gives one of the best textbook accounts of all psychoses. For a shorter presentation, see F.C. REDLICH and D.X. FREEDMAN, *The Theory and Practice of Psychiatry* (1966). More specialized works include: S. ARIETI, *Interpretation of Schizophrenia* (1955), a comprehensive account of the major notions on the schizophrenic psychosis; L. BELLAK and L. LOEB (eds.), *The Schizophrenic Syndrome* (1969), a review of the literature on the subject; R. CANCRO, *The Schizophrenic Reactions: A Critique of the Concept, Hospital Treatment and Current Research* (1970), a conceptual view of the disorder; A.T. BECK, *Depression: Clinical, Experimental and Theoretical Aspects* (1967); and L. BELLAK *et al., Manic-Depressive Psychosis and Allied Conditions* (1952). S. ARIETI, "Some Socio-Cultural Aspects of Manic-Depressive Psychosis and Schizophrenia," in J.H. MASSERMAN and J.L. MORENO (eds.), *Progress in Psychotherapy,* vol. 4 (1959); and J.W. EATON and R.J. WEIL, *Culture and Mental Disorders* (1955), study the importance of cultural factors in relation to manic-depressive psychosis. F.J. KALLMANN, *Heredity in Health and Mental Disorder* (1953); E. KRINGLEN, *Heredity and Environment in the Functional Psychoses* (1968); D. ROSENTHAL (ed.), *The Genain Quadruplets* (1963), and with S.S. KETY, *The Transmission of Schizophrenia* (1968), discuss the origins of psychoses from a genetic point of view. P. HOCH and J. ZUBIN (eds.), *Schizophrenia* (1965); and D.W. WOOLLEY, *The Biochemical Bases of Psychoses* (1962), approach the problem from a biochemical point of view. J.A. HAMILTON, *Postpartum Psychiatric Problems* (1962), is a good book on the psychiatric complications of childbirth. Of special interest are: S. ARIETI, *The Intrapsychic Self: Feeling, Cognition, and Creativity in Health and Mental Illness* (1967), a study of psychoses particularly from the point of view of intrapsychic mechanisms; and R.G. HEATH, *Serological Fractions in Schizophrenia* (1963), a study of schizophrenia from a biological point of view.

(S.Ar.)

Ptolemy

Although Ptolemy (Claudius Ptolemaeus) was a celebrated astronomer, geographer, and mathematician whose work had a profound influence on subsequent generations, virtually nothing is known about his life. The period during which he flourished (2nd century AD) is known from the dates of certain of his astronomical observations. It was from these observations and from the work of previous astronomers that Ptolemy evolved his detailed description of an Earth-centred (geocentric) universe, a revolutionary but erroneous idea that governed astronomical thinking for over 1,300 years. His influence in geography was equally long-lasting.

According to Theodorus Meliteniota of Byzantium, Ptolemy was born in the Hellenistic city of Ptolemais Hermii. He did his major work at Alexandria between 127 and 145 and may have been active as late as 151.

Astronomical work

Ptolemy's astronomical work was enshrined in his great book HE; *mathēmatikē syntaxis* ("The Mathematical Collection"), which eventually became known as *Ho megas astronomos* ("The Great Astronomer"). During the 9th century, however, Arab astronomers used the Greek superlative *Megistē* to refer to the book. When the definite article *al* was prefixed to the term, its title then became known as the *Almagest,* the name still used today.

The *Almagest* is divided into 13 books, each of which deals with certain astronomical concepts pertaining to stars and to objects in the solar system (the Earth and all other celestial bodies that revolve around the Sun). It was, no doubt, the encyclopaedic nature of the work that made the *Almagest* so useful to later astronomers and that gave the views contained in it so profound an influence. In essence, it is a synthesis of the results obtained by Greek astronomy; it is also the major source of knowledge about the work of Hipparchus, most probably the greatest astronomer of antiquity. Although it is often difficult to determine which findings in the book are those of Ptolemy and which are those of Hipparchus, Ptolemy did extend some of the work of Hipparchus through his own observations, apparently using somewhat similar instruments. For example, whereas Hipparchus had compiled a star catalog (the first of its kind) containing 850 stars, Ptolemy expanded the number in his own catalog to 1,022 stars.

On the motions of the Sun, Moon, and planets, Ptolemy again extended the observations and conclusions of Hipparchus—this time to formulate his geocentric theory, which is popularly known as the Ptolemaic system. In the first book of the *Almagest,* Ptolemy describes his geocentric system and gives various arguments to prove that, in its position at the centre of the universe, the Earth must be immovable. Not least, he showed that if the Earth moved, as some earlier philosophers had suggested, then certain phenomena should in consequence be observed. In particular, Ptolemy argued that since all bodies fall to the centre of the universe, the Earth must be fixed there at the centre, otherwise falling objects would not be seen to drop towards the centre of the Earth. Again, if the Earth rotated once every 24 hours, a body thrown vertically upwards should not fall back to the same place, as it was seen to do. Ptolemy was able to demonstrate, however, that no contrary observations had ever been obtained. As a result of such arguments, the geocentric system became dogmatically asserted in Western Christendom until the 15th century, by which time detailed observations had made the system so complex that its validity had lo be seriously questioned. In 1543 the geocentric system was supplanted by the heliocentric (Sun-centred) system of Nicolaus Copernicus (*q.v.*), a Polish astronomer.

Ptolemy accepted the following order for celestial objects in the solar system: Earth (centre), Moon, Mercury, Venus, Sun, Mars, Jupiter, and Saturn. He realized, as had Hipparchus, that the inequalities in the motions of these heavenly bodies necessitated either a system of deferents and epicycles or one of movable eccentrics (both systems devised by Apollonius of Perga, the Greek geometer of the third century BC) in order to account for their movements in terms of uniform circular motion. In the Ptolemaic system, deferents were large circles centred on the Earth, and epicycles small circles whose centres moved round the circumferences of the deferents. The Sun, Moon and planets moved round the circumference of their own epicycles. In the movable eccentric, there was one circle; this was centred on a point displaced from the Earth, with the planet moving round the circumference. These were mathematically equivalent schemes. Even with these, all observed planetary phenomena still could not be fully taken into account. Ptolemy therefore exhibited brilliant ingenuity by introducing still another concept. He supposed that the Earth was located a short distance from the centre of the deferent for each planet and that the centre of the planet's deferent and the epicycle described uniform circular motion around what he called the equant, which was an imaginary point that he placed on the diameter of the deferent but at a position opposite to that of the Earth from the centre of the deferent—*i.e.*, the centre of the deferent was between the Earth and the equant. He further supposed that the distance from the Earth to the centre of the deferent was equal to the distance from the centre of the deferent to the equant. With this hypothesis, Ptolemy could better account for many observed planetary phenomena.

Although Ptolemy realized that the planets were much closer to the Earth than the "fixed" stars, he seems to have believed in the physical existence of crystalline spheres, to which the heavenly bodies were said to be attached. Outside the sphere of the fixed stars, Ptolemy proposed other spheres, ending with the *primum mobile* ("prime mover"), which provided the motive power for the remaining spheres that constituted his conception of the universe.

Work in mathematics and geography

As a geometrician of the first order, Ptolemy performed important work in mathematics. He devised new geometrical proofs and theorems; and, in a book entitled *Analemma* (Greek *Peri analēmmatos*; Latin *De analemmate*), he discussed the details of the projection of points on the celestial sphere (an imaginary sphere extending outward from the Earth for an infinite distance and on whose surface the objects in space appear to be located) onto three planes at right angles (90") to each other—the horizon, the meridian, and the prime vertical. In another book, the *Planisphaerium*, Ptolemy is concerned with stereographic projection—the delineation of the forms of solid bodies on a plane—and here he used the south celestial pole as his centre of projection.

Ptolemy also prepared a calendar that gave, in addition to weather indications: the risings and settings of the stars in the morning and evening twilight. Other mathematical publications include a work, in two books, entitled *Hypotheseis ton planōmenōn* ("Planetary Hypothesis") and two separate geometrical works, one of which is concerned with proving that there cannot be more than three dimensions of space; the other contains an attempted proof for a postulate on parallel lines that had been devised by Euclid. According to one authority, Ptolemy wrote three books on mechanics; another authority, however, credits him with only one mechanical work, *Peri ropōn* ("On Balancing").

Ptolemy's work on optical phenomena appeared in *Optica*, the original edition of which consisted of five books. In the last book, he deals with a theory of refraction (the change in direction of light and other energy waves when they pass obliquely from a medium of one density into a medium of different density), and he discusses the refraction suffered by light from celestial bodies at various altitudes. This is the first recorded attempt at a solution of this observational problem. Ptolemy also wrote a three-book treatise on music known as the *Harmonica*.

As a geographer, Ptolemy's reputation rests mainly on his *Geōgraphikē hyphēgēsis* (*Guide to Geography*), which was divided into eight books; it included information on how to construct maps and lists of places in Europe, Africa, and Asia tabulated according to latitude and longitude. There were, however, many errors in the *Guide*—*e.g.*, the Equator was placed too far north, and the value used for the circumference of the Earth was nearly 30 percent less than a more accurate value that had already been determined—as well as some contradictions between the text and maps. Moreover, as a whole, the *Guide* cannot be considered "good geography"; it does not mention anything about the climate, natural products, inhabitants, or peculiar features of the countries with which it deals, and Ptolemy's treatment of the geographical importance of such factors as rivers and mountain ranges is careless and of little use.

In spite of its faults, the *Guide* is an important work from a historical point of view because, like the *Almagest,* it exerted a great influence on later generations. Christopher Columbus, for example, used it to strengthen his belief that Asia could be reached by travelling westward because Ptolemy had indicated that Asia extended much farther east than it actually does. Even as late as 1775, it was believed that the Indian Ocean was bounded by a southern continent, as Ptolemy had suggested; the return voyage from the Southern Hemisphere of Capt. James Cook in July of that year proved otherwise.

Although it is not known exactly when Ptolemy died, Arabian traditions claim that he was 78 years old at the time of his death.

BIBLIOGRAPHY. There is no biography of Ptolerny. His astronomical work is described in some detail in J.L.E. DREYER, *A History of the Planetary Systems from Thales to Kepler* (1905; republished as *A History of Astronomy,* 2nd ed., 1953). A full English text of *Almagest* which, incidentally, shows how clearly Ptolemy expressed himself, is the translation by R CATESBY TALIAFERRO published in "Great Books of the Western World," vol. 16 (1952). Ptolemy's geographical work is discussed by J.O. THOMSON in *History of Ancient Geography* (1948); and his mathematics in THOMAS L. HEATH, *A Manual of Greek Mathematics* (1931, reprinted 1963). There are no popular texts about Ptolemy or his work; those mentioned above are somewhat specialized although all are readable by the nonspecialist. Ptolemy's work is also described in the context of early astronomy as it leads to modern times in COLIN A. RONAN, *Discovering the Universe* (1971), which is a more popular text.

(C.A.R.)

Ptolemy I Soter

Ptolemy I Soter, friend and general of Alexander the Great, was a man of remarkable military, diplomatic, and organizational abilities, who rose from the lower Macedonian gentry to become king of Egypt, where he ruled from 323 (king from 305) to 285/283 BC. The Ptolemaic dynasty, which he founded, reigned longer than any other dynasty established on the soil of the Alexandrian empire and only succumbed to the Romans in 30 BC. He introduced the cult of Alexander the Great, whom he buried in Egypt, and thus inaugurated a new epoch marked by the worship of the Hellenistic kings. He also originated the cult of Sarapis, a Greco-Egyptian deity, which outlasted his heirs to the throne by centuries.

By courtesy of the trustees of the British Museum

Ptolemy I Soter, portrait on a silver tetradrachrn. In the British Museum.

Early life and career

Ptolemy, born about 367/366 or 364 BC, was the son of the nobleman Lagus, a native of the Macedonian district of Eordaea whose family was undistinguished until Ptolemy's time, and of Arsinoe, who was related to the Macedonian Argead dynasty. He was probably educated as a page at the royal court of Macedonia, where he became closely associated with Alexander. He was exiled in 337, along with other companions of the crown prince. When he returned, after Alexander's accession to the throne in 336, he joined the King's bodyguard, took part in Alexander's European campaigns of 336–335, and in the fall of 330 was appointed personal bodyguard (*sōmatophylax*) to Alexander; in this capacity he captured the assassin of Darius III, the Persian emperor, in 329. He was closely associated with Alexander during the advance through the Persian highland. As a result of Ptolemy's successful military performance on the way from Bactria (in northeastern Afghanistan) to the Indus River (327–325), he became commander *(tridrarchos)* of the Macedonian fleet on the Hydaspes (modern Jhelum in India). Alexander decorated him several times for his deeds and married him to the Persian Artacama at the mass wedding at Susa, the Persian capital, which was the crowning event of Alexander's policy of merging the Macedonian and Iranian populations.

Ptolemy, who distinguished himself as a cautious and trustworthy troop commander under Alexander, also proved to be a politician of unusual diplomatic and strategic ability in the long series of struggles over the throne that broke out after Alexander's death in 323. Convinced from the outset that the 'generals could not maintain the unity of Alexander's empire, he proposed during the council at Babylon, which followed Alexander's death, that the satrapies (the provinces of the huge empire) be divided among the generals. He became satrap of Egypt, with the adjacent Libyan and Arabian regions, and methodically took advantage of the geographic isolation of the Nile territory to make it a great Hellenistic power. He took steps to improve internal administration and to acquire several external possessions in Cyrenaica (the easternmost part of Libya), Cyprus, and Syria and on the coast of Asia Minor; these, he hoped, would guarantee him miiitary security. Although he pursued a friendly policy toward Greece that secured his political influence there, be also succeeded in winning over the native Egyptian population.

Egypt of

In 322 Ptolemy, taking advantage of internal disturbances, acquired the African Hellenic towns of Cyrenaica. In 322–321, as a member of a coalition of "successors" (*diadochoi*) of Alexander, he fought against Perdiccas, the ruler (*chiliarchos*) of the Asiatic region of the empire. The coalition was victorious and Perdiccas died during the fighting. Ptolemy's diplomatic talent was put to the test during this war. When the satrapies were redistributed at Triparadisus in northern Syria, Antipater, the general of the European region, became regent of the Macedonian empire and Ptolemy was confirmed in possession of Egypt and Cyrene. He further strengthened his position by marrying Eurydice, the third daughter of Antipater.

About 317 he married Berenice I, the granddaughter of Cassander, the son of Antipater, who, at his father's death in 319, refused to accept the new regent, made war upon him, seized part of the empire, and in 305 assumed the title of king of Macedonia. In the coalition war of 315–311, Ptolemy obtained possession of Cyprus. In this war he scored his most important victory in the battle near Gaza in 312, in which the Egyptian contingents were decisive. But war broke out anew in 310, and he lost Cyprus again in 306. He temporarily lost Cyrene as well and was unable to hold the important Greek positions of Corinth and neighbouring Sicyon and Megara, which he had captured in 308. He ultimately suffered overwhelming defeat in 306 in the naval battle near Salamis on Cyprus. The victor in this battle, Antigonus I Monophthalmus, who was assisted by his son, Demetrius Poliorcetes, assumed the title of king in 306. The remaining satraps, led by Ptolemy after he successfully resisted Antigonus' attack on Egypt, also took the title of King in 305–304.

King of Egypt

After naming himself king, Ptolemy's first concern was the continuing war with Antigonus, which was now focussed on the island of Rhodes. In 304 Ptolemy aided the inhabitants of Rhodes against Antigonus and was accorded the divine title Soter (Saviour), which he was commonly called from that time. The dissolution of Alexander's empire was brought to a close with the battle near Ipsus in Asia Minor in 301. During this battle Antigonus was defeated by the other kings. This led to the attempt by the remaining successors of Alexander to define their kingdoms. For this reason a dispute arose between Ptolemy and Seleucus I Nicator of Babylon over Syria, particularly the southern Syrian ports, which served as terminal points for the caravan routes. This quarrel, however, was temporarily settled peacefully through compromise. In addition to Coele Syria (Palestine), Ptolemy apparently also occupied Pamphylia, Lycia, and part of Pisidia in southern Asia Minor.

During the last 15 years of his reign, because of the defeats he suffered between 308 and 306, Ptolemy preferred to secure and expand his empire through a policy of alliances and marriages rather than through warfare. In 300 he concluded an alliance with Lysimachus of Thrace (modern Bulgaria) and gave him his daughter Arsinoe II in marriage in 299/298. At approximately the same time he married his stepdaughter Theoxena to Agathocles, the tyrant of Syracuse (southeastern Sicily). About 296 he made peace with Demetrius Poliorcetes, to whom he betrothed his daughter Ptolemais. To Pyrrhus of Epirus, Demetrius' brother-in-law, who was at the Egyptian court as a hostage, he gave his stepdaughter Antigone. He finally brought rebellious Cyrene into subjection in 298, and in approximately 294 he gained control over Cyprus and the Phoenician coastal towns of Tyre and Sidon.

In a last coalition war in 288–286, in which Ptolemy, Seleucus, Lysimachus, and Pyrrhus opposed Demetrius, the Egyptian fleet participated decisively in the liberation of Athens from Macedonian occupation. During this war Ptolemy obtained the protectorate over the League of Islanders, which was established by Antigonus Monophthalmus in 315 and included most of the Greek islands in the Aegean. Egypt's maritime supremacy in the Mediterranean in the ensuing decades was based on this alliance.

Ptolemy was able to evaluate the chaotic international situation of this post-Alexandrian era, which was characterized by constantly renewed wars with shifting alliances and coalitions, in realistic political terms. Adhering to a basically defensive foreign policy, he secured Egypt against external enemies and expanded it by means of directly controlled foreign possessions and hegemonic administrations. He did not, however, neglect to devote attention to the internal organization of the country and to provide for a successor. In 290 he made his wife Berenice queen of Egypt and in 285 (possibly on June 26) appointed his younger son Ptolemy II Philadelphus, who was born to Berenice in 308, coregent and successor. The provision for the succession, which was based on examples from the time of the pharaohs, made possible a peaceful transition when Ptolemy died in the winter of 283–282. The early Ptolemies were occupied with the economic exploitation of Egypt, but, because of the lack of first-hand information, the details of Ptolemy's participation in the process cannot be determined. It is certain, however, that discrimination against the Egyptians took place during his reign. The only town he founded was Ptolemais in Upper Egypt. He probably placed Macedonian military commanders alongside the Egyptian provincial administrators and intervened unobtrusively in legal and financial affairs. In order to regulate the latter, he introduced coinage, which until that time was unknown in Egypt.

He found it necessary from the outset, however, to pursue a conciliatory policy toward the Egyptians, since Egyptians had to be recruited for his army, which initially numbered only 4,000 men. Ptolemy won over the Egyptians through the establishment in Memphis of the Sarapis cult, which fused the Egyptian and Greek religions; through restoration of the temples of the pharaohs, which had been destroyed by the Persians; and through gifts to the ancient Egyptian gods and patronage of the Egyptian nobility and priesthood. Finally, he founded the Museum (Mouseion), a common workplace for scholars and artists, and established the famous library at Alexandria. Besides being a patron of the arts and sciences, he was a writer himself. In the last few years of his life Ptolemy wrote a generally reliable history of Alexander's campaigns. Although it is now lost, it can be largely reconstructed through the extensive use made of it later by the historian Arrian.

Several times during his life Ptolemy was proclaimed a deity by certain classes of people. After his death he was raised to the level of a god by all the Egyptians.

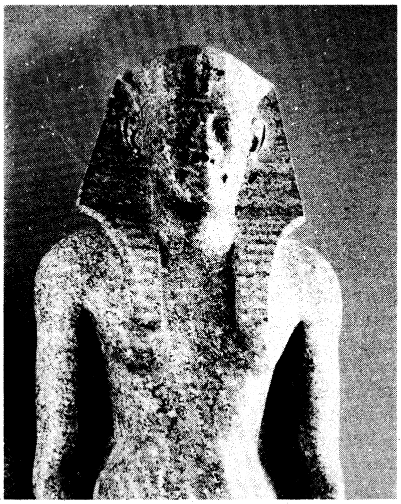
BIBLIOGRAPHY. E.R. BEVAN, The House of *Ptolemy:* A History of Egypt Under the Ptolemaic Dynasty, rev. ed. (1968), a general account, clearly arranged; MAX CARY, *A* History of the Greek World *from 323* to *146 B.C.,* 2nd ed. rev. (1951), a brief summary; P. JOUGUET, "La Politique intérieure du premier Ptolémée," *Bulletin de l'Institut Français d'Archéologie Orientale,* 30:513–536 (1931), a reliable, accurate article; JAKOB SEIBERT, *Untersuchungen zur Geschichte Ptolernaios' I.* (1969), includes the most recent scholarly research; W.W. TARN, *The Cambriage Ancient History,* vol. 6, ch. 15 and vol. 7, ch. 3 (1927–28), a classical presentation with bibliography.

(R.We.)

Ptolemy II Philadelphus

Ptolemy 11 Philadelphus, a masterful diplomat and a famous patron of the arts, was the first king of the Macedonian dynasty of the Ptolemies in Egypt to set up the Hellenistic divine cult of the living ruler. Born on the island of Cos in the Aegean in 308 BC, Ptolemy II was king of Egypt and its foreign possessions from 285 to 246 BC. Reigning at first with his father Ptolemy I Soter, he became sole ruler in 283–282 and purged his family of possible rivals. This dynastic strife led also to the banishment of his first wife, Arsinoe I, daughter of king Lysimachus of Thrace. Ptolemy then married his sister, Arsinoe II, an event that shocked Greek public opinion but was celebrated by the Alexandrian court poets. Taking advantage of the difficulties of the rival kingdoms of the Seleucids and Antigonids, Ptolemy II extended his rule in Syria, Asia Minor, and the Aegean at their expense and asserted at the same time his influence in Ethiopia and Arabia. Egyptian embassies to Rome as well as to India reflect the wide range of Ptolemy's political and commercial interests.

By courtesy of the Vatican Museum



Ptolerny II Philadelphus, upper portion of a colossal red granite statue from Heliopolis, c. 260 BC. In the Vatican Museum.

Militaıy leader

While a new war with the Seleucids (from 274 to 270) did not affect the basic position of the rival kingdoms, the so-called Chremonidean War (268?–261), stirred up by Ptolemy against Antigonus II Gonatas, king of Macedonia, resulted in the weakening of Ptolemaic influence in the Aegean and brought about near disaster to Ptolemy's allies Athens and Sparta. Ptolemy was no more successful in the Second Syrian War (c. 260–253), fought against the coalition of the Seleucid king Antiochus I and Antigonus Gonatas. The unsuccessful course of the military operations was compensated for, to a certain degree, by the diplomatic skill of Ptolemy, who first managed to lure Antigonus into concluding a separate peace (255) and then brought the war with the Seleucid Empire to an end by marrying his daughter, Berenice—provided with a huge dowry—to his foe Antiochus I. The magnitude of this political masterstroke can be gauged by the fact that Antiochus, before marrying the Ptolemaic princess, had to dismiss his former wife, Laodice. Thus freed for the moment from Seleucid opposition and sustained by the considerable financial means provided by the Egyptian economy, Ptolemy 11 devoted himself again to Greece and aroused new adversaries to Antigonid Macedonia. While the Macedonian forces were bogged down in Greece, Ptolemy reasserted his influence in the Aegean, making good the setback suffered during the Chremonidean War. He further improved his position by arranging for the marriage of his son (and later successor) Ptolerny 111 Euergetes to the daughter of King Magas of Cyrene, who had proved so far a very troublesome neighbour. Not aiming at outright hegemony (even less imperialistic conquest) in the Hellenistic world of the eastern Mediterranean, Ptolemy II tried nonetheless to secure for Egypt as good a position as possible, holding at large his rivals beyond a wide buffer zone of foreign possessions. Without being completely successful, he managed to let his allies bear the brunt of the heaviest reverses, healing his own military wounds with diplomatic remedies. The influence on Ptolemy of his wife and sister Arsinoe II, particularly in foreign affairs, was certainly substantial, though not as extensive as claimed by some contemporary authors.

Domestic innovator and patron of the arts

Ptolemy II's record in domestic affairs is no less impressive. From pharaonic times onward, agriculture and the work of artisans in Egypt had been highly organized. Under Ptolemy's supervision and with the help of Greek administrators, this system developed into a kind of planned economy. The peasant masses of the Nile Valley provided cheap labour, so that the introduction of slavery on a broad basis was never considered an economic necessity in Ptolemaic Egypt. Ptolemy II became a master at the fiscal exploitation of the Egyptian countryside; the capital, Alexandria, served as the main trading and export centre. Ptolemy II displayed a vivid interest in Greek as well as in Egyptian religion, paid visits to the sanctuaries in the countryside, and spent large sums erecting temples. Anxious to secure a solid position for, and religious elevation of, his dynasty, the King insisted upon divine honours not only for his parents but also for his sister and wife Arsinoe II and himself as *theoi adelphoi* ("brother gods"). He thus became one of the most ardent promoters of the Hellenistic ruler cult, which in turn was to have a far-reaching influence on the cult of the Roman emperors.

Under Ptolemy II, Alexandria also played a leading role in arts and science. Throughout the whole Mediterranean world the King acquired a reputation for being a generous patron of poets and scholars. Surrounding himself with a host of court poets, such as Callimachus and Theocritus, he expanded the library and financed the museum, a research centre founded as a counterweight to the more antimonarchial Athenian schools. Learning there was not confined to philosophy and literature but extended also to include mathematics and natural sciences. The age of Ptolemy II coincided with the apex of Hellenistic civilization; its vigour and glamour were a result of the still fresh forces of Greek leadership in the eastern Mediterranean. Ptolemy 11 was no man of peace, but neither was he one of the warlike Hellenistic soldier-kings. A prudent and enlightened ruler, he found his strength in diplomatic ability and his satisfaction in a vast curiosity of mind.

BIBLIOGRAPHY. H. VOLKMANN, "Ptolemaios II. Philadelphos," in PAULY-WISSOWA *Realencyclopädie,* vol. 46, col. 1645–66 (1959), provides an informative and up-to-date general account with extensive references to ancient sources and modern literature. W.W. TARN, *Antigonos Gonatas* (1913, reprinted 1969), puts the Macedonian king in the framework of the Hellenistic world and includes many statements on the policy of Ptolemy II. Relevant information may also be found in C. PREAUX, *L'Économie royale des Lagides* (1939), fundamental for an understanding of the economy of Ptolemaic Egypt; E. WILL, *Histoire politique du monde hellénistique* (323–30 *av. J.-C.*), vol. 1 (1966), the most recent survey of politics in the Hellenistic world, with a detailed analysis of Ptolemaic foreign policy (see pp. 133–186); G. LONGEGA, *Arsinoe II* (1968), an interesting biography (in Italian), but overstresses the infiuence of Arsinoe on her husband Ptolemy II; and H. HEINEN, *Untersuchugen zur hellenistischen Geschichte des* 3. *Jahrhunderts v. Chr.* (1972), which includes an extensive account of the Chremonidean War.

(He.H.)

Public Administration

Public administration, traditionally defined, comprises those activities involved in carrying out the policies and programs of governments. The term is also used today in a broader sense, for public administration is often regarded as including some responsibility—varying widely in degree among governments and departments—in determining what the policies and programs of governments should be, as well as in executing them. But public administration focusses principally on the planning, organizing, directing, coordinating, and controlling of government operations.

As an occupational field, public administration is common to all nations, whatever their system of government. Whether monarchical, totalitarian, socialist, parliamentary, or congressional-presidential, all countries require machinery to put into effect the policies of the government. Within nations public administration is practiced by the central government, by local governments, and, in federal systems, by intermediate provinces and states. The interrelationship of different levels of government within individual nations is in fact a continuing and growing problem for public administration generally.

Public administration as a profession

The professional identity of those engaged in public administration has been strengthened in European countries, and, by colonial imposition or diffusion, in most of the rest of the world, through the establishment of more or less elite administrative, executive, or directive classes in civil services. In the United States and a few other countries, the elitist class connotation was consciously abandoned or avoided, with the result that professional recognition has come slowly and only partially.

Public administration, as well as being a profession, is also a distinct field of study, and the middle decades of the 20th century have seen a considerable increase in research and in the education and training of public administrators.

Historical background. Public administration is an ancient activity. It must have existed for as long as there have been organized societies, for few group determinations in society are self-executing. The earliest records of civilization in south Asia, China, and Egypt contain references to what one would now call public administration.

Modem public administration is a by-product of the emergence of nation-states from the feudal societies of Europe (see also BUREAUCRACY). With the growth and centralization of power and responsibility in monarchical courts came the need for a full-time, stable, and qualified corps of public administrators, who became increasingly specialized in different fields of national activity. One of the earliest systematic manifestations of public administration was known as cameralism, which developed during the 17th and 18th centuries in Prussia and Austria. Cameralism was designed to provide efficient management of highly centralized, paternalistic states characterized by mercantilist economies. It required university training in such fields as public finance and its administration, police science, economics, and agriculture and forestry. The 18th century also witnessed the development in France of a high degree of emphasis and proficiency in technical and engineering activities and saw the establishment of national professional schools, primarily to provide qualified technicians for the public service.

Legalistic approach to public administration

Though some of the elements of cameralism and French technology have had significant impact on public administration in various countries down to the present time, the emphases upon them could not survive the French Revolution and the Napoleonic era. The weight given to the rights of individuals and the obligation of states to protect those rights, the introduction of laissez-faire economics, the codification of law, and other developments led to a quite different view of the state and of its administration. The essence of the latter came to be obligation and loyalty to the state through the interpretation and fair-handed application of law, legitimately enacted to express the will of the state. Such a view suggested that senior permanent officers should be trained in law. The nation-state was sovereign, centralized, and durable; and the legally oriented bureaucracy served the function of providing permanence and stability and of expressing and preserving its "will," despite changes of government and even systems of government. This legalistic view of the state and its bureaucracy persists in much of western Europe and, to a lesser degree, in parts of eastern Europe, as well as in many of the newer countries that were once colonies of continental powers.

Great Britain and the United States followed paths quite different from each other and from continental Europe in the development of their systems of public administration. Neither adopted the somewhat mystical European view of the state, and neither abandoned its common-law tradition for codification. Britain had long entrusted the administrative responsibilities of its government to representatives drawn from its aristocracy of unspecialized, often well-educated gentlemen. Until the Industrial Revolution of the late 18th and early 19th centuries, most of the aristocracy came from rural estates. Following the reform of the civil service in the 19th century, most administrators came to be drawn from the growing mercantile and business classes of the cities. For the last century, they have been selected primarily on the basis of stiff competitive examinations of university graduates, mainly from Oxford and Cambridge. These examinations tested neither administrative law, as on the continent, nor any other specialization directly related to public administration but concentrated on the classics and humanities. This method of recruitment to the British administrative class persisted, with only temporary changes during periods of crisis, into the late 1960s. It was designed to produce generalist administrators—intelligent, broad-gauged men free of parochial, professional perspectives. They would learn administration and the activities they were administering on the job. Administration was perceived more in terms of providing policy advice to ministers and less in terms of internal management than in most other countries. In general, British administrative practice was highly centralized, with controls over the system itself largely concentrated in the Treasury. In 1968 a committee of inquiry into the civil service under the chairmanship of Lord Fulton made recommendations that would result in much broader recruitment into the upper reaches of the civil service.

Through the Indian and other colonial civil services, the British concepts of administration were conveyed to the colonies, and they continue to have impact, though with varying force, in the Commonwealth.

Public administration in the American colonies and in the states and national government that succeeded them also began on the model of the mother country. Administration was by the gentry or landed aristocracy in the South, and by the increasingly commercial and industrial gentry in the North. Public administration was not perceived as a distinct and separable kind of activity or occupation, and the term was not used in the U.S. Constitution. There were three basic structural differences from the British system. One was the federal system and particularly the limited powers of the national government. Second was the separation of executive from legislative powers at the national, state, and city level of government. Third, growing out of the fear and distrust of concentrated executive powers over which the American Revolution had been fought, was the tendency to scatter executive power among a wide variety of commissions, boards, or elected or appointive officials who were either autonomous or responsible to the legislative body. Among other consequences of these structural differences are that to this day, the primary problems of American public administration are these: bridging the separation of executive and legislative powers; establishing more effective and cooperative relationships between and among the national government, the states, and a vast array of local governments; and integrating executive powers in single executives at each level.

Since the Revolution three major developments have significantly affected American public administration.

Major influences on the development of public administration

One has been the two-party political system and, especially during the middle decades of the 19th century, the invasion by political parties of the administrative offices of governments under the patronage system. This practice of rewarding political service with public office effectively thwarted the development of stable administrative systems and made the career prospects of administrators uncertain, for with each change in political control a new group of officeholders would replace the old. Party affiliation, rather than merit or ability, was the criterion for appointment. But the so-called spoils, or patronage, system had a corollary and more wholesome effect: it was a sweeping expression of egalitarianism in its weakening of the hold of the gentry on public offices and its opening of public service to the common man. This move toward egalitarianism preceded by more than a century steps in the sarne direction that were underway in many European and other countries in the 1970s. The depredations of the spoils system gave rise in the late 19th century to a second development: a fervent movement to reform the civil service on the basis of merit and without regard to political affiliation or to social class background. Civil service reform is still far from complete, but it has given rise to many of the features and developments of public administration throughout much of the 20th century. American civil service reform was originally molded on the British pattern, but its effects were entirely different, for it prevented the development of an elite administrative class.

The third development affecting the character of American public administration was the accelerating specialization, diversification, and professionalization of occupations that began about the turn of the 20th century in the United States and within its governments. This development took place all around the world, but its impact in the United States on public administration has been somewhat different. The public agencies of the United Slates had no entrenched amateur, gentlemen administrators as in Britain, nor administrative lawyers, as in continental Europe. Specialists assumed substantial control over the agencies appropriate to their skills: engineers for public works, agriculturists for agriculture, doctors for public health, lawyers for regulatory programs, and so forth.

Russia inherited some of the legal tradition associated with the Napoleonic era, acquired in its case largely thraugh international osmosis rather than occupation. The Russian Revolution of 1917 superimposed a very different set of values and institutions in public administration: subordination of the individual citizen in favour of dominance of the state; a high degree of concentration of power; the one-party system and party control not only of basic policy but of administrative and industrial agencies at every level. The goals of the state are determined by the Communist Party, and these have been primarily concerned with the expansion of political, economic, and military power. The means toward this end are seen as rapid industriaiization, high productivity, discipline, and the suppression of elements in society that might retard or endanger such development. The Soviet Union has thus moved in the direction of a technocratic administration. Its administrative leadership consists increasingly of engineers, production managers, and scientists. They are usually party members, but of a distinctly different type from the older revolutionaries. The Soviet Union may thus be regarded as providing a fourth main source of administrative thought and practice, its system of public administration contrasting strongly with the British, American, and ccntinental patterns. The importance of these four patterns of public administration has been greatly increased as a result of their impact on many other governments of the world.

An important factor in the spread of particular approaches to public administration is the education and training of future leaders in foreign countries, Many Communist political, as well as administrative, leaders around the world were trained in the Soviet Union. A good many public administrators in Africa, Latin America, and Asia were trained in Britain, the United States, and France. Finally, and related to such educational activity, is the conscious imitation of the administrative systems of other countries, often coupled with a request for assistance in establishing such a model. During the 19th century, for example, Japan imitated, and was assisted by, Germany; Thailand was aided similarly by Britain and France. Since World War II, most of the former colonies have been assisted and influenced by their erstwhile mother countries. The United States has had the largest technical assistance program, and American practices are consequently becoming increasingly prevalent. This may be due in part to the greater ability of the United States to provide such assistance and in part to a belief that the United States is most advanced in this field. Public administration is one of the major fields of technical assistance of the United Nations, and such assistance is also provided by a number of other international organizations and by some private foundations.

Factors hindering reform

Current efforts to improve public administration, which seem to be common to almost every country in the world, take different forms in different lands. Improvement is everywhere identified with greater effectiveness or productivity in responding to public needs, but particular facets of the different administrative systems retard such improvement in a variety of ways. The continental legal psychology, for example, is more conducive to stability than to innovation, and the dominance of a legal class has not encouraged the employment of other kinds of specialists and scientists. To a considerable extent the same is true of the British administrative-class generalists. Insofar as both groups are drawn principally from established upper class minorities, their domination is not coriducive to what has been called "democratization" of the public services. Ey way of contrast, a major problem of public administration in the United States is to coordinate the efforts of numerous separate departments at the federal, state, and local levels, each dominated by its particular brand of specialists. In addition, the long-standing American orientation toward business, free enterprise, and economics has discouraged the development of integrating mechanisms for social planning. And the severity and discipline of the single party in most of the Communist states has. according to many observers, provided disincentives toward either innovation or productivity.

Basic approaches. A central theme of public administration throughout the 20th century has been reform. That is, it has been assumed that current administrative processes can and should be improved. The study and practice of public administration has been essentially pragmatic and normative rather than theoretical and value free. This may explain why public administration, unlike some social sciences, developed without much concern about an encompassing theory. Not until the middle decades of the 20th century and the dissemination of the German sociologist Max Weber's theory of oureaucracy was much interest stimulated concerning a theory of public administration. Most of the recent development of bureaucratic theory, however, has been addressed to organizations in the private sector, and, with a very few exceptions, there has been little effort to relate organizational theory with political theory. The main emphasis remains upon pragmatic reform (see BUREAUCRACY).

A second theme of public administration has been economy and efficiency; that is, the provision of public services at the minimum cost, This has usually been the stated objective of administrative reform. Despite growing concern abcut other kinds of values such as responsiveness to public needs, justice and equal treatment, and citizen involvement in government decisions, efficiency continues to be a major goal.

A third feature of public administration has been its emphasis on the structure of formal organization. Most efforts at administrative reform have included as their central element the modification of organizational structures. They reflect a basic faith, held by administrators, politicians, and the educated public, that administrative ills can be at least partly corrected by reorganization guided by logical rules or "principles." Many of the principles were initially borrowed from the military, a few from private business. They include, for example: (1) organizing departments, ministries, and agencies on the basis of common or closely related purposes; (2) grouping like activities in single units; (3) equating responsibility with authority; (4) ensuring unity of command (only one supervisor for each group of employees); (5) limiting the number of subordinates reporting to a single supervisor; (6) differentiating line (operating or end-purpose) activities from staff (advisory, consultative, or support) activities; (7) employing the principle of management by exception (only the unusual problem or case is brought to the top); and (8) having a clear-cut chain of command downward and of responsibility upward.

Some critics have maintained that these and other principles of public administration are useful only as rough criteria for given organizational situations. They believe that organizational problems differ and that the applicability of rules to various situations also differs. Nonetheless, and despite much more sophisticated analyses of organizational behaviour in recent decades, such principles as those enumerated above continue to carry force.

A fourth feature of public administration has been its stress upon personnel. In most countries administrative reform has been preceded or accompanied by civil service reform. Historically, the direction of such reform has been toward what has been labelled "meritocracy"—the best individual for each job, competitive examinations for entry, and selection and promotion on the basis of merit. Attention has increasingly been given to factors other than merit, including personal attitudes, incentives, personality, personal relationships, and collective bargaining.

A fifth feature of public administration has been the development of the budget as a principal tool in planning future programs, in making decisions as to program priorities and the allocation of resources, in managing current programs, in linking executive with legislature, and in developing control and accountability. The public budget and thecontests for its control have along history, particularly in the Western world, that began with centuries of struggle between monarchs and representatives of their subjects for control over public finances. The modern executive budget system in which the executive recommends, the legislature appropriates, and the executive oversees expenditures originated in 19th-century Britain. In the United States during the 20th century, the budget became the principal vehicle for legislative surveillance of administration, for executive control of departments, and for departmental control of subordinate programs. It has been assuming a similar role in many of the developing countries of the world.

Recent developments. The classical approach to public administration described above probably reached its furthest development in the United States during the 1930s. Since that time, through educational and training programs, technical assistance, and the work of international organizations, it has become standard doctrine in many countries, although some of its elements have been resisted by governments with British or continental-legal perspectives. During the 1930s this rather simplistic, mechanistic approach was also challenged from a number of different directions. The field, in practice and in study, has since been greatly enlarged and enriched by new perspectives and insights. It has also been somewhat confused as a result of certain inconsistencies in approach.

Changes in the basic premises of public administration

The orthodox doctrine rested on the premise that administration was simply the implementation of public policies determined by others. According to this view, administrators should seek maximum efficiency but should be otherwise neutral about values and goals. During the Great Depression of the 1930s, and even more so during World War II, however, it became increasingly evident that most new policies originated within the administration; that policy and value judgments were implicit in most significant administrative decisions; that many administrative officials worked on nothing except policy; and that, insofar as public policies were controversial, such work inevitably involved administrators in politics. The presumed separation of administration from policy and politics was seen to be artificial. Since the 1930s there has thus been increasing concern in public administration about the determination of policies and the development and use of techniques to improve policy decisions. Although the concept of a value-free, neutral administration is regarded by many as a thing of the past, no fully satisfactory substitute has been offered. How to ensure that responsible and responsive policy decisions are made by career administrators, and how to coordinate their work with the policies of politically elected or appointive officials, remain vital problems, especially in democratic states.

A second challenge to the old concepts of public administration was a result of a direct response to the Depression and the accompanying responsibility of national governments to restore, stimulate, and stabilize their economies. With these responsibilities came new informational devices, including the system of national income accounting and the emergence of gross national product as a major index of economic health. The practices of fiscal and monetary policy have become established specializations of public administration. Economists occupy key posts in the administrations of most nations, and many other administrators must have at least elementary know!-edge of the economic implications of government operations. Great Britain, Sweden and other Scandinavian nations, and the United States were among the leaders in the development of tools and techniques for economic planning. Such planning has become a dominating concern of public administration in many of the developing countries.

Modem economists have introduced new methods of analyzing the costs and benefits of government programs. An extension of cost-benefit analysis known as Planning Programming Budgeting System (PPBS) was developed in the U.S. Department of Defense and was later adopted by civil agencies of the U.S. government during the 1960s. Though officially abandoned by the national government in 1971, it has been increasingly applied in U.S. states and cities. Most of the nations of western Europe, Japan, and some of the developing nations have introduced the PPBS system or some aspects of it, although sometimes under a different name.

Under a PPBS system long-range objectives are defined as specifically as possible, and alternative programs for the attainment of such objectives are analyzed and compared from the standpoint of their predicted costs and benefits. PPBS is objective, quantitative, and economic in its orientation. Whether or not it will survive as a "system" is conjectural, but its emphasis upon concrete evaluation and analysis of programs has made a significant contribution to decision making in the administrative sphere, and it seems likely that this influence will continue to grow.

PPBS has proved to be a forerunner of a more profound shift in emphasis in public administration toward systems and systems analysis and away from structures and processes. As indicated above, public administration has traditionally been concerned with the legitimacy and efficiency of the organizations and procedures by which public decisions are reached and executed. In the emerging systems approach, public problem areas—such as weapons systems, housing, employment, and education—have come to be perceived and analyzed as systems that consist of interacting parts and that are responsive in predictable ways to influences and "inputs" from outside. Public administrators require more and diierent kinds of data from those that are customarily available to governments, and new kinds of information systems providing data on the condition of society have been developed.

Most of the domestic activities of governments have been directed to the improvement of the quality of life in a society. Quantitative economic measurement has been useful to a certain extent in this respect, but the value of human life, of freedom from sickness and pain, of safety on the streets, of clean air, and of opportunity for achievement have hardly been measurable in monetary terms. Public administration has thus increasingly concerned itself with finding and developing better social indicators, quantitative and qualitative—that is, better indexes of the effects of public programs, and new techniques of social analysis.

A concurrent, but entirely different, development in public administration has been the movement known generally as human relations. Its seeds were also planted in the 1930s when research, conducted over a period of several years and involving the workers and management of an industrial plant near Chicago, was reported. The research brought out, among other things, the importance of social or informal organization, of good communications, of individual and group behaviour, and of attitudes (as distinct from aptitudes). Concurrent and subsequent studies, especially in Great Britain and the United States, explored group dynamics, the process of attitude change, and differing styles of leadership and their effects.

The human relations movement was initiated by interest among social scientists—psychologists, social psychologists, sociologists, cultural anthropologists, and psychiatrists. During the middle decades of the 20th century their views had an increasing impact upon business management and a somewhat less but still significant influence upon public administration. The human relations approach raised questions concerning many of the oldest dogmas of administration: hierarchy; directive styles of leadership; clear and set definitions of duties; treatment of employees as impersonal "units" of production; and monetary incentive systetns (see also INDUSTRIAL AND ORGANIZATIONAL RELATIONS).

By the early 1970s the human relations approach had developed into a field known as "orgznization development." Its primary goal and orientation was tcward change: change in the attitudes, values, and structures of organizations so as to better adapt them to rapid change in their environment and in the demands placed upon them. Its tools included the use of trained consultants, usually from outside the organization, intensive interviewing of managers and employees, sensitivity training, confrontation meetings, and many others. Unlike the rationalistic PPBS approach, organization development stressed the identification of personal with organizational goals, the "self-actualization" of workers and managers, effective interpersonal communication, and broad participation in decision making. Its direct use within governmental agencies has been limited and has not always been successful, but it has had considerable indirect influence upon the way administrators think and behave.

Another modern movement in public administration, which is at least philosophically related to organization development, has been directed toward the greater participation of citizens in governmental decision making and action. It was stimulated during the 1950s and 1960s by a growing feeling that governments were not responding to the needs of their citizens, particularly minority groups and the poor. This disaffection contributed to widening protests against public, agencies and institutions. A variety of experiments to involve citizens or their representatives in governmental decision processes were begun in the 1960s that involved the delegation of decision-making power from central to local offices and, at the local level, the sharing of authority with citizen groups. This challenge to governmental bureaucracy could lead to fundamental changes in the nature and style of public administration. At the least it will probably contribute to greater concern on the part of administrators about the effects of their actions on citizens.

Comparative public administration. Until World War II there was rather little cross-filtration or even communication of public administration between nations except, as noted earlier, within the confines of imperial systems. As early as 1910 a professional organization, which eventually became the International Institute of Administrative Sciences (IIAS), was established. At first, however, its membership consisted principally of scholars and practitioners of administrative law in the countiies of continental Europe. In the early 1980s the IIAS had a membership drawn from about 50 countries. Its triennial congresses have covered all aspects of the field.

The period since World War II has witnessed an extraordinary burst of interest in the administrative systems of other countries. This was precipitated by the necessity of cooperation among the allied countries during the war; by the formation of international organizations; bv the occupation of conquered nations and the administration of economic recovery programs for Europe and the Far Bast; and by aid and technical assistance programs for developing countries. One by-product of aid and assistance programs was a renewed appreciation of how crucial effective administration was to national development. Another by-product was a realizztion of how parochial and culture-bound public administration thinking and practice had often remained within individual countries.

International interest in public administration

Another effect of this international communication and sharing of experiences has been the realization that development is not exclusive to the so-called underdeveloped countries. All countries have continued to develop, and public administration has increasingly been perceived as the administration of planned change in societies that themselves have undergone rapid change. not all of it planned. Government has no longer been merely the keeper of the peace and the provider of basic services: in the postindustrial era government has become a principal innovator, a determinant of social and economic priorities, and an entrepreneur on a major scale. On virtually every significant problem or challenge—from unemployment to clean air—people have looked to the government for solutions or assistance. The tasks of planning, organizing, coordinating, managing, and evaluating modern government have likewise become awesome in both dimension and importance.

Education and training for public administration. It is probably a safe judgment that every nation in the world has suffered a shortage of well-trained public administrators. European universities have been producing administrative lawyers for their governments for more than a century and a half, but legal skills are hardly adequate for handling contemporary problems. U.S. universities more than half a century ago began graduate programs, and by the 1970s there were moie than 60 university programs in public administration. Their total output has been, however, insufficient to meet the needs of government, and very few of the scientists and other specialists who become administrators in their fields attend such programs.

Shortage of trained public administrators

There has been a burgeoning of training programs around the world since World War II, many of them set up or sponsored by national governments, and some of them attached to universities. As one of its civil service reforms of 1946–47, France established an École Nationale d'Administration, which provided an extensive course for recruits to the higher civil service. It was not until 1969 that Britain established a Civil Service College under the new Civil Service Department. In the United States, the government established a variety of educational and training programs during the 1960s, including the Federal Executive Institute and the Executive Seminar Centers. Many of the developing countries have established one or more schools or training institutes, and by 1970 centres for the training of public administrators existed in more than 100 countries.

BIBLIOGRAPHY. The most comprehensive treatment of' administration is B.M. GROSS, ***The Managing of Organizations: The Administrative Struggle,*** 2 vol. (1964). Among the classics in the field are F.W. TAYLOR, ***The Principles of Scientific Management*** (1911; reprinted in ***Scientific Manugernent.*** 1947); HENRI FAYOL, ***Administration industrielle et générale*** (1916 and 1941; Eng. trans.. ***General and Industrial Management,*** 1949); and MAX WEBER, ***Wirtschaft und Gesellschaft*** (1922: Eng. trans. of pt. 1, ***The Theory of Social and Economic Organization,*** 1947, reprinted 1964).

The traditional approach to public administration and its principles were set forth in L.H. GULICK and L. URWICK (eds.), ***Papers on the Science of Administration*** (1937); and L. URWICK, ***The Elements of Administration,*** 2nd ed. (1947). Challenges to the principles, as well as efforts to build a theory of decision making as central to administration, appeared in C.I. BARNARD, ***The Functions of the Executive*** (1938, reprinted 1958); and H.A. SIMON, ***Administrative Behaviour: A Study of Decision-Making Processes in Administration Organization*** (1947). DWIGHT WALDO provided a thoughtful review of the evolution of public administration in its relation to society in ***The Administrative Stare: A Study of the Political Theory of American Public Administration*** (1948). The challenge to the traditional dichotomy between policy and administration was

expressed cogently in various works of P.H. APPLEBY, most notably in his ***Policy and Administration*** (1949). Later developments of similar views are contained in D.B. TRUMAN, ***The Governmental Process*** (1951, 2nd ed. 1971); E.S. REDFORD, ***Democracy in the Administrative State*** (1969); and H. SEIDMAN, ***Politics, Position and Power: The Dynamics of Federal Organization*** (1970). The economic systems approach to public program decisions and PPBS are set forth by two of their principal architects: C.J. HITCH, ***Decision-Making for Defense*** (1965); and C.L. SCHULTZE, ***The Politics and Economics of Public Spending*** (1968). The prophet of the human relations movement was Mary Parker Follett, some of whose writings were published by H.C. METCALF and L. URWICK (eds.), in ***Dynamic Administration: The Collected Papers of Mary Parker Follett*** (1960). More recently, the derivative movement (now called organization development) has been treated by C. ARGYRIS, ***Integrating the Individual and the Organization*** (1964); RENSIS LIKERT, ***New Patterns of Management*** (1961); and W.G. BENNIS, ***Organization Development*** (1969). A significant and growing literature has been developed since World War II in case studies of actual administrative experience. Pioneered and led by the American Inter-University Case Program, the use of cases has spread to many other countries. An example of the use of cases in comparative analysis is F.C. MOSHER (ed.), ***Governmental Reorganizations: Cases and Commentary*** (1967).

On the administrative systems of different countries, the following are of special note: B. CHAPMAN, ***The Profession of Government*** (1959); F.F. RIDLEY (ed.), ***Specialists and Generalists: Civil Servants at Home and Abroad*** (1968); F.W. RIGGS, ***Administration in Developing Countries*** (1964); M. BERGER, ***Bureaucracy and Society in Modern Egypt*** (1957); and J.G. LaPALOMBARA (ed.), ***Bureaucracy and Political Development*** (1963).

(F.C.M.)

Public Debt

Kinds of public debt

Public debt may be defined as sums owed by governments, in the form of bonds, notes, bills, etc., requiring the payment of specified amounts of money to the holders at designated times. For the most part public debt differs from private debt only in that it is an obligation of government rather than of private individuals or corporations. Public debt may be classified according to various criteria.

Internal and external debt. If the debt is held by persons outside of the issuing jurisdiction, it is called external; if it is held within the jurisdiction, it is called internal. The U.S. national debt is almost entirely internal, while the debts of many developing countries and of subordinate units of government in the United States are largely external.

Maturity period. Public debt ranges in maturity from infinity to periods of a month or even a few days. Debt instruments without a maturity date, requiring merely the payment of interest, are often called consols. The name originated in Great Britain, where the first important indeterminate-period debt issue was one that consolidated a number of separate issues.

A large portion of government debt consists of bonds with specific maturities of five to 99 years or more. Twenty- and 30-year periods are common. These are often known as long-term or funded debt.

Debt of maturity less than five years is often called short-term or floating debt and may take several forms: notes, with maturities from one to five years; treasury bills, with maturities from one month to a year and often sold at auction; and certificates of indebtedness, with similar maturities but available at a fixed interest rate.

The length of the maturity period affects what is known as the liquidity of the debt—*i.e.*, the quickness with which it can be converted into money. Securities with very short maturity periods are constantly repayable in money and thus have maximum liquidity or "moneyness." As the period maturity increases, the liquidity falls and the pure debt characteristic increases; consols have minimum liquidity because they have no date of maturity.

Type of issuer. Government debt may be directly issued by a government or by semi-autonomous governmental organizations. Examples of the latter would include railways, provincial power authorities in Canada, and various federal lending agencies in the United States. Their issues may be guaranteed by the government (general obligation bonds) or may rest solely upon the enterprises themselves, to be paid out of their revenues. In the United States this type of obligation is known as a revenue bond.

Marketability. The great bulk of all government debt consists of marketable securities. These securities are negotiable and sell freely on the market. They are usually issued in relatively large denominations, $1,000 or higher, and interest is paid periodically by check or coupon. Since they are salable, their price fluctuates from time to time, going above maturity value when the current market interest rate falls below the interest rate that they bear, and falling below the maturity value when the current rate rises or when fear about the ability of the government to pay interest develops.

Other bonds bought by the public are not marketable but can be redeemed, after a specified period, for their principal plus accrued interest. Various savings bonds, including those of the United States, are of this kind.

Other characteristics. Bondholders may receive current interest either by redemption of coupons attached to the bonds or by check from the government. Alternatively, interest may be receivable only upon maturity or redemption of the bond, as in the case of savings bonds. Interest and principal are usually payable in fixed monetary units, but they may be payable in amounts with fixed purchasing power based on changes in price levels.

THE ECONOMICS OF PUBLIC DEBT

The effects of debt financing upon the economy

Government borrowing is likely to have effects upon the economy that are substantially different from those of other methods of financing, and the existence of a sizable debt may likewise have important consequences. The effects of retiring the debt may also be significant. National government borrowing has the greatest impact, but that of subordinate units may also have some influence.

Borrowing. Government borrowing in the strict sense includes only borrowing from the private sector of the economy—from individuals, corporations, and various financial institutions, including banks. When the government obtains its funds from the central bank (the Bank of England, the Bank of Italy, the Bank of Japan, or the Federal Reserve System in the United States) it is really creating money rather than borrowing it, since the purchasing power is made by the central bank and no obligations to the public are created.

When a government borrows, funds are transferred from the lender to the government, the lender exchanging its money for government securities. The effect is to reduce the liquidity of the lender—its command over cash—to an extent dependent upon the nature of the securities. The reduction in liquidity is small with short-term securities and greatest with nonsalable, nonredeemable securities—a type seldom issued.

Funds loaned to the government almost certainly come from savings, unlike, for example, funds paid in higher taxes, which are more likely to come out of consumption. To pay higher taxes, many individuals are forced to reduce their consumption since they have no margin of savings and are unable or unwilling to go into debt; others do so as a matter of choice, in an effort to keep their savings intact. Lending, on the other hand, is entirely voluntary. People who buy government securities are not likely to increase their rate of saving in order to do so or to decrease their consumption. If government borrowing raises the market rate of interest, this may in turn encourage the diversion of additional money to saving, as may government securities that offer additional attractions—such as small denominations or redeemability—not possessed by other securities. Both effects in total, however, are not likely to be significant.

The net effect of government borrowing on total spending and thus on employment and national income will depend upon its influence on real investment—the purchase of new capital goods. In a period of unemployment, when savings will be available in greater quantity than investors will be prepared to use, government borrowing will not compete with private investment nor make it more costly. In effect, the government will be absorbing funds that would otherwise be idle.

In periods of full employment the situation is substantially different. With banks loaned up to the limit of their reserves and real investment absorbing all of savings, government borrowing will restrict private spending as much as an increase in taxation will under the same conditions.

Government borrowing is of economic significance in several other respects. First, the buying and selling of government securities provides the central bank with a means of influencing the money supply, essential for effective monetary policy. Second, borrowing avoids the adverse effects that taxes may have on incentives, particularly if the taxes are raised sharply above levels to which persons have become accustomed. Third, borrowing permits government expenditures to be higher than would otherwise be feasible, as explained in a subsequent section. Finally, the foreign borrowing of some governments gives them access to a greater quantity of foreign exchange, which enables them to finance the import of capital goods essential for economic growth. This consideration is not of concern to highly developed countries.

Effects of the debt. The existence of a government debt is of economic significance in itself, as distinct from the effects of the borrowing. In the first place, individuals who hold government securities regard them as a portion of their personal wealth and accordingly consider themselves to be wealthier than they would if they had only tax receipts. This may make them spend more on consumption and save less than they otherwise would. The additional consumption may reduce the rate of capital formation and economic growth; it may also increase the level of employment over what it would otherwise be.

Second, because government securities are more liquid than most other investments, their holders are able to increase consumption out of accumulated savings more easily than they could otherwise. This may contribute to inflationary pressures.

Third, it is often argued that a large national debt makes equity capital (stocks) somewhat cheaper. The larger amount of fixed-income securities will lead to a relatively greater willingness to invest in stocks, given investors' relative preferences between loan and equity investments.

Fourth, if investors, and particularly the business community, regard the national debt as a source of potential economic instability, their willingness to undertake real investment will be lessened. At times, particularly in the 1930s, there has been widespread fear of government debt even though there was, in reality, little basis for the fear. A similar phenomenon, with more logical foundation, may arise in the case of subordinate units of government. A large debt may discourage expansion of economic activity because of the fear of high taxes in the future and the realization that the large debt may prevent borrowing for urgently needed local improvements.

When governments borrow they must meet interest obligations, and these are usually paid out of taxes. The payment of interest on government debt thus involves a transfer of wealth from taxpayers to bondholders. The taxes may have adverse effects upon incentives while receipt of the interest will provide no offset to these adverse effects. The tax-and-interest-payment program is also likely to redistribute wealth in favour of higher income groups, since government bonds are likely to be held to a greater extent by those groups. The effect may be to increase saving and reduce consumption.

Finally, large interest obligations lessen the ability of the government to finance other governmental activities. This effect is particularly obvious at the local level, where there are limited tax potentials.

Retiring the debt. The retirement of debt has effects opposite from those of borrowing. Bondholders receive money in exchange for their bonds; though they could increase their consumption, they are more likely to put the funds into other securities and, as a consequence, security prices will rise and money capital will become more readily available for business investment. Whether it will be used for that purpose depends, of course, on the general economic situation.

The money for retirement must be obtained from some source. If it is simply created, there will be no repressive effect on consumption or investment, and total spending in the economy will rise—although by an amount relatively small compared to the total retirement. If, as is more common, the debt is retired from tax revenues, consumption will be reduced in substantial measure; the remainder of the tax will be absorbed from savings, and real investment may be reduced. The net curtailment in spending from the program of debt retirement is almost certain to reduce total spending in the economy. Elimination of the debt has one other effect: while current taxes will be increased, future taxes required to meet interest and principal obligations will be reduced.

The burden of the debt. It is often said that borrowing shifts the burden of governmental activities to future generations, since those generations will be assessed higher taxes to pay the interest and principal. Economists generally regard the argument as invalid, for future generations will inherit both the bonds and the obligations to pay them and collectively will be neither richer nor poorer than if the debt had not been incurred, except as a result of the difficulties incident to the debt and its retirement noted in preceding sections. Regardless of the methods of financing, the real cost of any governmental activity, war or otherwise, is borne in the form of reduced consumption and investment and harder work or the like during the period in which it is carried on. The only burden on the future is that arising from the depletion of natural resources or slower rates of investment, and these are not affected by methods of financing.

The effects of borrowing upon future generations

Several writers have pointed out that this reasoning ignores one significant element: if governmental activities are financed by borrowing, consumption will be reduced less and savings more than with tax financing, and also, because of the existence of the debt, persons will continue to consume more and save less. Assuming full employment, the rate of capital formation will therefore be less than with tax financing; future generations will inherit a somewhat smaller stock of capital goods, and per capita real incomes will be somewhat less than they would have been if tax financing instead of borrowing had been used.

Some writers question the doctrine that the burden, even apart from its effects on capital formation, is not shifted forward to the future. They argue that the burden can be thought of as meaningful only in terms of individual satisfaction. In the period in which the borrowing occurs, according to this view, the purchasers of government bonds obviously suffer no burden, in the sense of loss of personal satisfaction, because they bought the bonds voluntarily, and, having exchanged money for the bonds, presumably regard themselves as better off. But, in the future, when the debt is repaid, taxpayers will suffer a loss since they must pay larger amounts of taxes than they otherwise would and will therefore lose satisfaction, while the bondholders will experience no net increase in satisfaction since they will merely re-exchange the bonds for money. Thus a net decline in personal satisfaction will have occurred in the future generation and the burden will actually have been shifted to the future.

The difference between these two lines of reasoning hangs upon the definition of burden. If burden is defined as the reduction of output in the private sector of the economy when resources are transferred to the public sector, obviously the burden occurs at the time the governmental activities are undertaken. If burden is defined in terms of individual sacrifices, the burden occurs in the future. The first definition is useful in stressing the fact that resources cannot be pulled from one generation to another (except through the effect on the rate of capital formation), and that therefore borrowing does not make real income greater in the present and less in the future. The second definition is relevant if persons believe that borrowing does shift the burden to the future, for then, during the period of the borrowing, fewer private-sector goods will be available for consumption and capital formation, but individuals will experience an equivalent increase in their personal wealth in the form of government bonds. Thus as long as they disregard the future tax liabilities created by the borrowing, they do not feel a burden.

The argument of the preceding paragraphs is applicable, of course, only to domestically held national debt. Foreign-held debt gives a claim to foreigners against national output; payment of interest and principal will require the export of goods and, thus, a reduced domestic standard of living. Similar reasoning applies to state or local debts held outside the community, in which cases payment of interest and principal will constitute a decline in the real incomes of the people of the area as a whole, and the burden will thus be effectively shifted to future generations.

SHOULD GOVERNMENTS BORROW

Objections to borrowing. The desirability of government borrowing has been debated for centuries. The traditional argument against borrowing is, of course, the interest burden to which it gives rise, an argument applicable equally to private and governmental borrowing. These interest obligations require either higher levels of taxes, with possibly adverse effects on the economy, or reduced expenditures for other purposes. The payment of interest may easily result in a transfer of purchasing power to higher income groups, contrary to accepted standards of equity.

Arguments for and against government borrowing

As noted in the preceding section, the financing of expenditures by borrowing instead of taxation and the debt itself, once incurred, tend to produce higher prices and other inflationary effects in periods of full employment because they increase total spending. During periods of full employment, any increase in government expenditures not offset by an equivalent decline in private spending for consumption or business expansion will be inflationary. This is the usual argument made against the use of borrowing instead of taxation from the standpoint of the goal of economic stability. It is primarily relevant to national government borrowing because the national government must assume the primary responsibility for lessening economic instability. But state and local borrowing is, of course, equally inflationary.

Borrowing, if freely employed, can easily lead to increases in government expenditures beyond levels regarded by society as the optimum and may reduce the pressures for efficiency and elimination of waste. As governments consider expenditure levels, the adverse reaction to taxation serves as an offset against the favourable response to increased services that will have to be paid for by taxation and thus facilitates the attainment of a balance between government-produced services and privately produced services. But if borrowing replaces taxation and is generally accepted as a suitable routine method of financing, the pendulum will swing in the direction of increased governmental activity, and appropriate balancing wlll be lost. The best evidence of this danger is to be found in the history of state and local government finance in the early 19th century in the United States, when large sums of money were borrowed for purposes of limited usefulness to society. The basic trouble with borrowing is its relative painlessness as long as people disregard the future consequences.

There is, of course, the further possibility that a government may accumulate so much debt that it will lose the confidence of the people and may reach a point at which it cannot meet its obligations. Loss of confidence will make borrowing of additional funds in an emergency difficult and may check economic development by lessening optimism on the part of businessmen. If a government goes so far into debt that it cannot meet its obligations, its credit is seriously impaired. A national government, with its extensive taxing potentialities and its control over the banking system, is hardly liable to find itself in a position in which it cannot meet interest obligations or sell bonds. But if debt were to exceed certain levels, the government might be unable to sell bonds extensively to individuals and thus be forced to resort to borrowing from the central bank, the most inflationary form of financing.

These arguments are believed to constitute a significant case against general reliance on borrowing. The case is felt to be strongest in periods of full employment; for local governments, when it exceeds levels at which interest can be met easily from current sources; and in all circumstances in which governments are relatively irresponsible with regard to future problems created by present policies.

The case for borrowing. There are three primary circumstances in which government borrowing is usually regarded as justifiable:

1. In a period of unemployment or depression or both, the basic argument against the use of government borrowing—that it tends to generate inflation—is not relevant. At such times, any expansionary effect that the financing of government expenditures by borrowing instead of taxation may have will tend to bring about an increase in output rather than an increase in the general price level. Accordingly, the use of borrowing may be regarded as justifiable, in terms of the goal of lessening the severity of depressions and attaining greater economic stability. However, the other objections against borrowing may still remain and suggest the need for a restraint on the borrowing, particularly at the level of units smaller than national governments. Indeed some persons regard these objections as serious enough to overrule the justification for depression borrowing as a means of encouraging recovery. But this has now become a minority point of view. In a severe depression borrowing could be avoided only by a sharp curtailment in activities and an increase in taxes, all of which would not only aggravate the depression but would be politically intolerable. Even if the point of view that governments cannot succeed in bringing about recovery from depression by deliberate action is accepted, it must still be recognized that some national reliance on borrowing in depressions is inevitable and justifiable to avoid aggravating the depression.

An alternative to borrowing in time of depression is the creation of money through the central bank. The government borrows from the central bank, which creates the required funds and thus enlarges the money supply. Money creation offers two significant advantages: no obligations to repay debt in the future are created, and there is no direct depressing effect on private spending. The funds spent by government are merely created, rather than being taken from anyone. A secondary advantage is the increase in commercial bank reserves that results from the process; the banks are made more liquid and can expand loans, thus stimulating investment.

In practice, however, governments have been unwilling to use money creation as a means of financing deficits in periods of unemployment. The primary reason for this is the widespread belief that money creation is irresponsible and a source of inflation. This fear arises from the black memories of runaway inflations that have occurred in various countries in the past when governments resorted to excessive creation of money. But with responsible use of the technique, most economists believe that the dangers are slight.

2. The use of borrowing is regarded as inevitable in periods of major war. If taxes were increased sufficiently to finance all war costs, they could seriously impede the war effort by impairing incentives to work and by reducing the overall morale of the people. The limits of economically and politically tolerable taxation may well be below the maximum feasible allocation of resources to the war effort. Adequate tax increases would also aggravate the inequities of the tax structure; an overall level that would reduce total consumer spending to a level equal to the rate of output of consumer goods might well push some persons below subsistence levels and make it impossible for others to meet fixed commitments. While the use of borrowing as a method of war finance makes the control of inflation more difficult, there appears to be no escape from the necessity.

3. When governmental activities require capital outlays far in excess of usual expenditures and of a nonrecurrent character, then borrowing is not only virtually imperative if the outlays are to be made but is entirely justifiable. This rule is of primary concern to local governments because at higher levels the ratio of nonrecurrent expenditures to total expenditures is such that tax financing is possible without undue fluctuation in tax rates or hardships to taxpayers in particular years. But local governments must often make expenditures far in excess of usual annual income on projects that will last over a period of years. In such cases, the failure to use the loan method would tend to hold expenditures for such purposes below the optimum level because taxpayers will fight such sharp increases in taxes, particularly if the capital outlay will tend to benefit taxpayers in future years. The argument is particularly strong in the case of self-liquidating projects such as power or water system expansions and is valid even when the outlays are to be paid off out of tax revenues. It must be emphasized that the characteristic feature justifying the use of borrowing is the nonrecurrent element; the mere fact that the object of a particular expenditure will last a number of years is not in itself a justification if roughly the same amounts will be spent each year. If a city is to repave 20 blocks of streets annually, there is no need for borrowing, and a good case can be made against it, even if each particular repaved street may last 25 years.

A final justification for borrowing, relevant only to short-term loans, is a lack of exact correlation in the time when revenues are received and expenditures are made. Unless a government is to build up a surplus, it must borrow occasionally to meet expenditures at a date earlier than the receipt of the tax revenue.

Debt limitation laws. Efforts have been made in some countries to set restrictions on government borrowing through legislative acts.

Efforts to restrict government borrowing

In the United States, fear of excessive borrowing has resulted in restrictions on the amounts the executive, and even the legislative branches of government, can borrow. When many states found themselves in financial difficulties after borrowing heavily to provide funds for canals and railroads in the middle of the 19th century, public debt provisions were written into the constitutions of all but seven states. The provisions limiting borrowing differ widely. In most jurisdictions a maximum, usually expressed as an absolute dollar sum and one relatively low in terms of present-day expenditure levels, is set. Either this figure cannot be exceeded at all (except by amending the constitution) or it can be exceeded only with the approval of the voters at an election. In some places all bond issues require approval by popular vote and in some instances by more than a bare majority. The purposes for which funds may be borrowed and the duration of the issue are also frequently restricted. These constitutional restrictions have unquestionably lessened state borrowing; in so doing they have, perhaps, reduced waste, but they have also sometimes prevented urgently sought improvements. The limits have likewise greatly increased the use of revenue bonds, which are normally not subject to the restrictions. Unfortunately, the interest rate on these bonds is higher than the rate on other bonds.

Restrictions on municipal borrowing in the United States are almost universal. The restrictions, established either in the state constitutions or by state legislation, limit the total sum to be borrowed by any particular unit to a certain percentage (from 2 percent to over 20 percent) of the total assessed value of its property. The limits vary for different types of local units (city, county, school district, etc.). They usually do not apply to debts incurred for self-liquidating enterprises. In many states every bond issue must be approved by popular vote, in some instances by a two-thirds majority. In other states the limits established may be exceeded by popular vote, often with a requirement beyond a mere majority. Legislation also controls maximum interest rates that may be paid, the duration of the issues, the purposes of the borrowing, the establishment of means of retiring the bonds, etc. Several states exercise review over local bond issues, but only North Carolina requires specific state approval of all issues. Like the states, the local governments have found means of escaping the restrictions. Special taxing districts with their own debt limits are often formed when a city has reached its limit. Revenue bonds are also employed. In some states, such as Pennsylvania, there has been widespread creation of special authorities. A special school building authority, for example, is established with the power to finance the building of schools by issuing revenue bonds. In turn, the authority pays interest and principal on the bonds from rentals obtained from the school districts using the buildings.

While there are no constitutional limits on federal borrowing powers in the United States, Congress for many years has restricted borrowing by the Treasury Department. Before 1917 borrowing was permitted only upon specific authorization by Congress. After 1917 maximum figures were set at first for each type of loan and then, after 1938, as an overall total. The 1938 figure of $45,000,000,000 was gradually increased to a high of $300,000,000,000 in 1945 and reduced to $275,000,000,000 in 1946. Buttressed by a strong belief in Congress that failure to raise the limit would check growth in government spending, the limit remained at the 1946 level until 1954. Eventually, pressure on the limit became so great that various government bodies such as lending agencies were forced to borrow on their own at higher interest rates, and a series of increases was made in the 1960s. The limit was $377,000,000,000 as of Jan. 1, 1970. Experts differ in their estimates of the usefulness of the federal limit. Some believe that it curtails government waste and unjustified increases in expenditures, while others argue that it reduces flexibility in meeting emergencies, checks needed increases in various activities, could prevent quick action to stave off a depression, and leads to uneconomical forms of borrowing.

In Canada, neither the dominion nor provincial governments are subject to debt limitations. Local government limits are comparable to those in the United States, and in several provinces bond issues must receive the approval of a provincial agency. In Great Britain, borrowing by local governments is subject to control, specifically by the Ministry of Housing and Local Government, and limits are usually established in terms of a ratio of debt to total ratable value (assessed value of property). After World War II much local borrowing was channelled through the Public Works Loan Board, and thus was subject to additional control.

EVOLUTION OF GOVERNMENT BORROWING

The history of government debt in various countries

The evolution of government borrowing was very slow. The extensive use of loans by governments became possible only after the ruler had become differentiated from the state and after the fact of the continuity of the state had been separated from the persons of the rulers. Other factors were also required: the development of a regular revenue source to provide funds for repayment of loans, a monetary system, and an organized money market. The earliest loans of medieval times were either forced loans or personal borrowing by the sovereign. Government borrowing in its modern form first occurred in medieval Genoa and Venice when the city governments borrowed on a commercial basis from the newly developed banks.

France. Throughout much of French history public borrowing has been of major dimensions. Ministers of finance in the 17th and 18th centuries found the problem of managing the debt almost insuperable. During the Revolution that began in 1789, about two-thirds of the accumulated debt was repudiated, and the remainder was refunded in new securities in 1800 at a total of 926,000,000 francs. The sum increased by only 340,000,000 francs during the Napoleonic period because Napoleon's military expenditures were financed mainly by foreign levies. A large increase occurred during the Second Empire, when the debt rose from 5,516,000,000 francs in 1852 to 12,310,000,000 francs in 1870. The Franco-Prussian War, which ended in defeat for France, and the consequent imposition of an indemnity of 5,000,000,000 francs by the victorious Germans raised the French public debt to over 21,000,000,000 francs in 1873. Most of the increase was financed by four bond issues. After 1878 the debt increased further as a result of public works expenditures and France's colonial expansion until it stood at 34,204,000,000 francs at the outbreak of World War I. The war and its effects multiplied the debt, although at the same time inflation reduced the value of the franc by half.

The inflationary trend continued throughout the interwar years, and by 1960 the franc had lost more than 99 percent of its 1914 value. The increase of the public debt in this period to 8,404,000,000,000 francs has to be seen, therefore, in the context of the continuing inflation. The issuance in 1960 of a new franc equalling 100 old francs automatically reduced the nominal value of the public debt to 1 percent of its previous figure.

Great Britain. Government borrowing in Great Britain dates back to the end of the 17th century. In 1692, legislation pledged the receipts from beer and liquor taxes as security for a loan of £1,000,000. As Table 1 indicates, the trend of the debt was upward throughout the next 150 years largely because of wars; by 1802 it had reached £523,000,000 and by 1840, £827,000,000. The second half of the 19th century saw gradual reduction of the debt to £610,000,000 in 1900, while the amount of debt still remaining became less significant because of the growth of the economy in the same period. World War I brought a tremendous increase, the 1920 figure being £7,828,000,000. The 1920s showed little reduction and the figure rose slightly during the Depression years. World War II brought the level to £21,366,000,000 in 1945, and the figure rose in the postwar period—partly as a result of nationalization of industry—to over £34,000,000,000 in the late 1960s.

Table 1: Growth of National Debt
(in 000,000 of currency units shown)

year*	Great Britain and Northern Ireland† (pound sterling)	U.S.‡ (dollar)	France§ (franc♂)	Germany‖ (Deutsche Mark)	Japan¶ (yen)	Canada♀ (dollar)
1697	14	—	—	—	—	—
1757	77	—	—	—	—	—
1781	187	—	—	—	—	—
1791	—	75	—	—	—	—
1802	523	81	—	—	—	—
1815	834	100	1,272	—	—	—
1820	800□	91	3,590	—	—	—
1825	—	84	4,123	—	—	—
1830	—	49	4,890	—	—	—
1835	832◇	▲	4,557	—	—	—
1840	827	4	4,682	—	—	—
1845	818	16	5,810	—	—	—
1850	804	63	5,426	—	—	—
1855	789	36	6,965	—	—	—
1860	799	65	10,262	—	—	—
1865	790	2,678	13.865	—	—	—
1870	768	2,436	12,310	487	5	116
1875	742	2,156	21,185	120	56	152
1880	730	2,091	21,597	388	247	195
1885	703	1,579	24,943	587	240	265
1890	678	1,122	26,152	1,242	255	286
1895	627	1,097	27,258	2,205	299	315
1900	610	1,263	30,080	2,421	506	346
1905	756	1,132	30,610	3,327	1,292	378
1910	720	1,147	32,750	5,014	2,605	471
1915	1,104	1,191	39,023	16,955	2,506	700
1920	7,828	24,299	240,242	184,864	3,278	3,042
1925	7,597	20,516	418,075	2,413	4,901	2,818
1930	7,469	16,185	480,173	10,375	6,003	2,545
1935	7,687	28,701	324,013	14,253	9,613	3,206
1940	9,083	42,968	708,715	52,060	23,481	4,027
1945	21,366	258,682	1,831,859	323,615+	150,795	15,713
1950	25,802	257,357	4,133,000	6,672	316,800	16,751
1955	26,934	274,374	5,867,600	20.131	857,400	17,951
1960	27,733	286,331	8,404,000	25,634	542,000	20,986
1965	30,441	317,274	84,947	33,600	471,000	26,564
1966	31,341	319,907	80,047	—	746,000	27,482
1967	32,001	327,300	90,630	45,308	1,476,000	30,340
1968	34,194	347,660	97,360	49,445	2,213,000	32,924
1969	33,983	352,900	—	—	2,735,000	35,852
1970	—	370,094	—	—	—	38,150

*Fiscal years vary between countries and over the periods covered for the same countries. †Data for 1697–1828 are from E. L. Hargreaves, *The National Debt*, p. 291 (London 1930). Data for 1836–1935 are net national debt compiled from British official "Accounts and Papers" by subtracting exchequer balances from total debt. Data for 1940 and 1945 are from the United Nations, Department of Economic Affairs, *Public Debt, 1914–1946* (1948). Data for 1950–69 are from the United Kingdom Central Statistical Office, *Annual Abstract of Statistics*. ‡Data for 1791–1850 are from U.S. Treasury Department, *Annual Report of the Secretary of the Treasury on the State of the Finances for Fiscal Year Ended June 30, 1903*, p. 63. Data for 1855–1935 are from *ibid, June 30, 1943*, pp. 562–563. Data for later years are from subsequent *Annual Reports* and *Treasury Bulletins*. §Data are from the French Institut National de la Statistique et des Études Économiques, *Annuaire Statistique de la France*. ‖Data for 1870–1935 are from the *Statistisches Jahrbuch fiir das Deutsche Reich;* for 1940, Bank for International Settlements, *Thirteenth Annual Report* (May 1944); for 1944 United Nations, Department of Economic Affairs, *Public Debt 1914–1946*. Data for 1950–68 are for the Federal Republic of Germany and are from the German Statistisches Bundesamt, *Statistisches Jahrbuch fiir die Bundesrepublik Deutschland*. ¶Data for 1870–1945 are computed from Japanese government sources, especially the Department of Finances, *Financial and Economic Annual of Japan*. Data for 1950–64 are from the United Nations, *Statistical Yearbook*. ♀Data are from the Dominion Bureau of Statistics, *The Canada Year Book*, and Canadian Tax Foundation, *The National Finances*, annual. ♂After 1965, data reported are in new francs, equal to 100 old francs. □1828. ◇1836. ▲The total gross debt of the U.S. on Jan. 1, 1835, was $33,733.05. +Sept. 30, 1944.

United States. In the United States, when the federal government was formed, it assumed the debts of the states and various other obligations incurred during the American Revolution, all of which were funded into a single debt issue of $75,000,000 in 1790. The government was highly successful in avoiding additional borrowing in the early years, except for the War of 1812, and during 1835 all federal debt was eliminated (see Table 1). The years 1835 and 1836 were the only ones in the history of the country during which there was no federal debt at all. The American Civil War, only 25 percent of which was financed by taxation, pushed the debt to a total of $2,678,000,000 in 1865. Most of this debt was retired by budget surpluses during the following decades; debt reduction proceeded so far that bonds available for security behind national bank notes became inadequate. The debt remained relatively constant in the 1890s and during the early 1900s. World War I brought an increase to $26,000,000,000, consisting in part of short-term and intermediate-term securities and in part of Liberty Loan bonds. In the 1920s the government was able to reduce the debt; the low point reached was $16,185,000,000 in 1930, primarily by budget surpluses. Interest costs were materially reduced through replacement of old issues by new ones at lower interest rates.

The 1930s brought budget deficits because of the Depression and the efforts to stimulate recovery. Despite extensive borrowing, which raised the total debt to $42,968,000,000 by 1940, interest rates fell sharply as a result of the surplus of money capital and federal reserve action. A substantial part of the borrowing was on a short-term basis, partly because the interest on such loans was extremely low. With the outbreak of World War II, borrowing rose sharply and by 1946 the debt had reached $269,000,000,000.

In the postwar period the debt fell to a low of $252,000,000,000 in 1948, then gradually rose to $371,000,000,000 in December, 1970. This increase was caused by budget deficits arising primarily from a high level of defense spending and the unwillingness of Congress to hold taxes to rates high enough to meet the expense and in some years from a desire to stimulate economic activity.

The states incurred substantial debts in the early part of the 19th century, largely for public improvements, and some found themselves in financial difficulties. As a result, borrowing came nearly to an end until after 1900; after that date there was further borrowing, particularly for highways. After 1945 the state debt increased sharply and reached a total of $39,500,000,000 by 1969. Much of this additional borrowing was for highway purposes. The local governments have traditionally borrowed more than the states, largely because of the nature of their functions. Local debt in the 20th century increased steadily and had reached $94,000,000,000 by 1969. Table 2 shows the growth of federal, state, and local debt in the United States.

Canada. Canada's debt began with $75,000,000 at the time of confederation in 1867, when certain obligations were taken over from the provinces. The figure grew slowly until 1915, largely because of government railroad financing. World War I pushed the figure to $3,042,000,000 by 1920; the total rose as the Canadian National Railways was developed, fell slightly in the late 1920s, rose to $5,000,000,000 with Depression borrowing, and reached $15,713,000,000 at the end of World War II. Some debt fluctuation then took place and the figure reached about $17,000,000,000 by 1950. By April

Table 2: Government Debt in the United States, Selected Years, 1902–70

fiscal year	federal*	state	local	total	total ($) per capita
	(in $000,000,000)				
1902	1.2	.2	1.9	2.3	42
1913	1.2	.4	4.0	5.6	58
1922	23.0	1.1	9.0	33.1	300
1932	19.5	2.8	16.4	38.7	310
1934	27.1	3.3	15.7	46.0	364
1936	33.8	3.4	16.1	53.3	416
1938	37.2	3.3	16.1	56.6	436
1940	43.0	3.6	16.7	63.3	479
1942	72.4	3.3	16.4	92.1	683
1944	201.0	2.8	14.7	218.5	1,579
1946	269.4	2.4	13.6	285.3	2,018
1948	252.3	3.7	15.0	270.9	1,848
1950	257.4	5.3	18.8	281.5	1,856
1952	259.1	6.9	23.2	289.2	1,842
1954	271.3	9.6	29.3	310.2	1,910
1956	272.8	12.9	36.0	321.6	1,923
1958	276.3	15.4	42.8	334.5	1,931
1959	284.7	16.9	47.2	356.3	1,976
1960	286.3	18.5	51.4	364.0	1,979
1961	289.0	20.0	55.0	379.0	1,968
1962	298.2	22.0	58.8	379.4	2,042
1963	305.9	23.2	61.9	390.9	2,073
1964	311.7	25.0	67.2	403.9	2,111
1965	317.3	27.0	72.5	416.8	2,150
1966	319.9	29.6	77.5	427.0	2,180
1967	326.2	32.5	81.2	439.9	2,223
1968	347.6	35.7	85.5	468.7	2,345
1969	353.7	39.5	94.0	487.3	2,413
1970	370.1	—	—	—	—
1971†	396.0	—	—	—	—

*Public debt of the federal government; excludes guaranteed obligations issued by the Federal Housing Administration and nonguaranteed debt of federal agencies. †Estimate.
Source: U.S. Department of Commerce, Bureau of *the* Census. ***Governmental Finances in the United States, 1902 to 1957,*** and annual report, ***Governmental Finances; Treasury Bulletin.***

1969 it had risen to $35,800,000,000 as a result of deficits. The path of provincial and local borrowing in Canada was similar to that in the United States, though with a slower rate of growth.

Germany. The German Reich, founded in 1871, began as a confederation of sovereign states. Most financial powers remained with the individual states until the Weimar Republic was established in 1919. A French war indemnity of 1871 was used largely to reduce the public debts of the states. As late as 1913 the debt of the Reich (4,900,000,000 marks) was less than half that of Prussia (9,900,000,000 marks) and substantially less than the aggregate debt of all the other federal states (6,300,000,000 marks). The country's defeat in World War I led to financial chaos. In 1925, after the stabilization of the new Reichsmark, the public debt was 2,413,000,000 marks. In the 1930s the public debt rose, going to 52,060,000,000 marks by 1940. The Second World War was financed mainly by borrowing, from both the private sector and the central bank; by 1945 the debt stood at more than 300,000,000,000 marks. Most of this was wiped out by the postwar currency reform of 1948.

Japan. The rise of the modern Japanese state began in the latter part of the 19th century. The government began to issue bonds in 1870. The cost of financing the war with China in 1894–95 and a subsequent buildup of its army and navy raised Japan's public debt from 255,000,000 yen in 1890 to 506,000,000 in 1900. The war with Russia in 1904–05 cost about 1,500,000,000 yen, which was mainly raised by foreign borrowing. The financial burden of growing empire was henceforth largely covered by taxation, so that public debt did not increase substantially from 1907 until the end of World War I. Between 1918 and 1930, however, the debt doubled. In these years a large proportion of the debt was in foreign-owned bonds. In the 1930s the government adopted heavy spending policies, mainly for military purposes, and in 1940 the debt was more than three times what it had been in 1930. Since World War II the debt has risen from 150,795,000,000 yen to over 750,000,000,000.

Local governments in Japan have always been heavy borrowers. This has continued to be true in the postwar years, when prefectures, cities, towns, and villages issued bonds on a scale approaching that of the national debt. Much of the local indebtedness was used to finance large public works programs.

Comparative size and burden of government debt

Debt and **national** income. The absolute figures of growth in government debt exaggerate the actual growth in the debt relative to the economy as a whole. In the first place, the general price level has increased significantly over recent decades; since debt obligations are stated in fixed monetary terms, the relative magnitude goes down as the price level goes up. The general rise in prices over a period thus reduces the problems created by the debt for the government and the magnitude of the adverse effects of the interest payments on the economy. The gain occurs at the expense of the bondholders whose real economic position is worsened by the change.

Secondly, the rise in national income reflecting an increase in output reduces the real significance of a fixed sum of debt for the economy. The combined effects of the real and monetary influences can be illustrated by expressing the size of the debt as a ratio to gross national product over a period of years. Data for the United States are shown in Table 3. The ratio fell from 129 percent in 1946 to 38 percent in 1969. The ratio of interest payments to national income likewise fell, despite a rising interest rate level over this period. In Great Britain the ratio of national debt to gross national product fell from 221 percent in 1952 to 136 percent in 1958, although the size of the debt increased slightly over the period.

Magnitude of debt in various countries. An adequate comparison of debt burdens in various countries is difficult to make. The reported figures are by no means en-

Table 3: Ratios of Interest-Bearing National Debt to Gross National Product and Interest Payments to National Income, United States, 1929–69

calendar year	Gross National Product	interest-bearing debt at midyear	ratio of debt to Gross National Product (percent)	national income	interest paid by federal govern-ment*	ratio of interest to national income (percent)
	(in $000,000,000)			(in $000,000)		
1929	103.1	16.6	16	86,795	678	0.8
1930	90.4	15.9	18	75,382	659	0.9
1931	75.8	16.5	22	59,669	612	1.0
1932	58.0	19.2	33	42,785	599	1.4
1933	55.6	22.2	40	40,312	689	1.7
1934	65.1	26.5	41	49,515	757	1.5
1935	72.2	27.6	38	57,208	821	1.4
1936	82.5	33.0	40	65,013	749	1.2
1937	90.4	35.8	40	73,650	866	1.2
1938	84.7	36.6	43	67,372	926	1.4
1939	90.5	39.9	44	72,564	941	1.3
1940	99.7	42.4	43	81,124	1,041	1.3
1941	124.5	48.4	39	104,222	1,111	1.1
1942	157.9	72.0	46	137,065	1,260	0.9
1943	191.6	135.4	71	170,322	1,808	1.1
1944	210.1	199.5	95	182,592	2,609	1.4
1945	211.9	256.4	121	181,485	3,617	2.0
1946	208.5	268.1	129	181,879	4,722	2.6
1947	231.3	255.1	110	199,018	4,958	2.5
1948	257.6	250.1	97	224,178	5,211	2.3
1949	256.5	250.8	98	217,494	5,339	2.5
1950	284.8	255.2	90	241,074	5,750	2.4
1951	328.4	252.9	77	277,978	5,613	2.0
1952	345.5	256.9	74	291,380	5,859	2.0
1953	364.6	263.9	72	304,734	6,504	2.1
1954	364.8	268.9	74	303,138	6,383	2.1
1955	398.0	271.7	68	331,018	6,370	1.9
1956	419.2	269.9	64	350,799	6,787	1.9
1957	441.1	268.5	61	366,096	7,244	2.0
1958	447.3	274.7	61	367,762	7,607	2.1
1959	483.7	281.8	58	400,025	7,593	1.9
1960	503.7	283.2	56	414,522	9,180	2.2
1961	520.1	285.7	55	427,341	8,957	2.1
1962	560.3	294.4	53	457,687	9,120	2.0
1963	590.5	302.0	51	481,927	9,895	2.1
1964	631.7	307.4	49	517,281	10,666	2.1
1965	681.2	313.1	46	559,020	11,346	2.0
1966	749.9	315.4	42	620,600	12,014	1.9
1967	793.5	322.3	41	654,000	12,953	2.0
1968	865.7	344.4	40	714,400	15,404	2.2
1969	931.4	351.7	38	769,500	17,087	2.2

'Fiscal year.
Source: U.S. Department of Commerce, ***Survey of Current Business;*** U.S. Treasury Department, ***Annual Report of the Secretary of the Treasury.***

tirely comparable because they vary in their treatment of debt incurred for various commercial enterprises, loans from foreign countries, special issues, and the like. The relative importance of the national debt and the debt of subordinate units of government also varies, and figures for the latter are not available for many countries. Any comparison of absolute figures of debt in monetary terms is of limited value and may be very misleading because of problems of conversion to a common monetary unit. The only meaningful figure is the ratio of national debt to national income, and the significance of these figures is greatly lessened by the inaccuracy of national income data for many countries. Figures of national debt and of the ratios of debt to Gross National Product are shown in Table 4.

Table 4: Gross National Debt of Various Countries

country	currency unit	year	Gross National Product at market prices (in 000,000s of own currency unit)	Gross National Debt (in 000,000s of own currency unit)	Gross National Debt as a percentage of GNP
Africa					
Ghana	cedi	1968	1,790	1,021	57
Kenya	pound	1967	433	119	27
Rhodesia	pound	1968	392	267	68
South Africa	rand	1968	9,607	3,218	33
Zambia	kwacha	1967	889	230	26
North America					
Canada	dollar	1969	71,454	35,852	50
Costa Rica	colón	1967	4,588	1,236	27
Guatemala	quetzal	1967	1,445	162	11
Honduras	lempira	1968	1,192	145	12
Mexico	peso	1966	276,000	28,253	10
Nicaragua	córdoba	1968	4,900	111	2
Panama	balboa	1966	711	155	22
United States	dollar	1969	932,100	362,100	39
South America					
Argentina	peso	1967	5,197,000	444,258	9
Brazil	cruzeiro	1967	44,396,000	7,711,000	17
Chile	escudo	1966	32,467	3,721	11
Colombia	peso	1968	83,500	12,877	16
Ecuador	sucre	1967	25,000	5,192	21
Peru	sol	1965	114,900	8,630	8
Venezuela	bolivar	1969	41,900	3,713	9
Asia					
Sri Lanka (Ceylon)	rupee	1968	8,800	6,045	69
India	rupee	1969	231,000	161,912	70
Iran	rial	1966	443,000	73,875	17
Iraq	dinar	1968	872	170	19
Japan	yen	1967	41,670,000	1,476,000	4
Philippines	peso	1969	27,000	5,547	21
Thailand	baht	1968	105,600	15,000	14
Europe					
Austria	schilling	1968	276,000	39,875	14
Belgium	franc	1967	970,000	528,916	55
Denmark	krone	1967	84,400	3,491	4
Finland	markka	1967	30,100	4,431	15
France	franc	1967	571,000	90,630	16
Germany (Fed. Rep.)	Deutsche Mark	1967	485,000	80,734	17
Ireland	pound	1967	1,103	792	72
Italy	lira	1968	41,601,000	6,972,000	17
The Netherlands	guilder	1968	81,600	28,591	35
Norway	krone	1967	60,300	14,066	23
Portugal	escudo	1966	132,200	21,931	17
Spain	peseta	1967	1,622,000	209,622	13
Sweden	krona	1968	123,600	25,034	20
Switzerland	franc	1967	67,800	4,918	7
United Kingdom	pound	1968	39,100	68,803	176
Oceania					
Australia	dollar	1968	24,200	3,600	15
New Zealand	dollar	1967	3,452	2,412	70

Source: *United Nations Statistical Yearbook,* 1968, supplemented by information from annual statistical yearbooks of various countries.

DEBT RETIREMENT POLICY

The question whether a government should seek to pay off its debt must be considered separately for the national government and for the state and local governments.

National debt. Retirement of the national debt has the obvious advantage of eliminating the undesirable features noted in an earlier section: the adverse effects of the taxes necessary to pay interest; the redistribution of income resulting from interest payments; the inflationary effects of the debt under certain circumstances; and the confidence-disturbing effects, which are of particular importance in a depression. Furthermore, an overall program calling for a budget surplus accompanied by debt retirement aids in checking inflationary pressures. While it is true that the accumulation of a budget surplus without retirement of debt would exercise still more anti-inflationary pressure, such a policy is not usually regarded as politically feasible. Furthermore, the accumulation of surpluses without debt retirement tends to bring increases in government expenditures; it is politically difficult for a government to maintain an accumulated surplus for any length of time.

Methods and consequences of national debt retirement

Debt retirement is also likely to increase the rate of capital formation in the country; that is, investment in business expansion will increase as long as full employment of resources is maintained. The taxes necessary to finance debt retirement will tend to reduce consumption to some extent, while the sums paid out to the bondholders will be made available, in large measure, as money capital for business expansion. The net effect is to reduce the percentage of total national product used for consumption and increase the percentage used for expansion of the economy.

Retirement of national debt is not always a necessarily good policy. Maintaining a rigid program of debt retirement regardless of business conditions could easily produce a decline in business activity, which in turn would make further repayment virtually impossible for the duration of the slump. If economic stability is regarded as an important goal, then debt retirement is possible only in years of inflationary pressures. Thus, few national governments attempt to retire their debt systematically, year by year. Another objection to debt retirement is that strong popular resistance to tax increases beyond certain levels could result in a situation in which the debt retirement program is carried on at the expense of urgently needed current activities of government, rather than through higher taxes. Certain other arguments can be advanced against retirement. Most national debt is incurred for the conduct of wars; since the effects of these wars extend indefinitely into the future, it may be argued that no one generation of taxpayers should be made to shoulder the burden of debt retirement. Closely related is the argument that continued expansion in the economy will gradually reduce the significance of the debt, so that failure to retire it is no cause for alarm.

If debt retirement is carried on as part of an anti-inflationary program, the net effectiveness of the program will be influenced by the nature of the debt being retired. Maximum effect is obtained by retiring debt held by the central banking system since the repayment will create no inflationary influence at all and, by reducing reserves of the banks, will tend to check bank lending. Retiring debt held by the commercial banks will produce the next best results; although the banks' reserve positions will be improved by the retirement, they will not extend new loans to the extent of the debt retired because other investments will be less liquid than government bonds and the banks will feel obligated to hold greater reserves.

An alternative to debt retirement at the national level is monetization of the debt, that is, payment of the debt by new monetary issues. In a period of depression, business activity would be stimulated by such an act, and the burdens created by the debt would be lessened. Such a policy is rarely favoured for the same reasons that it is not favoured as a means of raising revenue in lieu of borrowing in the first place. If there were assurance that governments would not overissue money and that the policy of monetization would not cause a general loss in confidence, this alternative would be more acceptable.

Finally, national debt may be retired by direct repudiation. Since a sovereign government cannot be sued without its permission, it can, if it wishes, cancel outstanding debt without legal interference. Such a policy has usually been followed only after a government has been overthrown by revolution and the new government has disavowed the debt of the old. Under ordinary circumstances such a policy cannot be seriously contemplated because of the injustice to the bondholders that would result, the loss of prestige by the government, and the difficulty of selling bonds in the future.

The necessity of retiring local government debt

State and local debt. The question of debt retirement is a different one for units of government below the national level. Their debts are incurred primarily for the financing of capital improvements with a limited life, and if they are not retired during the period of use of the improvements, the burden will rest upon generations of taxpayers who will not enjoy the benefits of them. Furthermore, failure to retire debt will soon get a state or local government into a position in which it will be unable to borrow for new improvements because the total amount of debt will exceed the capacity of the government, at tolerable tax rates, to meet interest payments. These levels of government are not faced with the same degree of responsibility for maintaining economic stability as is the national government.

In view of these factors, state bond issues in the U.S. normally provide for a systematic program of retirement, and state law generally requires that the local governments take such action. The length of life of the bonds is limited to the expected life of the improvements. There are two systems for retirement. The older, the sinking fund system, requires that during each year of the life of the issue the government make payments into a sinking fund so that the interest on the accumulating sums, plus the original payments, will be sufficient when the bond issue matures to pay it off in full. Unfortunately, sinking funds are sometimes used for other purposes, and scheduled payments are not always made into them. As a consequence, governments have tended to shift to the serial bond system, under which the maturity dates of serial bonds are spread over a period of years, so that a certain amount will fall due each year, and the issue will thus be redeemed gradually, without the establishment of a sinking fund. Though investors were at first inclined to resist this approach, by the 1960s it had become more or less standard in many countries.

The states of the United States have the power to repudiate debts if they wish, since they have inherent sovereign powers. On some occasions, notably in the South after the ousting of the carpetbagger governments following the American Civil War, states have repudiated debts outright. Local governments, however, being technically corporations, have no sovereign powers and thus can be sued and cannot repudiate debt.

Distinct from the question of outright repudiation is that of unintended default. In the 1830s and again in the 1930s, several states of the U.S. and provinces of Canada and large numbers of municipalities were unable to meet their obligations at tax rates that were politically acceptable. As far as the states and the Canadian provinces were concerned, bondholders could not take legal action and simply had to wait until tax collections improved. At the municipal level, legal action was possible, as provided by state law; in some instances bondholders could obtain a court order requiring the levying of additional taxes. In other instances this was either impossible or futile, given the depressed conditions of the period. In four states provision was made for a state agency to assume administrative receivership of municipalities in default to straighten out their finances. In 1934 the U.S. Congress enacted a municipal bankruptcy act that provided for a revision of local government debts under court supervision and with the concurrence of various percentages of the debt holders. The original law was held unconstitutional, but a new law enacted in 1936 was upheld. Government bondholders can never foreclose on government-owned property in the event of default and normally cannot proceed against the property of taxpayers or municipal officials, except to a limited extent in the New England states. It is interesting that in Great Britain in the 1930s, where the Depression caused acute distress, no local authority defaulted.

DEBT MANAGEMENT

Government debt takes various forms, and the combination of them is the province of debt management. There are two sets of problems: new borrowing and the conversion or refunding of existing securities into new issues.

Alternative methods of financing government debt

Policies on new borrowing. Governments face a number of alternatives when undertaking new borrowing. National governments of large, developed countries seldom consider borrowing from foreign sources, since they are able to obtain desired funds within the country. For developing countries, however, the issue is a significant one. Frequently potential sources of domestic funds are limited. Savings are small and not easy to mobilize through financial institutions. Foreign borrowing also offers the important advantage of access to additional foreign exchange, thus permitting the importation of goods important for economic development. But such loans create obligations to pay interest and principal to foreign debt holders in the future; these payments may be difficult to make if the country's export potential is limited. Foreign loans often have restrictions attached that governments are reluctant to accept and may bring foreign intervention. Accordingly the governments of developing economies differ in the degree of their willingness to seek foreign funds. The development of international lending agencies such as the International Bank for Reconstruction and Development (World Bank) has increased the availability of foreign loans to developing countries without the danger of foreign interference.

The general rule for subordinate governments is that of adjusting the borrowing in such a manner as to minimize interest and other casts. National governments must strike a balance between low interest costs and attainment of the goals of economic stability. The precise policies that must be followed in the light of this objective will depend on the relative importance attached to the interest and economic stability goals, the relative effectiveness of monetary policy and other methods of attaining economic stability, and the state of business conditions. Policies appropriate for depressions are clearly different from those suitable for inflationary periods. While the principles themselves are widely accepted, there are serious problems of interpretation in specific cases.

Voluntary vs. compulsory loans. One question of concern in time of war is the choice between voluntary and compulsory loans. Canada, Great Britain, and several other countries used a compulsory lending program to some extent during World War II; the United States considered it but relied on the voluntary approach. The compulsory system is undoubtedly more anti-inflationary, per dollar obtained, since a greater amount of the money will come from persons who would otherwise spend it in consumption. But it is much less acceptable politically.

Compulsory lending offers several advantages over tax increases. Because the sums will eventually be paid back, political opposition to the program and the possible adverse effects on incentives will be less than they would be with tax increases. The program may be more equitable than with a tax increase for the lower income groups because the overall final burden may be distributed to a greater extent on the higher income levels than would be possible with the tax increase. The return of the funds may cushion a postwar depression. However, in neither Great Britain nor Canada was the World War II experience very satisfactory, because the taxpayers regarded the payments essentially as another tax rather than as savings that they would ultimately get back, and therefore the hoped-for advantages were lost.

Compulsory lending must be distinguished from a true compulsory savings program. The former method requires persons to place a certain amount of money (usually related to income) in government bonds each year. The latter method requires persons to save—*i.e.*, reduce their consumption—a certain amount during the year and place it in government bonds. A compulsory savings plan has not been attempted because of problems of enforcement. The liquidation of existing assets to offset the new saving and thus defeat the intent of the program would be very difficult to detect. But if the program could be made effective it would be much more anti-inflationary than even a compulsory lending system.

In a period in which borrowing is necessitated by war or high defense spending, the government may appeal to the public to buy bonds on patriotic grounds. During and after World War II, the U.S. government followed this program extensively, particularly to get the bonds into the hands of individuals rather than banks and thus to minimize inflationary pressures. In a severe emergency such a program may have merit, but it is difficult to maintain the "crisis" atmosphere indefinitely. If, however, persons can be induced to commit themselves to buy bonds on a payroll deduction basis, the forces of inertia are placed on the side of recurrent bond purchases, and the net effect may be to increase materially the total purchase of bonds by the lower income groups and thus to reduce consumer spending.

To a considerable degree, the national government can influence the extent to which the bonds are acquired by different types of purchasers. It may "borrow" directly from the central bank, if it wishes and the law permits, and thus create money. It may borrow from the commercial banks by direct sales of bonds to them, and it can, in fact, exert strong pressure on them to buy; it may sell to individuals, even restricting bank purchases of the issues. The choice of lender depends in large measure on the state of economic activity. In a period of depression, the goal is to obtain funds with the least possible contracting effect on individual and business spending. Thus, if money creation is ruled out, sale to commercial banks, which usually have excess reserves in depressions, is almost as good but gives rise to greater interest cost. In a period of full employment and inflationary tendencies, the goal is the reverse—to borrow in such a manner as to restrict private spending to the maximum extent. Thus during World War II, governments stressed their savings bond programs.

Short-term versus long-term borrowing. Subordinate units of government are usually unable to borrow substantial sums on a short-term basis, and they thus confine their borrowing to the issuance of bonds, usually running for 10, 20, or more years. The only exception may involve small amounts of short-term borrowing in anticipation of tax revenue. But national governments have a much greater choice between the two methods. Traditionally, short-term borrowing has been regarded as indicative of financial irresponsibility. Beginning in the 1930s, however, many national governments began to make increased use of this form of credit, primarily because of the much lower rate of interest. In depression periods, a substantial volume of loanable funds exists whose owners are willing to invest them only on a short-term basis, with maximum liquidity and minimum risk.

From the standpoint of economic stability, the primary disadvantage of the short-term loan is the greater degree of liquidity to the investor. In inflationary periods, borrowing of this sort is less effective than long-term borrowing in reducing liquidity. Short-term borrowing also creates a continuous problem of refunding the issues as they mature, increases the danger that large amounts of debt may come due in a period of financial stringency and embarrass the government, and renders the government's interest burden highly vulnerable to changes in the market interest rate. Short-term borrowing is particularly disadvantageous if interest rates are expected to rise in the near future. On the whole, short-term borrowing is most justifiable when interest rates are expected to fall—*e.g.*, at the top of an inflationary boom.

When inflation continues for a period of time, many persons become reluctant to buy any type of fixed-money-return security. Accordingly, it has been suggested that in such periods bonds might be sold with fixed purchasing power; that is, interest and principal payments would be adjusted in terms of changes in the price level so that the holder would be protected from the effects of inflation. At least one city in the United States has issued bonds of this type. But such a policy would not be without difficulties if adopted by a national government. If there were strong fear of inflation, a great shifting from existing securities to the new ones would occur and could disorganize the bond markets. The change would serve as a notice that the government expected inflation to continue, and that in itself would be a spur to increased purchasing, which would stimulate inflation. Payment of interest and principal on the basis of higher prices would feed the inflationary pressures and increase the future financial problems of the government. One more "escalator" would be added to the economy, and inflationary pressures would be increased.

Most government bonds are salable, have a market price, and therefore may be transferred on the market. On the other hand, they are not usually redeemable until maturity. In the United States, Canada, Great Britain, and some other countries, savings bonds were made nonsalable at the outbreak of World War II, primarily to insure the purchasers against any decline in market value such as occurred to similar bonds after World War I. This feature also enabled the government to control the amounts individuals purchased and prevented a mass shift from regular issues to the higher-interest-rate bonds. But since they were nonsalable it was virtually imperative to have them redeemable at will in order to make them attractive to the purchaser. Persons were given an incentive to hold them, however, because the interest rate was higher the longer the bonds were held. Use of the discount method of paying interest avoided the need for making large numbers of very small payments, delayed the receipt of interest until after the major inflationary pressures were expected to be over and facilitated the scheme of varying the rate according to the period the bonds were held in order to discourage persons from cashing them in.

Interest rate policy. Governments can borrow only at the market interest rates as determined by current demand and supply for loanable funds. A government will attempt to ascertain the rate at which a new issue can be sold by examining yields on existing issues and consulting investment banking firms. Normally an effort is made to set a rate at which the bonds will sell at par (maturity value). If the rate is set too low, the bonds will have to be sold at a discount and the effective rate will reflect the market figure; the result of selling at a discount is often a net actual rate slightly higher than otherwise.

Occasionally governments deliberately set rates higher than the market in order to attract certain classes of investors. A primary example of this was the establishment of an unnecessarily high rate on U.S. savings bonds during World War II in order to get as much of the debt as possible into the hands of individuals rather than the banks. Banks were not permitted to buy these issues.

While state and local governments are completely at the mercy of the money market insofar as interest is concerned, a national government can affect the rate of interest through its influence over the central banking system. The national government can thus reduce its interest burden if it wishes or can prevent an increase. If such a policy is followed in an inflationary period, however, it will interfere with effective anti-inflationary measures by the central banking system.

When the central banking system raises interest rates in periods of inflation, state and local borrowing is made more expensive. These governments then seek special assistance from the national government in obtaining funds more cheaply. Such aid was provided after World War II in Great Britain through the Public Works Loan Board, which made loans to local authorities. There was extensive demand for this sort of assistance in Canada after 1955. Similar difficulties arose in the U.S. in 1966–68. The problem illustrates one of the basic difficulties of general credit control, namely, the nonselective restriction on borrowing.

A government can make its bonds more acceptable to buyers, and thus obtain a lower rate of interest, by exempting the interest on its bonds from tax. This was the practice of the U.S. federal government before 1941; after that date only state and local bond interest were exempted from federal tax. Many countries do not provide such an exemption.

For a particular level of government, if tax rates are progressive, the tax loss will exceed the gain from lower interest rates unless there is sufficient market among persons in the highest tax bracket to absorb all of the bonds. In this instance the interest rate savings would more or less offset the tax loss. But this is unlikely to be the situation, and in all higher tax brackets than the one at which the two elements are in balance, the government will lose in tax more than it gains through lower interest, and persons in lower brackets will not, presumably, buy the bonds. The shift in federal policy in the United States in 1941 was therefore desirable.

But when the units of government involved are at different levels, as in the case of the exemption of state and local bond interest from federal tax in the United States, the units whose bonds are exempt experience a net gain from lower interest and thus will oppose any effort by the other government to make the interest taxable. The United States originally exempted interest on state and local bonds because the 1894 income tax law had been held invalid in part because the Supreme Court considered the inclusion of state and local bond interest to violate the constitutional division of powers between the federal and state governments. Since the attitude of the Supreme Court on this general question has changed materially (federal taxation of the salaries of state employees having been held valid in 1939), it is now thought likely that the court would accept the taxation of interest on state and local bonds. But the strong opposition of the states and local governments has prevented action by Congress, although it is obvious that these governments gain less than the federal government loses. Furthermore, the exemption provides a simple avenue of tax avoidance and encourages wealthy persons to put their money into these safe investments instead of into business expansion. The 1969 tax reform legislation did not alter the tax status of state and local bonds, although 1968 legislation did restrict somewhat the tax-exempt status of bonds issued to provide facilities for new industry.

Adjusting the structure of debt

Management of existing debt. Debt management also involves adjustment of the existing debt: replacement of short-term securities by long-term securities, refunding of maturing issues, transfer of securities from one type of investor to another (banks to individuals, for example), conversion of issues, and the like. This is a problem rarely faced by local and state governments because their debt is retired on a systematic basis and there is little opportunity to change the form before the date of maturity. About the only possibility of change is that of repayment before maturity if surplus funds are available or if replacement of an old issue by a new (as lower interest rates) is feasible. To facilitate such action without the necessity of buying the securities in the market, the issues are sometimes made callable so they can be paid off at maturity value (or with some premium) at any time beyond a certain year.

Maturity structure. **A** national government can, if it wishes, materially change the relative size of the short-term and long-term elements in the debt. Short- and intermediate-term issues, which are constantly maturing in large volume, can be replaced by long-term bonds, or long-term bonds can be replaced by short-term issues, either as they mature or through purchase or call.

Replacement of short-term by long-term issues, advocated by many persons in the period after 1945 and attempted by the U.S. government on several occasions during that period, offers several advantages in periods of inflationary tendencies. First, it reduces the overall liquidity of the debt structure, making it more difficult for investors to use wealth in government securities for spending on consumption or business expansion. Second, increased use of long-term securities lessens the nuisance and cost of constant refunding operations and makes less frequent the danger that extensive refunding may coincide with a stringency in the money market, thus necessitating central bank action to support the government's borrowing. This action may run directly counter to the interests of economic stability. Lengthening the maturity structure of government debt also tends to raise long-term interest rates relative to those on short-term loans. This should lessen inflationary pressures. But the actual magnitude of the effect is doubtful; a deflationary monetary policy pursued by the central banking system may be far more effectiveand simpler.

The primary disadvantage of increased reliance on long-term securities is the higher interest rate that must usually be paid, especially if the change occurs in periods of high rates. On rare occasions short-term rates may exceed long-term rates, but normally the reverse is true. It is this consideration that discouraged the U.S. government from continuing its program of lengthening the maturity structure of the debt in the late 1950s. Part of the difficulty lay in the reduced competitive position of long-term government bonds compared with other securities. The rapid fluctuations in government bonds, as monetary policy was used more intensively, made them less attractive investments, while private securities became more attractive while business activity remained at a high level. The increased supply of government-guaranteed mortgages also lessened interest in government bonds.

In terms of cost considerations alone, the appropriate time for a shift toward long-term issues is in a depression. While short-term funds become very cheap in such periods, it is more advantageous for the government to take advantage of the low figures on long-term issues. On the other hand, short-term borrowing keeps the lenders' investments highly liquid and facilitates their spending the sums for consumption or business expansion whenever they wish. Thus cost and stability considerations suggest opposite policies. The former requires shifting from short-term to long-term loans in depressions, while the latter dictates this shift in periods of inflation. But the liquidity considerations are probably not very important in depression periods, nor is the gain from the shift toward long-term securities in inflationary periods. The important consideration is that the overall length of issues be relatively long when inflation does arise. There is an increased trend toward a policy of relying on other policies than changes in debt maturity as a means of lessening economic instability and of lengthening debt maturities in periods when interest rates are low.

Structure of holdings. The government may at times find it advantageous to shift the relative holdings of different types of investors. In periods of inflationary pressures, increased holdings of savings bonds by the public and reduced holdings of other bonds by commercial banks would lessen spending in the economy and reduce the liquidity of individuals — the ability to increase spending sharply if they wish. It is for this reason that during the inflationary years of the postwar period the U.S. government continued to push the savings bond program even when it was not doing any net borrowing. Table 5 shows the changing ownership of the U.S. national debt between 1940 and 1970.

On the other hand, during a period of depression, the repayment of savings bonds through borrowing from the commercial banks would tend to increase total spending. Some of this effect will be attained automatically as persons cash in their savings bonds when they lose their jobs.

Government securities are held, apart from those in government trust funds and the central banking system, by three principal groups of investors — commercial banks, other financial institutions, and individual investors. In the United States, some tailoring of securities to the particular needs of various investors has been undertaken through the use of savings bonds, which have been restricted to individuals. Some authorities have suggested that much more of this should be done, in the interests of stability and reduced cost to the government. In general such a program could greatly reduce the volume of refunding and free the central bank from constant concern with this problem to the neglect of credit control policy. The direct effect of debt management on stabilization could be increased through varying redemption and volumes of issues for various classes of investors.

Distinct from the question of adjustment in the structure of the debt is that of conversion designed simply to lower interest cost, without change in the nature of the debt involved. When the interest rate has fallen sharply, conversion may be highly advantageous. It is not possible, however, unless the bonds are callable, except in the rare

Table 5: Estimated Ownership **of** United States Federal Securities
(par values—in $000,000,000)*

year (as of December 31)	total federal securities outstanding†	held by Federal Reserve banks	held by commercial banks‡	U.S. government investment accounts	held by private nonbank investors						
					total	individuals§	insurance companies	mutual savings banks	corporations‖	state and local governments	miscellaneous investors7
1940	50.9	2.2	17.3	7.6	23.9	10.6	6.9	3.2	2.0	.5	.7
1945	278.7	24.3	90.8	27.0	136.6	64.1	24.0	10.7	22.2	6.5	9.0
1950	256.7	20.8	61.8	39.2	134.9	66.3	18.7	10.9	19.7	8.8	10.5
1955	280.8	24.8	62.0	51.7	142.3	64.7	14.6	8.5	23.2	15.4	15.6
1960	290.4	27.4	62.1	55.1	145.8	64.7	11.9	6.3	18.7	18.7	24.2
1962	304.0	30.8	67.2	55.6	150.4	64.5	11.5	6.1	18.6	20.1	28.0
1963	310.1	33.6	64.3	58.0	154.3	66.2	11.3	5.8	18.7	21.1	29.3
1964	318.7	37.0	63.7	60.6	157.4	68.2	11.1	5.7	17.9	21.2	31.2
1965	321.4	40.8	60.8	61.9	158.0	72.3	10.4	5.4	15.5	22.9	31.4
1966	329.8	44.3	57.5	68.8	159.3	75.5	9.6	4.7	14.7	23.8	30.9
1967	344.7	46.7	63.8	71.6	158.7	74.0	8.6	4.1	12.2	25.1	34.7
1968	372.0	52.2	65.5	76.1	163.0	75.3	8.0	3.6	14.6	27.1	30.8
1969	368.0	54.1	56.5	84.8	165.6	79.3	7.1	2.9	15.8	27.1	33.4

'United States savings bonds, Series A–F and J, are included at redemption value. †Securities issued or guaranteed by U.S. government, excluding guaranteed securities held by the Treasury. ‡Consists of commercial banks, trust companies, and stock savings banks in U.S. and territories. $Includes partnerships and personal trust accounts. Nonprofit institutions and corporate pension funds included under "miscellaneous investors." ‖Exclusive of banks and insurance companies. ¶Consists of savings and loan associations, nonprofit institutions, corporate pension trust funds, dealers and brokers, and Investments of foreign balances and international accounts.
Source: U.S. Department of the Treasury, Treasury ***Bulletin.***

instance of forced conversion. This latter type of policy is in a sense partial repudiation and is not followed by democratic governments. If the bonds involved are callable and conversion is decided upon, the government will give the bondholders notice of conversion and give them the choice of repayment in cash if they wish.

Debt management in the usual sense does not include central bank policy designed to influence the general interest rate level, which is regarded as a distinct form of stabilization policy. The two are, however, closely related. Central bank policy controls the interest rate that the government must pay; the interest costs of the national debt can always be reduced by central bank action that increases the supply of money capital through credit creation. But such a policy is not likely to be consonant with the goal of economic stabilization.

British **debt policy.** In recent years particular attention has been given to debt management by the government of Great Britain. Its debt policy in the 1960–70 decade was dominated by three influences: the theoretical and practical implications of the work of the Radcliffe Committee, the world trend in interest rates, and the problems posed by the balance of payments.

The Radcliffe Committee

The Report of the Committee on the Working of the Monetary System (1959) questioned the traditional Bank of England view that the demand for long-term government debt was perverse so that higher interest rates lowered the demand for gilt-edged securities instead of increasing it. The reasoning turned on the expectation of investors in the bond market. If interest rates rise and bond prices fall, investors may expect further declines in the bond market and delay purchases or even sell bonds, thus precipitating a further decline. The implication of this traditional view was that the monetary authorities, if they wished to sell long-term debt, had to maintain an orderly bond market; this, in turn, meant supporting prices if they were falling fast and dampening them if they were rising quickly.

The Radcliffe Committee challenged this policy, pointing out that permanent damage could be done to the market's appetite for government bonds if the authorities were continually dragging themselves up behind inexorably rising rates. Despite the Report, this happened in the 1960s when the bank rate rose from 4 percent to 8 percent. Over the decade the average bank rate was *6* percent and the average yield on 2½ percent non-redeemable bonds was 6% percent. Within this ten-year period there were three cycles. The bank rate rose to 7 percent by July 1961 and then fell to 4 percent by January 1963; it rose again to 7 percent by July 1966 and was reduced to 5% percent by August 1967; finally, with the balance of payments crisis in late 1967, it was pushed to 8 percent and by April 1970 was back to 7 percent.

During the period when rates were rising to 8 percent, the authorities were consistent sellers of bonds. At the same time they were obliged to convert gilt-edged securities maturing at a rate of about £1,500 million a year. The difficulties of the authorities as net sellers of debt in a period of rising rates forced them to rely on methods other than open-market operations to control credit. These included requests made by the central bank to the clearing banks to limit advances or even to reduce them and the requirement that the banks maintain special deposits at the Bank of England in addition to their cash reserves. By 1970 the authorities found themselves pushed toward a more active open-market policy.

As the Radcliffe Committee had suggested, serious consequences occurred in the bond market with a large reverse yield gap (an investment in bonds yielding more than it would in stocks) and the authorities often in the position of supporting a falling bond market. Late in 1968 bond prices were allowed to fall sharply. This appeared to follow the Radcliffe prescription.

Another part of the Radcliffe legacy was what has become known as the "new-orthodoxy" controversy. This gave short-term government treasury bills a central part to play in influencing the volume of banks' liquid assets, which in turn set a ceiling to the expansion of bank deposits, The central bank by selling Treasury Bills can reduce banks' cash and thereby their deposits and advances.

From 1959 to 1969 the Treasury Bill portfolio of the clearing banks and the money market fell from 22 percent of gross deposits to only 12 percent; but this fall was partially offset by a rise in commercial bills (through export credits) unexpected by the Radcliffe Report. Also Treasury Bills were replaced for the nonbank public by short-term local authority claims.

This debate over short-term debt policy was conducted against the background of a shortening of the average maturity of the debt and a substantial increase in official holdings (*i.e.*, funds held by government agencies, the Bank of England, and the Exchange Equalization Account). In 1959 about 50 percent of the total U.K. debt was held in bonds maturing in over fifteen years, and 17.5 percent in bonds maturing in under five years. By 1969 the long-term debt had dropped to **46** percent, medium debt had been further squeezed, and short-term debt had risen to 32 percent of the total. Short-term floating debt had remained at 17 percent of the total debt. Official holdings increased both as a proportion of total debt held and in their commitment to short-term bonds. Between 1959 and 1969 official holdings rose from 25 percent of marketable securities to 29 percent; but official holdings of short-term debt rose from 12 percent of total official holdings to 32 percent.

Finally, the persistent deficits on the current balance of trade and the accumulated international indebtness of the United Kingdom over the decade also affected debt policy. In 1967, after the pound was devalued, the authorities implied that they meant to exercise greater control over the money supply. This new emphasis on restricting the money supply was accompanied by a substantial effort at debt repayment.

BIBLIOGRAPHY

Economics of government borrowing: J.M. BUCHANAN, ***Public Principles of Public Debt*** (1958), a thorough statement of conflicting points of view about the shifting of debt burden to future generations; ***The Public Finances,*** 3rd ed., pt. 6 (1970), a survey of public debt theory; C.S. SHOUP, ***Public Finance,*** ch. 18 (1969), a survey of government borrowing and debt theory; J.M. FERGUSON (ed.), ***Public Debt and Future Generations*** (1964), a collection of papers on the theory of the shifting of debt burden to the future; and H.C. ADAMS, ***Public Debt: Essays in the Science of Finance*** (1887), the classic work on public debt.

Debt management: T.C. GAINES, ***Techniques of Treasury Debt Management*** (1962), a survey of the objectives and techniques of debt management, including a history of policies; W.L. SMITH, ***Debt Management in the United States*** (1960), the best available analysis of debt management policy; E. NEVIN, ***The Problem of the National Debt*** (1954), an analysis of various alternative policies, with emphasis on Great Britain; and H.C. MURPHY, ***The National Debt in War and Transition*** (1950), a detailed discussion of World War II experience.

State and local debt: B.U. RATCHFORD, ***American State Debts*** (1941), the principal study, now out-of-date; and A.J. HEINS, ***Constitutional Restriction Against State Debt*** (1963), the history of state borrowing in the U.S. and the effects of debt limitation laws.

(J.F.D.)

Public Enterprises

Public enterprises

The term "public enterprise" denotes an organization operating (or supposed to be operating) on commercial principles, wholly or partly owned and effectively controlled by a public authority. An enterprise of this kind may be a new creation or owe its existence to the nationalization of a privately owned concern. It may have as its main function the provision of some "infrastructural" service (*e.g.*, power or transport), the direct manufacture of a commodity, or the extension of certain forms of assistance (*e.g.*, credit, marketing) to enterprises in the private sector. As a commercial entity engaged in the sale of its goods and services to private or corporate consumers, it normally needs to be organized and controlled in ways that are different from those applicable to ordinary, noncommercial government agencies. In particular it requires a certain measure of operational freedom and an immunity from persistent governmental intervention in its current decision-taking processes.

CHARACTERISTICS OF PUBLIC ENTERPRISES

There are many possible reasons for the creation of a public enterprise. Perhaps the most ancient of motives is the fiscal, exemplified by the establishment of "state monopolies" in commodities such as tobacco, salt, gunpowder, and alcohol. In certain countries such enterprises are still important revenue earners, although an equivalent yield to the public purse might be obtained by the more simple expedient of imposing excise duties upon these commodities. Another reason for locating an enterprise in the public sector is the strategic. For instance, military considerations strongly influenced Bismarck's nationalization of the German railway system, and in many countries today the government has decided that prudence demands that it should assume direct responsibility for certain types of defense production (*e.g.*, arms and ammunition). The existence of a monopoly, or the desirability of creating one, offers another justification for the public enterprise. If, for social reasons, it is not considered appropriate to break up a monopoly so as to subject the producers to the automatic regulator of free competition. there are the alternatives of subjecting it to state regulation or bringing it under state ownership. The latter is frequently chosen because it seems to be simpler and less productive of tension. In some cases public enterprise has even been used to create a monopoly where none previously existed by bringing together independent units into one large undertaking in order to achieve economies of scale, as with the nationalized coal and electricity industries in Britain. Cultural and educational factors may also be responsible for the choice of public as distinct from private enterprises. Such examples are the creation of the British Broadcasting Corporation and the Swedish wine and spirits enterprise known as the AB Vin und Spritcentralen.

Public enterprises have traditionally been regarded as acceptable alternatives to private enterprises in the field of "utilities" such as transport and power. In some countries, however, of which the United States provides the best-known example, even utilities are uredominantlv run bv private companies operating under strict legal regulation. By contrast, in Communist countries public enterprises embrace not only utilities but almost the whole range of industrial production, commerce, and finance. Here there prevails an anticapitalist ideology that strongly favours the public enterprise as an essential constituent of a socialist pattern of society. Ideological predispositions toward public enterprise are also found in varying degrees of strength outside the Communist countries and particularly in certain developing countries such as Egypt, Syria, Iraq, Algeria, Guinea, Sudan, Tanzania, Burma, and—more doubtfully—India. Socialist parties in the non-Communist developed countries, although frequently committed by their- founding charters to a massive extension of vublic enterprise through nationalization., nowadays tend to adopt a much more pragmatic attitude toward the public-private balance, particularly when they occupy governmental office.

The mixed economy

Indeed, except where Communist governments are in power, the principle of the "mixed economy" universally prevails. Although the constituents of the mixture vary widely from country to country, it usually contains a much larger public enterprise element than would previously have been regarded as either necessary or tolerable. Developing countries, especially, have often found it necessary to use public enterprises on a large scale as a result partly of their urgent need for infrastructural investment—which in these days can rarely be undertaken by private agencies—and partly of the unwillingness or incapacity of their private entrepreneurs and investors to initiate and bring to fruition projects that are regarded as essential constituents of national development programs. For reasons of this kind Pakistan, although predisposed to favour private enterprise, has built up a relatively larger public enterprise sector than is to be found in India, a country which, at least in theory, has adopted the "socialist pattern."

The actual extent of the public enterprise sector, however, is not always easy to define, whether one measures it by its contribution to the gross national product (GNP), its absorption of investment resources, or its generation of productive employment; for the sector includes not only public utilities and other directly productive undertakings but also (1) state-controlled agencies that provide financial and other services for the benefit of private industrial and agricultural undertakings and (2) state-controlled agencies charged with the task of entering into active collaboration with the private entrepreneur and investor in the development of certain areas of the economy. In most countries with mixed economies there has been an increasing interpenetration between public and private enterprise. The most obvious form that this takes is the creation of "mixed" companies whose shares are partly state-owned and partly privately owned. The state shareholdings in such companies frequently are managed by large "holding" corporations, such as the Istituto per la Ricostruzione Industriale in Italy, the Industrial Development Corporations in Pakistan, or the Corporación de Fomento de la Producción in Chile, which are themselves public enterprises. Another form of interpenetration is illustrated by the Gezira Corporation in Sudan, through which individual peasant cultivation is supervised, serviced, and partly financed by a government agency. Indeed, there are some countries where the public-private links provided by financing institutions, development corporations, mixed companies, and organized cultivation schemes are so close that the question "What is a public enterprise?" becomes difficult to answer. A United Nations survey in 1969, using figures provided by a limited number of mixed-economy countries, estimated that the amount of employment generated by public enterprises varied from 1.3 percent in Ceylon (now Sri Lanka) to 7 percent in France, its share of gross fixed investment from 14.6 percent in Greece to 39 percent in Ghana, and its contribution to GNP from 2.5 percent in Ceylon to 12.3 percent in Italy. Although these percentages should not be treated too seriously, they may be taken as representing rough orders of magnitude.

The role played by public enterprise in the context of a national economy is as variable as its size, but there is no necessary correspondence between these two criteria. As has been indicated, it would be a mistake to regard the existence of a large public enterprise sector as indicative of socialistic proclivities. Conversely, a small one provides no evidence of devotion to "economic freedom" in its traditionai sense. Sweden, for instance, although widely regarded as a successful example of a social-democratic state, has comparatively few public enterprises, whereas Italy, which has never been ruled by a socialist government, has many. Nor does a large public enterprise sector invariably mean that the governments responsible for its creation have adopted or intend to adopt a system of centralized, "command-type" planning. Although the operation of such a system is obviously facilitated by having most productive undertakings in state ownership, planning via the price and credit mechanisms remains a viable alternative, as the Yugoslav example has shown.

It should also be noted that, particularly in developing countries, the balance between public and private enterprise has been a shifting one. Both policies and perspectives have undergone change as a result not only of movements in political opinion but of the accumulation of experience. Both Turkey and Pakistan, for instance, found themselves compelled by circumstances to give public enterprise a much larger role in their economies than was originally envisaged for it. Mexico, on the other hand, became less enthusiastic about public enterprise as its revolutionary period receded and political conditions were established that opened up opportunities to its relatively numerous would-be private entrepreneurs. What seems to be well established, however, is that in the modern world even the less socialistically inclined of the developing countries find that, at a certain stage in their progress toward economic maturity, they need to make fairly extensive use of public enterprise as a catalytic agent.

THE GROWTH OF PUBLIC ENTERPRISE SECTORS

Nationalization

Of the developed European countries operating mixed economies, Britain, France, and Italy provide the most interesting examples of the evolution of reasonably successful and well-integrated public enterprise sectors. In both Britain and France there were extensive nationalizations during the period immediately following World War II, but neither country lacked previous experience of running public enterprises. During the interwar period, Britain had nationalized broadcasting, the long-distance transmission of electricity, London passenger transport, and civil airways. These enterprises, together with the Post Office, the Royal Ordnance Factories and Admiralty Dockyards, the Port of London, and the numerous electricity, gas, and transport enterprises owned by local authorities, constituted a small but significant public enterprise sector of the economy. In France, successive governments had established the National Nitrates Office, acquired a majority of shares in the Compagnie Générale Transatlantique and the railways, nationalized the aircraft manufacturing companies, and become a majority or minority participant in the equity of a number of other undertakings. Neither country, therefore, was entirely unprepared administratively for the massive nationalizations that followed the end of World War II, which were placed on the statute books by governments of a "leftist" kind—the Labour Government of 1946–50 in Britain and the short-lived Socialist-Communist-MRP coalition in France. Both countries nationalized coal, gas, electricity, and civil airlines, to which Britain added railways (already nationalized in France), long-distance road haulage, and steel. France, however, undertook a much more extensive nationalization than Britain of financial enterprises. While Britain brought the Bank of England under public ownership, France nationalized not only the Bank of France but the four great deposit banks, 34 insurance companies, and a variety of other finance houses. In addition, the state took over some important manufacturing concerns forfeited by their private owners for wartime collaboration with the Germans. The most famous of these was the Renault Motor Works. Since then there have been no further major nationalizations in either country. Britain, in fact, witnessed the denationalization of both steel and road haulage by a subsequent Conservative government. The steel industry, however, was renationalized by the Labour government that came into office in 1964. It may be noted that, to promote the deveiopment of the most important of all the new industries, which called for heavy investment with a comparatively distant and—at the time—uncertain yield, both countries established atomic energy commissions, with commercial as well as research and development responsibilities.

Reconstruction and development

Italy has had a very different experience. Whereas the extension of public enterprise in Britain and France took place, for the most part, through a series of deliberate legislative acts—in Italy it occurred largely as a result of an expansion, sometimes gradual, sometimes rapid, but never very well coordinated—of state shareholdings. Until the 1930s the Italian state had undertaken very few commercial functions, although the railways had been nationalized in 1905 and corporations subsequently created to take over the administration of the postal, telephone, and telegraph services; the state monopolies; and the state forests. In 1923 the state monopoly of life insurance, created in 1912, was disbanded before it had even begun effective operations. In 1933, however, the Fascist government was compelled by the severity of the impact of the world economic crisis on Italy's financial and industrial structure to mount a massive rescue operation by which the three main commercial banks, with their considerable holdings in industry, agriculture, and real estate, were taken over and a holding company, the Istituto per la Ricostruzione Industriale (IRI), created to administer the newly-acquired state shareholdings. The original idea was that the state should disinvest as soon as the economic situation sufficiently improved; but the IRI was retained, and during the postwar period it expanded its shareholdings to become Italy's greatest commercial enterprise, employing through its totally owned and partly owned subsidiaries a work force of some 300,000. It became one of a small group of comparable holding corporations, all working to what has become known as the "IRI formula," which evoked considerable interest and some imitation abroad. The distinctive features of the Italian public enterprise sector are that it enjoys a high degree of entrepreneurial autonomy and that its interpenetration with the private sector of the economy is close.

Among the developed countries, the U.S., the most developed of all, makes least use of public enterprise. Although the emergency requirements of the two world wars and the Great Depression of the 1930s produced a considerable proliferation of public enterprises, mainly for purposes of finance and procurement, most of these were subsequently disbanded. In this respect the policies pursued by the U.S. differed considerably from those of its neighbour, Canada, which not only retained in the public sector one of the most important of the commercial agencies created during World War II, Polymer Corporation, but subsequently created a number of new ones such as the Canadian Commercial Corporation (1946), the Canadian Sugar Stabilisation Corporation (1947), the Northern Transportation Company (1944), the North-West Territories Power Commission (1948), the Canadian Overseas Telecommunications Corporation (1949), and Atomic Energy of Canada, Ltd. (1952). Nevertheless, the U.S. has one of the most famous public enterprises in the world, the Tennessee Valley Authority (TVA) which, founded in 1933, has been paid the compliment of widespread imitation by countries wishing to undertake the integrated development of their great river valleys for purposes of irrigation, flood control, navigation, soil conservation, and the generation of hydroelectric power.

In general, the less developed countries have tended to copy the patterns of public enterprise already established in the more developed, although some of them have undertaken a pioneering role. Japan was the first of them systematically to use public enterprise for promotional purposes, while the U.S.S.R. pioneered its use for the building up of a socialist type of economy. Contemporaneously with Italy, Turkey experimented vigorously with the "holding company" type of enterprise through the creation of the Sümerbank and Eti Bank (which still occupy a very prominent place in its industrial structure) in the 1930s, while in Africa Egypt created a unique system by which a multitude of public enterprises were brought under the supervisory authority of 39 "Public Organizations."

Except in those developing countries where nationalization policies have been adopted for ideological or political reasons, the main role of the state has been to create new public sector enterprises for the expansion and diversification of the economy, rather than to take over existing private sector enterprises. Some countries, such as India, have attempted to define in advance which branches of their economies shall be reserved, for purposes of expansion and development, respectively to the public and private sectors.

FORMS OF PUBLIC ENTERPRISE

The decision to establish a public enterprise invariably raises questions of legal status and organizational structure. Before World War I there was a widespread assumption, among socialists and nonsocialists alike, that a public enterprise could be fitted into the "normal" departmental structure of the executive government. Post offices, in fact, provided the prototype. The justification for this assumption was that in those days governments rarely entered into the more competitive or technologically experimental types of business where entrepreneurial abilities were required of management. The postal, communications, and transport enterprises that, in most countries, constituted the bulk of the government's business responsibilities were monopolies or quasi-monopolies engaged in supplying standardized services at fixed prices. Broadly speaking, the same was true of the local enterprises that were owned and controlled by municipalities. Hence there seemed little advantage in the adoption of more "autonomous" forms of organization, which tended to create constitutional complications by breaking the normal chain of command that culminated in the absolute responsibility of the departmental minister.

These conditions have disappeared; today the demand is for a specific form of organization that enables those who are actually managing the enterprise to enjoy, in a measure compatible with ultimate public control, the innovatory freedom and discretionary authority normally possessed by the private businessman. Even where the "departmental" form of organization is retained—as it may well be where the "social service" responsibilities of an enterprise rival its commercial responsibilities in importance—it is invariably modified in various ways; *e.g.*, to confer exemption from normal budgetary and appropriation procedures and to provide greater flexibility in matters of personnel management. Where it is discarded—*e.g.*, for much of a government's business responsibilities—recourse is had to specialized organizational types, all more or less inspired by the joint-stock company.

The public corporation

One of the major institutional innovations is the public corporation, which is the form given to most of the British nationalized industries. Owned by the state, it is normally created by a special law defining its powers, prescribing its form of management, and specifying its relationships with superior governmental authorities. As a body corporate, it can hold property and sue and be sued in its own name. Although financed by treasury appropriations, treasury loans, or treasury-approved fixed-interest stock, it meets its current costs from the sale of its goods and services, makes normal commercial provision for depreciation and reserves, and may be authorized to reinvest its profits. Its budget is separate from the state budget; it is exempt from the normal regulations applicable to the expenditure of public funds; and its accounts are subject to a commercial-type audit, whether by public authority or by especially appointed private accountants. Its employees are not normally civil servants and are recruited, remunerated, promoted, and disciplined by the corporation itself, subject to whatever general legal regulations may be applied. Governmental powers over the public corporation are usually limited to those assigned by its constituting statute. These, however, may be widely or narrowly defined, and in any case the relevant minister, being responsible for the nomination of the corporation's top-level management (which is usually a board), can sometimes exercise an informal influence over its decisions that considerably exceeds the scope of his formal powers.

The distinction between the public corporation and the government department, however, may be much less precise than the above account suggests, since there are various degrees to which an enterprise, whatever it may be called, can be integrated with the normal machinery of government. For instance, the French form of public corporation, the ***entreprise publique***, is subject to considerably more detailed supervision and control than its British counterpart. Moreover, there are examples of enterprises that cannot be clearly categorized as either "departmental" or "corporate." Of these the best known are the Swedish trading agencies which, although constitutionally responsible to the king-in-council and staffed by civil servants, enjoy a degree of financial autonomy denied to other governmental agencies and are authorized to recruit, on a "temporary" basis, personnel to whom the ordinary civil service regulations do not apply.

The company

An alternative to the public corporation is the state company. When this is used, the law relating to ordinary joint-stock companies (sometimes embodied in a commercial code) is applied to the enterprise, and public control is ensured by the government's exercise of shareholding rights. Originally the state company device was developed in the European countries, but it is now widely employed throughout the world. The public corporation form is frequently used for utilities, while the state company form is favoured when the government enters the field of manufacturing industry. This tends to be the case, for instance, in India. It is often claimed that the "company" confers on the public enterprise concerned a greater degree of commercial flexibility. This may be so, but examples are not infrequent of the imposition on state companies of rigid patterns of organization and control through special laws or articles of association. The creation of a company may be effected by executive decree; often, however, it owes its existence to a development corporation, which is authorized to found subsidiaries organized on the joint-stock principle. When the intention of the government is that the development corporation shall eventually dispose of its subsidiaries to the private investor, their establishment in the form of companies has obvious conveniences.

The company form is even more clearly indicated when a government wishes to establish a "mixed" enterprise, through which it aims to enter into partnership with private investors. Such enterprises are found in almost all countries that have not decided to adopt fully socialized patterns of economic development. Sometimes they are regarded as transitional, insofar as the government intends to retain its participation only until such time as private investors prove willing to take up the whole of the share capital; but in other cases the public-private partnership is considered as a means of giving the government permanent influence or control over the enterprise's management. It has become conventional to consider **a** mixed enterprise as part of the public sector when the government owns 51 percent or more of the share capital, and as part of the private sector when it owns less than this. This distinction, however, is often unreal since a government that owns only a small number of shares in a company can usually control it if the remainder are widely scattered among members of the investing public. Moreover, it is by no means unknown for the government-appointed directors of a mixed enterprise, whether they be in a majority or in a minority, to be given powers superior to those possessed by the privately appointed directors. A true partnership is normally achieved only when a government makes special contractual arrangements with a major private undertaking, either indigenous or foreign. Such an arrangement sometimes provides for the management of the enterprise, either permanently or temporarily, by personnel nominated by the private partner. Otherwise, the main advantage of the mixed enterprise is that it enables the mobilization of private capital, which might not be available to an enterprise that did not enjoy the advantage of government support. Another circumstance that may justify its use is the need to rescue from bankruptcy, by the injection of government resources, a private enterprise whose continued existence is regarded as essential to the national economic welfare.

PROBLEMS AND ISSUES

Although the form given to a public enterprise is of undoubted importance, it has tended to receive excessive discussion at the expense of more vital issues; for there is rarely any close correspondence between form and performance, which depends on the efficiency of internal organization and procedures, the quality of personnel at all levels, the adequacy of the incentives that they are given, and the capacity of the management and willingness of the political authorities to ensure that, de facto as well as de jure, the enterprise actually enjoys the degree of commercial flexibility it is supposed to have.

Efficiency versus politics

The promotion of efficiency in public enterprises raises certain special problems. There is a widespread tendency to subject the enterprise to meticulous political control, incompatible with the development of managerial initiative and self-confidence, and to impose on it inappropriately bureaucratic forms of organization, usually adopted from the noncommercial branches of the public service. It is clear that, whatever legal powers over the enterprise the minister may possess, those powers need to be exercised with restraint and to be concentrated on the broader issues of policy such as capital requirements and long-term development plans. Only thus can the agency be simultaneously "clothed with the power of government" and "possessed of the initiative and flexibility of private enterprise," to quote the famous words used by Pres. Franklin Delano Roosevelt when inaugurating the Tennessee Valley Authority. It is also clear that a public enterprise needs to be equipped with a management at least as well trained and experienced (and hence as well remunerated) as the management of a private enterprise of comparable size and complexity and should be given the opportunity to install the most up-to-date managerial techniques. In the more developed countries marked progress has been made toward the achievement of both these desiderata; but in the less developed ones there is usually much room for improvement.

Efficiency will also partly depend on the manner in which an enterprise is fitted into the country's general administrative structure. Experience suggests that the best pattern is one that brings each enterprise under the jurisdiction of the "relevant" minister (*i.e.*, the minister of industry for industrial enterprises, the minister of agriculture for agricultural enterprises, the minister of finance for financial enterprises, etc.). In some cases, however, there may be strong countervailing arguments for placing most of the public enterprise sector under a minister of state enterprises. A special problem may arise when a large variety of individual enterprises have been created as subsidiaries by a development corporation, functioning partly as a holding company and partly as a superior managerial agency. The normal experience is that when one of these enterprises has become fully established and has reached a certain size, it needs to be detached from its parent body and brought under the direct supervision of the minister whithin whose area of jurisdiction it naturally lies. Conversely, there may be the problem of bringing together in a single organization a number of separate but related enterprises in order to simplify the process of control and to achieve economies of scale. The circumstances in which this is appropriate always require detailed investigation by the political and administrative authorities in collaboration with the enterprise managements likely to be affected. Some patterns of organization may create resistances to unification proposals of undoubted economic and administrative rationality. Yugoslavia, which has confided the management of its enterprises to workers' councils operating in association with local or regional government authorities, has experienced difficulties of this kind. In general, the importance of patterning the public enterprise system in terms of what Paul Appleby called "coherent missions" can hardly be overemphasized.

The assessment of performance

A further problem relates to the criteria by which the performance of a public enterprise is to be assessed. When it is operating in a fully competitive market situation, profitability obviously provides a reasonably adequate measurement, but when there is an element of monopoly, more sophisticated criteria based on cost-benefit calculations have to be brought to bear. The task of assessment may be further complicated by the imposition of special burdens on a public enterprise or the extension to it of special privileges. On the one hand, it may be denied the freedom to charge a fully economic price for its goods and services or may be instructed by superior authority to engage in operations of an allegedly socially beneficial kind that involve it in commercial loss; on the other hand, it may receive various forms of protection denied to its competitors or potential competitors or enjoy the use of its capital resources on especially favourable terms. From the standpoint of facilitating public control over the patterns of production and distribution, all this has great advantages, but its disadvantages are no less obvious. It is wide open to abuse; it may lower the morale or undermine the incentives of both managers and workers; and it immensely complicates the task of producing a reliable and operationally useful assessment of public enterprise performance. For these reasons, and particularly the last of them, some countries have established expert, independent agencies to develop tools of performance measurement, to keep public enterprises under more or less continuous review, and to advise both enterprises and government of the likely consequences of the pursuit of alternative policies. Of such agencies the most famous and most successful is the Commission de Vérification des Comptes in France. It should be noted, however, that no agency of this kind can be of much use unless the "targets" that a public enterprise is expected to achieve have been specified as clearly as possible by the responsible authorities. Of recent years, in both Britain and France, such "targeting" has made notable advances. In Britain, for instance, each enterprise is now given an objective, stated in financial terms, which takes into account its social responsibilities as well as its expected commercial performance. This exercise not only makes the process of assessment more rational; it also provides a series of parameters within which management may be left relatively free to go about its business.

Another "control" device that has achieved some success is the appointment of a special parliamentary committee to provide the legislature with information and advice on matters relating to public enterprise. The expectation is that this will enable parliament to exercise its supervisory functions with greater intelligence and discretion and restrain it from inciting the relevant minister to use his powers in harmful ways. The best-known examples are the Select Committee on Nationalised Industries in England and the Committee on Public Undertakings in India. Despite these advances, however, it would be an exaggeration to suggest that any country has as yet succeeded in finding the ideal balance between autonomy and control, the search for which raises all the main problems in this field of governmental activity.

BIBLIOGRAPHY. *Report of the United Nations Seminar on Organization and Administration of Public Enterprises* (1967), a document emphasizing major principles; W.G. FRIEDMAN (ed.), *The Public Corporation: A Comparative Symposiunz* (1954), a collection of country studies with a comparative essay by Friedman; R. TURVEY (ed.), *Public Enterprise: Selected Readings* (1968), a discussion of the economic problems of public enterprise, with special reference to British experience; A.H. HANSON, *Public Enterprise and Economic Development,* 2nd ed. (1965), a study of the use made of public enterprise in the development programs of the less developed countries, with special reference to India, Turkey, and Mexico; W.A. ROBSON, *Nationalized Industry and Public Ownership,* 2nd ed. (1962), the standard work on British nationalized industries; M. EINAUDI, M. BYE, and E. ROSSI, *Nationalization in France and Italy* (1955), the most useful work in English on the public enterprises of both countries; M.V. POSNER and S . ~WOOLF, *Italian Public Enterprise* (1967), an analysis of economic problems, with main concentration on the IRI; D.V. VERNEY, *Public Enterprise in Sweden* (1959), the only published work on the subject in English, short but comprehensive; H. SEIDMAN, "The Government Corporation in the United States" in *Public Administration,* vol. 37, pp. 103–114 (1959), a succinct description and analysis by an acknowledged authority; J.R. MOORE (ed.), *The Economic Impact of TVA* (1967), the most recent work on the TVA, containing references to previous works and a chapter by J. OLIVER, "The Application of TVA Experience to Underdeveloped Countries"; L.D. MUSOLF, *Public Ownership and Accountability: The Canadian Experience* (1959), an examination of the relationship of the public enterprises with the executive and legislature; OM PRAKASH, *Theory and Working of State Corporations* (1962), L. NARAIM, *Public Enterprises In India* (1968), two comprehensive works concerned primarily with India.

(A.H.H.)

Public Health Services

Public health has been defined as the art and science of preventing disease, prolonging life, and promoting physical and mental efficiency through organized community effort. Man is a social being, and humans characteristically associate with each other for their mutual protection and advantage. From the interactions involved in dealing with the many problems of social life, there has emerged a recognition of the importance of community action in the promotion of health and the prevention and treatment of disease. This recognition is expressed in the concept of public health.

The approach to and understanding of public health has changed from century to century and, more recently, from year to year. Insights into the nature of public health can be gained by examining these changes.

In primitive communities, society's defense against sickness was to isolate or destroy the sufferer. As knowledge of sources and modes of infection increased, governments attempted to control implicated environmental factors, such as public water supplies, milk and other foods, human waste, insects, and various forms of pollution. In the 20th century there has been an increasing interest in social legislation for the welfare of mothers and children and of the aged or physically handicapped; attention has been given to housing and special diseases, such as cancer, venereal disease, heart disease, and mental illness. Public health emphasis accordingly has changed from a large impersonal approach through environmental controls to a more personal approach through preventive medicine. In a number of countries there has been recognition of the inadequacies of personal medical care and increased emphasis on more accessible, efficient, and effective forms of medical care. The full range of health services—promotive, preventive, curative, and rehabilitative—is now seen to be inseparable, and the emphasis is on developing methods of integrating these sectors into comprehensive health-service systems aimed at both individuals and communities.

The range of health services

One way of defining the scope of public health is by consideration of the threats to the health of members of a community and the steps needed to protect man from those threats and to increase the likelihood that he will live a full and healthy life.

Diseases and hazards to health vary from region to region, and different countries handle their health problems in different ways. The greatest differences are between the less developed and the more developed nations.

This article is concerned with the historical development of public health, beginning in ancient times and emphasizing how various public health concepts have evolved. It deals with organizational and administrative methods of handling these problems in both the more developed and the less developed parts of the world. Special attention is given to the developing countries and to how the health problems, limitations of resources, education of health personnel, and other factors must be taken into account in designing health-service systems. Finally, there are descriptions of the most recent developments in public health, together with some indications of the problems still to be solved.

General history of public health

BEFORE THE INDUSTRIAL REVOLUTION

Beginnings in antiquity. Most of the world's primitive people have practiced cleanliness and personal hygiene, often for religious reasons, and, apparently, a wish to be pure in the eyes of their gods. For thousands of years epidemics were looked upon as divine judgments on the wickedness of mankind; but the idea that pestilence is due to natural causes, such as climate and physical environment, developed later—and gradually. This great advance in thought took place in Greece during the 5th and 4th centuries BC and represented the first attempt at a rational, scientific theory of disease causation. The association between malaria and swamps, for example, was established very early (503–403 BC), even though the reasons for the association were obscure. In the book *Airs, Waters, and Places,* thought to have been written by Hippocrates in the 5th or 4th century BC, the first systematic attempt was made to set forth a causal relationship between man's diseases and his environment. For hundreds of years, until the new sciences of bacteriology and immunology emerged, well into the 19th century, this book provided a theoretical basis for the comprehension of endemic disease (that persisting in a particular locality) and epidemic disease (that affecting a number of people within a relatively short period).

Greek theories of disease causation

The Middle Ages. In terms of disease, the Middle Ages can be regarded as beginning with the plague of 542 and ending with the Black Death (bubonic plague) of 1348. Diseases in epidemic proportions included leprosy, bubonic plague, smallpox, tuberculosis, scabies, erysipelas, anthrax, trachoma, sweating sickness, and dancing mania (see INFECTIOUS DISEASES).

The isolation of persons with communicable diseases first arose in response to the spread of leprosy. This disease, uncommon in antiquity, became a more serious problem in the Middle Ages and particularly in the 13th and 14th centuries.

The Black Death reached the shores of southern Europe from the Middle East in 1348 and in three years swept throughout Europe. The chief method of combatting plague was to isolate known or suspected cases as well as persons who had been in contact with them. The period of isolation at first was about 14 days and gradually was increased to 40 days (see QUARANTINE AND ISOLATION). Stirred by the Black Death, public officials created a system of sanitary control to combat contagious diseases, using observation stations, isolation hospitals, and disinfection procedures.

Major efforts to improve sanitation through these years included the development of pure water supplies, garbage and sewage disposal, and food inspection. These efforts were especially important in the cities, where people, though they lived in crowded conditions, did so in a rural manner with many animals around their homes.

During the Middle Ages a number of beginnings in public health were made: these included attempts to cope with the unsanitary conditions of the cities and, by means of quarantine, to limit the spread of disease; the establishment of hospitals; and provision of medical care and social assistance. These efforts were especially impressive because they were made in a period in which people were powerfully swayed by superstition, and scientific knowledge was inadequate to deal effectively with the public health problems of the day.

The Renaissance. Centuries of technological advance culminated in the 16th and 17th centuries in a number of scientific accomplishments. These included the discovery, by William Harvey, of the circulation of the blood: the growing use of experimental methods; progress in the study of disease in the individual and in populations; and the first consistent explanation—by the Veronese physician Girolamo Fracastoro—of the spread of disease by contagion.

Educated leaders of the time recognized that the political and economic strength of the state required that the population maintain good health. No national health policies were developed in England or on the Continent, however, because the government lacked the knowledge and administrative machinery to carry out such policies. As a result, public health problems continued to be handled on a local community basis, as they had been in medieval times and would be until well into the 19th century.

The great scientific advances of the 16th and 17th centuries laid the foundations of anatomy and physiology as medical sciences. Observation and classification made possible the more precise recognition of diseases. The idea that microscopic organisms might cause communicable diseases had begun to evolve. Such developments were to provide the basis for later change.

DURING AND AFTER THE INDUSTRIAL REVOLUTION

National developments in the 18th and 19th centuries. Nineteenth-century movements to improve sanitation that occurred simultaneously in several European countries were built upon foundations laid in the period between 1750 and 1830. From about 1750 the population of Europe increased rapidly, and with this increase came a heightened awareness of the large numbers of infant deaths and of the unsavoury conditions in prisons and in mental institutions.

This period also witnessed the beginning and the rapid growth of hospitals. Hospitals founded in Britain, as the result of voluntary efforts by private citizens, helped to create a pattern that was to become familiar in public health services. First, a social evil is recognized; then studies or experiments are undertaken through individual initiative, and these efforts mold public opinion and attract governmental attention. Finally, such agitation leads to governmental action.

This era was also characterized by efforts to educate people in health matters. In many countries appeals to reason were based on a belief in progress and the perfectibility of society.

The health of industrial workers

As the Industrial Revolution developed, the health and welfare of the workers deteriorated. When the discrepancy between the harsh social reality and the optimism of the prevailing economic philosophy was brought into focus, the serious need for coping with the problems of public health was also recognized.

In England, where the Industrial Revolution and its bad effects on health were first experienced, there arose in the 19th century a movement toward sanitary reform that finally led to the establishment of public health institutions. Between 1801 and 1841 the population of London doubled; that of Leeds nearly tripled. With such growth there also came rising death rates. Between 1831 and 1844 the death rate per thousand increased in Birmingham from 14.6 to 27.2; in Bristol, from 16.9 to 31; and in Liverpool, from 21 to 34.8. These figures were the result of an increase in the urban population that far exceeded available housing and of the subsequent development of conditions that led to widespread disease and poor health.

The Poor Law Commission, created in 1834, explored problems of community health and suggested means for solving them. Its report, in 1838, argued that "the expenditures necessary to the adoption and maintenance of measures of prevention would ultimately amount to less than the cost of the disease now constantly engendered." Sanitary surveys proved that a relationship exists between communicable disease and filth in the environment, and it was said that safeguarding public health is the province of the engineer rather than of the physician. Filth was declared to be a public enemy that endangers the health of whole communities.

A General Board of Health was established to furnish guidance and aid in sanitary matters to local authorities, whose earlier efforts had been impeded by lack of a central authority to which they could appeal for help and leadership. The board had authority to establish local boards of health and to investigate sanitary conditions in particular districts.

Advances in public health in England had a strong influence in the United States, where one of the basic problems, as in England, was the need to create effective administrative mechanisms for the supervision and regulation of community health. Catastrophe often precedes social change. In America recurrent epidemics of yellow fever, cholera, smallpox, typhoid, and typhus made the need for effective public-health administration a matter of terrifying urgency. The so-called Shattuck report, published in 1850 by the Massachusetts Sanitary Commission, reviewed the serious health problems and grossly unsatisfactory living condition in Boston. Its recommendations included an outline for a sound public-health organization based on a state health department and local boards of health in each town. A change from a haphazard administration to an efficient one became essential to the development of a complicated urban industrial society. In New York City (in 1866) such an administration was created for the first time in the United States.

Developments in Germany and France in the 19th century

Nineteenth-century developments in Germany and France pointed the direction in which future public health action would yo. During the first half of that century France was pre-eminent in the areas of political and social theory, and as a result the public-health movement in France was deeply influenced by a spirit of public reform. The greatest French contribution to the advancement of public health at that time, however, was in the application of scientific methods to the identification, treatment, and control of communicable disease.

Although many public health trends in Germany resembled those of England and France, the absence of a centralized government until after the Franco-Prussian War did cause significant differences. After the end of that war and the formation of the Second Reich, a centralized public-health unit was formed. Another development was the emergence of hygiene as an experimental laboratory science. In 1865 the creation at Munich of the first chair in experimental hygiene signalled the entrance of science into the field of public health just at the time it was also becoming involved in clinical medicine.

During the 19th century, knowledge of the health problems of industry was slowly accumulating, but it was not until the 20th century that substantial advances in this area were achieved where the combined efforts of medical research, administrative action, and social reform were focussed on the problem.

There were other advances. The use of statistical analysis in handling health problems emerged. The forerunner of the United States Public Health Service came into being, in 1878, with the establishment in the United States of port quarantine on a national basis and with assignment of enforcement of the quarantine to the Surgeon General of the Marine Hospital Service. (Port quarantine was the isolation of a ship at port for a limited period to allow time for the manifestation of disease.)

A number of investigations that were pursued during the 19th century suggested that communicable disease is caused by living organisms.

Developments from 1875 to 1950. A growing understanding of infection has resulted in either the virtual eradication or the effective control of many communicable diseases.

The work of an Italian bacteriologist, Agostino Bassi, with silkworm infections early in the 19th century prepared the way for the later demonstration that specific organisms cause a number of diseases. Some questions, however, were still unanswered. These included problems related to variations in transmissibility of organisms and in susceptibility of individuals to disease. Light was thrown on these questions by discoveries of human and animal carriers of infectious diseases.

Nineteenth-century developments in bacteriology

In the last decades of the 19th century the French chemist Louis Pasteur, the Germans Ferdinand Julius Cohn and Robert Koch, and others developed methods for isolating and characterizing bacteria; the English surgeon Joseph Lister developed concepts of antiseptic surgery; the English physician Ronald Ross identified the mosquito as the carrier of malaria; a French epidemiologist, Paul Louis Simond, provided evidence that plague is primarily a disease of rats spread by rat fleas; and two Americans, Walter Reed and James Carroll, demonstrated that yellow fever is caused by a filterable virus carried by mosquitoes. Thus, modern public health and preventive medicine owe much to the early medical entomologists and bacteriologists. A further debt is owed bacteriology because of its offshoot, immunology.

In 1881 Pasteur established the principle of protective vaccines and thus stimulated an interest in the mechanisms of immunity that has persisted to the present. The development of microbiology and immunology had immense consequences for community health. In the 19th century the efforts of health departments to control contagious disease consisted in attempts to improve environmental conditions. As bacteriologists identified the micro-organisms that cause specific diseases and learned how they produce their effects, however, progress was made toward the rational control of specific infectious diseases.

In the United States the diagnostic bacteriological laboratory was developed—a practical application of the theory of bacteriology, which evolved largely in Europe. These laboratories, established in many cities to protect and improve the health of the community, were a practical outgrowth of the study of micro-organisms, just as the establishment of health departments was an outgrowth of an earlier movement toward sanitary reform. And just as the health department was the administrative mechanism for dealing with community health problems, the public health laboratory was the tool for the implementation of the public health program. Evidence of the effectiveness of this new phase of public health may be seen in statistics surrounding immunization against diphtheria—in New York City the mortality rate due to diphtheria fell from 785 per 100,000 in 1894 to 1.1 per 100,000 in 1940.

Poverty and disease

During the first decade of the 20th century, new developments vastly broadened the horizons of public health workers. While improvements in environmental sanitation were of great value in dealing with some problems, they were of only limited usefulness in dealing with many health problems found among the poor of the community. In the slums of England and the United States, for example, malnutrition, venereal disease, alcoholism, and other diseases were widespread. Nineteenth-century economic liberalism held that increased production of goods would eventually bring an end to scarcity, poverty, and suffering. By the turn of the century, it seemed clear to many that deliberate and positive intervention by reform-minded groups, including the state, would also be necessary. For this reason social action was entered into by physicians, clergymen, social workers, other public-spirited citizens, and government officials. Programs that were initiated included organized efforts to prevent tuberculosis, lessen occupational hazards, and improve children's health.

The first half of the 20th century saw further advances in community health care, particularly in the welfare of mothers and children and the health of schoolchildren. This period also saw the emergence of the public health nurse and the development of voluntary health agencies, health education programs, and occupational health programs. Problems of medical care also took on increasing importance around the world. The need to provide medical care for the poor was recognized in England in the 17th and 18th centuries and expressed as early as the Elizabethan Poor Law of 1601. In the second half of the 19th century there were two significant attempts to provide medical care for large populations. One was by Russia, and took the form of a system of medical services in rural districts. After the Communist Revolution, this was expanded to include complete government-supported medical and public-health services for all the people. Similar programs have since been adopted by a number of European and Asian countries.

The other course was prepayment for medical care, a form of social insurance first adopted toward the close of the 19th century in Germany, where prepayment for medical care had long been a familiar concept. A number of other European countries adopted similar insurance programs.

In the United Kingdom, a royal-commission examination of the Poor Law in 1909 led to a proposal for a unified state medical service. This service was the forerunner of the 1946 National Health Service Act, which represented an attempt by a modern industrialized country to provide services to all people, a milestone in the history of community health action.

In the United States, the widespread private-practice, fee-for-service approach to medicine has long provided very good care for some, adequate care for many, and entirely inadequate care for a large segment of the population. Two efforts toward reform have emerged: first, reorganization of medical care through governmental action; and, second, private, prepaid medical care programs.

Recent situation of public health

ORGANIZATIONAL AND ADMINISTRATIVE PATTERNS

International organization. Since ancient times, the spread of epidemic disease from country to country had demonstrated the need for international cooperation for health protection. Early efforts toward international control of disease appeared in the form of national quarantines in Europe and the Middle East. The first formal international health conference, held in Paris in 1851, was followed by a series of similar conferences aimed at drafting international quarantine regulations. A permanent international health organization was established in Paris in 1907 to receive notification of serious communicable diseases from participating nations, to transmit this information to the member nations, and to study and develop sanitary conventions and quarantine regulations on shipping and train travel. This organization was ultimately absorbed by the World Health Organization in 1947.

In the Americas, the organization of international health probably began with a regional health conference in Rio de Janeiro in 1887. From 1889 onward there were several conferences of American states, which led ultimately to the establishment of the Pan-American Sanitary Bureau; this was made a regional office of the World Health Organization in 1949, when it became known as the Pan-American Health Organization.

The rise and decline of health organizations has been influenced by wars and their aftermaths. After World War I, a Health Section of the League of Nations was established and functioned until it was destroyed by World War II. After World War II, the United Nations Relief and Rehabilitation Administration (UNRRA) was set up; one of its important activities was to handle displaced persons in such a way as to prevent the spread of disease. It was responsible for the planning steps that led to the establishment in 1948 of the World Health Organization as a special agency of the United Nations.

WHO definition of health

Health is defined in the constitution of the World Health Organization (WHO) as "a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity."

The work of the organization is carried out under the direction of the World Health Assembly, which is composed of representatives from the member states. The first assembly selected a number of projects for the program of the organization. The first consideration was to be given to diseases that exist in large areas of the world and that lend themselves to international action. On this basis the following fields were given priority: (1) malaria control; (2) tuberculosis control; (3) venereal disease control; (4) the promotion of health by positive means, such as a concentration on measures to improve the health of mothers and children; (5) amelioration of environmental conditions responsible for a significant proportion of deaths throughout the world; (6) the improvement of nutrition, an essential condition to the betterment of health for both children and adults. Since this first selection was made, other areas of need have been included in the work of the World Health Organization.

Among the most important functions of the organization are the advisory services offered to governments through its regional staff. For knowledge in specialized fields, a number of expert committees were created for the advisory services, to consider specific questions and develop reports, which are made available to all member nations.

The World Health Organization maintains close relationships with other United Nations specialized agencies, particularly the United Nations International Children's Emergency Fund (UNICEF) and the Food and Agricultural Organization, and with international labour organizations.

The World Health Organization undertakes both short- and long-term projects to control or eradicate diseases on a world scale. A spectacularly successful short-term project on a local scale was a 1953 campaign against typhus in Afghanistan, when the World Health Organization helped government workers to dust 345,000 persons and 19,000 houses and other facilities with insecticides. The biggest long-term undertaking among WHO field projects was the worldwide campaign against malaria, begun in 1955.

Advanced nations. Forms and functions of health administration vary from country to country. Major health functions are frequently grouped in a department that is held responsible for health and for closely related functions. In the United Kingdom such functions are carried out by the Department of Health and Social Security; in the United States the Department of Health, Education, and Welfare has jurisdiction over the programs covered by national legislation.

Departments of health

Few central departments of health are all-embracing; other departments also operate medical programs of some sort. No country, for example, places the health services of its military forces under the central health agency. Because unity of control at the centre is impracticable, coordination is most important. Central administration is further complicated in federal systems. In the United States, for example, there are 50 states, no two of which have the same patterns of health organization.

Patterns shared. The official responsible for the administration of national health affairs is usually a member of the Cabinet, and his political leadership is of great value in the development of a national health program. Advisory councils are frequently used to bring the thoughts of leading scientists, health experts, and community leaders to bear on national health problems.

An organization that provides basic community health services under the direction of a medical officer is called a local health unit. It is usually governed by a local authority. Its functions may include maternal and child health, communicable-disease control, environmental sanitation, maintenance of records for statistical purposes, health education of the public, public health nursing, medical care, and, often, school health services. The local health unit can provide the administrative framework for a wider range of community health services, including the care of the aged, of the physically handicapped, and of the chronically ill, and mental health services. Although social welfare services may be provided by a separate agency, there are advantages in amalgamating health and welfare services, because a family's health and social problems tend to be intimately interrelated. In England, welfare and public health are often integrated at the local level, whereas in the United States they are almost always separate.

The population served by a local health unit may number only a few thousand or as many as several hundred thousand. There are substantially different problems involved in administering health services for a large rural area that is sparsely populated and a municipality with a population of one or two million.

One problem of administering local health services centres on the question of whether they should be governed by independent local authorities or organized on a regional basis to ensure coordination, effective referral, and lack of duplication of services. This leads to the further question of regional planning, which will be considered later in this section.

Types of public medical service

Medical care is provided as a public service to some degree in most countries. Public medical service may be limited to hospitalization of persons with certain ailments, such as mental disease, tuberculosis, chronic ill. ness, and acute infections. Comprehensive health services may be provided for certain population groups, as in Canada and the United States, where the federal government provides such care for Indians and Eskimos. Many countries have compulsory sickness insurance, and some combine the socialization of hospitals with sickness insurance covering general medical care, as in Denmark. Full-scale socialization of health services exists in a few countries, including the United Kingdom, the Union of Soviet Socialist Republics (U.S.S.R.), and New Zealand.

In countries such as The Netherlands and the United States, in which voluntary and nonprofit organizations support a considerable share of the health services and operate most of the general hospitals, there is bound to be pluralism in health administration. This makes coordination difficult, but strong voluntary effort has the advantages of involving citizens directly in the development of their health services and of infusing the administration with a spirit of experimentation often lacking in governmental services.

Historically, hospitals have tended to develop haphazardly with little relation to one another. There is often a duplication of services to some population groups and an absence of services to others. The growth of modern general hospitals has tended to increase institutional care at the expense of care of patients in clinic facilities and homes. While the quality of inpatient care has increased markedly, there has often been a neglect of health problems that cannot be cared for in hospitals. Community-level services often are underfinanced and function without coordination with hospital programs.

Levels of health care

There is a trend toward regional planning of comprehensive health services for defined populations. In a somewhat idealized plan, the first level of contact between the population and the health system, which can be called primary care, is provided by health personnel who work in community health centres and who reach beyond the health centres into the communities and homes with preventive, promotive, and educational services. At the next level of care, general specialists in community hospitals provide secondary care for patients referred from the primary-care centres. Finally, tertiary, or superspecialty, care is provided by a major medical centre, often a university teaching centre. The various levels of this regional scheme would be linked by a two-way flow of medical records, patients, and health personnel.

Regionalization has been most fully achieved in Europe and least so in North America, where voluntary hospitals provide most of the short-term general services and retain autonomy in their administration. Rising costs of medical care, together with acknowledged inadequacies of the medical care system, have resulted in substantial pressures for reform in the United States and in experimentation on the development of improved systems.

Variations. Among the more developed nations, there is substantial variation in the organization and administration of health services. Three examples may be given here: the United Kingdom, which has a National Health Service with substantial autonomy given to local government for implementation; the Union of Soviet Socialist Republics, in which health services are accepted as the responsibility of the national government and administration is strongly centralized; and the United States, with a pluralistic approach to health services, in which local, state, and national governments have varying areas of responsibility, with the private sector playing a prominent role.

During the first half of the 20th century in Britain, the emphasis shifted gradually from environmental toward personal public health. A succession of statutes, of which the Maternity and Child Welfare Act (1918) was probably the most important, placed responsibility for most of the work on county governments. National health insurance (1911) gave benefits to 16,000,000 workers and marked the beginning of a process upon which the National Health Service Act (1946) was to build.

The National Health Service Act provided comprehensive coverage for most of the health services, including hospitals, general practice, and public health. The service remained at the periphery, however, in three types of care: (1) Primary medical care is given by family physicians or general practitioners. This service is organized on a local basis by an executive council. Each general practitioner has a list of people registered with him for whom he is responsible for providing primary care. (2) Specialist consultation and outpatient and inpatient treatment are provided in hospitals under the direction of regional hospital boards and boards of governors of teaching hospitals. A later concept makes each district general hospital responsible for providing hospital services for a defined population. (3) Services, such as health visiting, home nursing, home helps, domiciliary midwifery, the prevention of illness, and the provision of health centres are the responsibility of local authorities. This tripartite structure of the National Health Service was subjected to substantial criticism. Under an alternative proposal, area boards would be created, each having responsibility for the three combined branches of service.

Health services in the U.S.S.R.

In the Union of Soviet Socialist Republics, the protection and promotion of public health is the responsibility of the state. It is based on free public access to all forms of medical care. The principles upon which the health services are based are complete integration of curative and preventive services; medicine as a social service; the predominance of preventive programs; health centres or polyclinics (clinics in which a variety of diseases are handled) as the basis of operation; and community participation.

The public health services for the country as a whole are directed by the Ministry of Health under a minister who is a member of the Council of Ministers of the U.S.S.R. The ministry, through the 15 union republic ministries of health, directs all medical institutions within its competence as well as the public health authorities and services throughout the country.

The administration is centralized, with little local autonomy. Each of the 15 republics has its own ministry of health, which is responsible for carrying out the plans and decisions established by the U.S.S.R. Ministry of Health. Each republic is divided into *oblasti,* or provinces, which have departments of health directly responsible to the republic ministry of health. Each *oblast,* in turn, has *rayony* (municipalities), which have their own health departments accountable to the *oblast* health department. Finally, each *rayon* is subdivided into *uchastoki* (districts) with some three thousand to four thousand people in each.

In most rural *rayony* the responsibility for public health lies with the chief physician, who is at the same time medical director of the central *rayon* hospital. This system ensures unity of public health administration and implementation of the principle of planned development according to the social and economic plan for the country as a whole.

In the *uchastoki,* there are three types of salaried physicians: general practitioners or internists; pediatricians; and public health officers. There are also liberal numbers of other health personnel, including nurses, feldshers (*i.e.,* paramedical personnel trained in medical care), and midwives. In the more rural areas, there tend to be fewer physicians and more feldshers.

Feldshers generally undergo four years of training after eight years of general education, or three years if they have had the ten years of education. They function either as assistants to physicians when the number of physicians is adequate or relatively independently if the supply of physicians is inadequate. The use of the feldshers has helped the U.S.S.R. to make medical care available to large numbers of people who would not otherwise receive any care.

There are well-established referral procedures, from the polyclinics and smaller hospitals in the *uchastoki* to the larger *rayon* hospitals, and from feldshers and other paramedical personnel to internists and pediatricians and, when necessary, to more highly specialized personnel.

Medical education in the U.S.S.R. has four streams, currently chosen by students before their entry into medical school, leading to practice in internal medicine, pediatrics, public health, and stomatology (oral medicine).

The health services of the United States can be considered at three levels: local, state, and federal.

At the local level in cities or counties, there is substantial autonomy to function within broad guidelines developed by the state. The size and scope of local programs vary, but some of their typical functions are control of communicable diseases; clinics for mothers and children, particularly for certain preventive and diagnostic services; public health nursing services; environmental health services; health education; vital statistics; community health centres, hospitals, and other medical care facilities; community health planning and coordination.

At the state level, a department of health is charged with overall responsibility for health, though a number of agencies may actually be involved. The state department of health usually has five functions: public health and preventive programs; medical and custodial care such as the operation of hospitals for mental illness; expansion and improvement of hospitals, medical facilities, and health centres; licensure for health purposes of individuals, agencies, and enterprises serving the public; and financial and technical assistance to local governments for conducting health programs.

At the federal, or national, level, the Public Health Service of the Department of Health, Education, and Welfare is the principal health agency, but several other departments have specific health interests and responsibilities. Federal health agencies accept responsibility for helping to improve state and local services, for controlling interstate health hazards, and for working with other countries on international health matters. The federal government also has the following specific responsibilities: (1) protecting the United States from communicable diseases from abroad; (2) providing for the medical needs of military personnel, veterans, merchant seamen, and American Indians; (3) protecting consumers against impure or misbranded foods, drugs, and cosmetics; and (4) regulating production of biological products, such as vaccines. In addition, the federal government promotes and supports vast medical research, health services, and educational programs throughout the country.

Voluntary effort is a significant part of health work in the United States. There are more than 100,000 voluntary agencies in the health field functioning mostly at the local level but also at state and national levels. Supported largely through private sources, these agencies contribute to programs related to education, research, and health services.

Channels for medical care

Medical care is provided and paid for through multiple channels. It is provided through public institutions, such as municipal, county, state, and federal health centres, hospitals, and medical care programs, and through private channels, such as private hospitals and private practitioners working either alone or, increasingly, in groups. Generally speaking, medical care is financed by public funds, voluntary health insurance, or personal payment. There is a trend away from the traditional fee-for-service payment to individual practitioners toward prepaid-care systems including health teams working at community, health centre, and hospital levels.

Thus, in the United States there is great variety in the content, scope, and quality of health services. These services are provided by a multiplicity of independent agencies. In effect, however, they constitute a working partnership for the protection and promotion of human health.

In recent years, two factors have contributed to rapid change in the orientation of health services in the United States. One of these is an increasing awareness that, while the existing system for providing health services provides high quality services for many. there are many others for whom the care is either lacking or unsatisfactory. The second factor is that of steeply rising costs of medical care. These two issues have led to radical reconsideration of the entire system of personal medical care and proposals for entirely new systems of providing and financing health care.

Developing nations. ***Health problems and obstacles.*** The difficulties of providing health services for the people of the developing nations involve a cluster of interrelated problems. These problems may arise from the nature of the diseases and hazards to health, insufficient and maldistributed resources, the design of health services systems, and the education of health personnel who work in those systems.

With respect to, first, the diseases and hazards to health, an extreme situation is found in middle Africa, where infective and parasitic diseases, malnutrition, and environmental deficiencies are so prominent and destructive as to displace all other issues from major concern. By contrast, in the more advanced nations the only communicable diseases of major significance are tuberculosis and venereal disease, and the leading problems are those of age and industrial civilization. Woven through the health programs of the developing nations and complicating them at both family and national levels are the pressures associated with rapidly growing populations.

There are differences not only in the kinds of diseases prominent in different countries but also in the rates at which they occur and in the age groups involved. Life expectancy in some countries is less than half that in others, principally because of high death rates among small children in the developing countries. It is the children of the developing world who carry the great burden of ill health.

Infant mortality rates

In much of Southeast Asia, for example, 40 percent of children die by their fourth year, a death rate not reached until age 60 in North America. The infant (under one year of age) mortality rate in Middle and South America is two to four times that in North America, and the death rate in children one to four years of age is as much as 25 times greater. The differences for central Africa are even more striking: infant mortality in some areas has been 12 times that in the United States, and the mortality in preschool children has been more than 60 times the United States figure.

The principal causes of sickness and death among small children in the developing world are diarrhea, respiratory infections, and malnutrition, all of which are diseases intimately related to culture, custom, and economic status. Malnutrition is as often due to food customs as it is to the lack of food—taboos and simple oversight lead to deprivation of children too small to help themselves. Gastroenteritis (inflammation of the lining of the stomach and intestines. usually with accompanying diarrhea) and respiratory infections are often due to infectious organisms that are not susceptible to antibiotics. The interrelationships of these diseases increase the complexity of dealing with them. Malnutrition is often the underlying culprit; not only does it cause damage itself, such as retardation of physical and mental development, but it also seems to set the stage for other illnesses. A malnourished child develops gastroenteritis, inability to eat, further weakness, and then dehydration. Weakened in this way, the child is susceptible to a lethal infection, such as pneumonia. Or, to complete the vicious circle, infection can affect protein metabolism in ways that contribute to malnutrition.

Another factor that contributes to this grim picture is family size. Malnutrition, with associated death and disability, occurs most often in children born into large and poorly spaced families. The resulting high death rate among small children often reinforces the tendency of parents to have more children. People are not inclined to limit the size of their families until it is apparent that their children have a reasonable chance of survival. Thus, there is a fertility–mortality cycle in which high fertility, reflected in large numbers of small children crowded into a poor home, leads to high childhood mortality, which, in turn, encourages high fertility. This is the basis of the arguments that population-control programs should include effective means of reducing unnecessary deaths among children.

Ratios of medical personnel to population

Among limitations of resources, shortages of trained manpower are among the most important; ratios of population to physicians, nurses, and beds provide an indication of the seriousness of these deficiencies and also of the great differences from country to country (Table 1). Thus, the proportion of population to physicians in developing countries varies drastically. These figures change from year to year, but the changes are not dramatic. Nationwide figures are misleading because of the clustering of resources in the major cities. The number of persons per physician in Bangkok, Thailand, for example, is less than 1,000, while in the remainder of the nation it is 17,000. Even this latter figure is misleading, because there is another level of clustering in the provincial cities. Thus, the ratio for the truly rural population is about 200,000 per physician. In most countries of the developing world this rural figure is 50,000 or more.

Table 1: Ratio of Population to Physicians, Nurses, and Beds in Selected Countries

	ratio of population to physicians	ratio of population to nurses	beds per 1,000
Thailand	7,600	6,070	0.8
Jamaica	2,200	950	4.3
Colombia	2,000	17,100	3.1
Sudan	29,000	44,000	1.0
Senegal	20,000	5,500	1.3
Nigeria	50,000	7,100	0.4
Malawi	76,000	46,000	0.8
Guatemala	3,600	8,800	2.4
Iraq	4,900	6,700	2.1
Sweden	960	200	14.3

Source: World Health Organization, ***World Health Statistic Annual 1962,*** vol. 3, ***Health Personnel and Hospital Establishments*** (1966).

Money is a crucial factor in health care—it determines how many health personnel can be trained, how many can be maintained in the field, and the resources that they will have to work with when they are there. Governmental expenditures on health care vary from the equivalent of 20 cents (U.S. currency) per person per year in Indonesia to $56 in the United Kingdom. In contrast, the current expenditure by various governmental agencies in the United States is in excess of $300 per person per year. Rates of national development and the proportion of national income allocated to health indicate that health-expenditure figures will not change rapidly (Table 2).

Table 2: Governmental Expenditures on Health Services in Selected Countries

	year	government expenditure (in percent)	expenditure per inhabitant, U.S. (in dollars)
Nigeria	1964	12.0	0.50
Thailand	1963–64	3.4	0.60
Indonesia	1963	2.8	0.20
Malawi	1964	5.8	0.64
Sudan	1963–64	—	1.42
Iraq	1964–65	5.5	3.10
Colombia	1964	11.0	3.50
Jamaica	1963–64	11.0	9.60
U.K.	1963–64	12.9	56.00

Source: World Health Organization, ***Official Records, Third Report on the World Health Situation, 1961–64,*** no. 155 (1967).

For decades to come there will be less than one physician for every 50,000 people in the developing countries and only a few cents or a few dollars of governmental funds per person per year will be available for health expenditures. These realities of health care have important consequences on what can be done.

As it attempts to provide health care for its people, a nation, on the one hand, must meet the urgent and complex problems, such as obstetrical and surgical emergencies for which hospital care is essential. On the other hand, it must reach into the communities and homes to find those who need care but do not seek it and must discover the causes of such diseases as malnutrition and gastroenteritis.

Patterns shared. Developing countries have sometimes been influenced in the approaches they take to their health problems by the developed countries that have had a role in their historical development. The countries in Africa and Asia that were once colonies of Britain, for example, have educational programs and health-care systems that reflect British patterns, though there have been substantial adaptations to local needs. The health services are partly derived from British systems but are largely planned to meet the needs of those countries. Similar effects may be observed in countries influenced by France, The Netherlands, and Belgium.

The typical administrative pattern

Despite substantial variations from country to country, a reasonably common, if somewhat idealized, administrative pattern may be drawn for developing countries. All health services, except for a small amount of private practice, are under a ministry of health, in which there are about five bureaus, or departments — hospital services, health services, education and training, personnel, and research and planning. Hospital and health services are distributed through the country. At the periphery of the system are dispensaries, or health outposts, often manned by one or two persons with limited training. The dispensaries are often of limited effectiveness and are upgraded to full health centres when possible. Health centres and the activities that emanate from them are the foundation of the health-services system. Health centres are usually staffed by auxiliaries who have four to ten years of basic education plus one to four years of technical training. The staff of a health centre might include a midwife, an auxiliary nurse, a sanitarian, and a medical assistant. The assistant, trained in the diagnosis and treatment of sickness, refers to a physician the problems that are beyond his own competence. Together, these auxiliaries provide comprehensive care for a population of **10,000** to **25,000.** Several health centres together with a district hospital serve a district of about **100,000** to 200,**000** people. All health services are under the responsibility of the district medical officer, who has the exacting job of integrating the various health efforts into a comprehensive health program. He is assisted by other professional and auxiliary personnel.

Of central importance is the distribution of responsibilities between auxiliaries and professionals. The auxiliaries, by handling the large number of relatively simple problems, allow the professionals to look after only the more complex problems, to supervise and teach the auxiliaries, and to plan and manage the programs.

The district hospital is dependent on a regional hospital, to which patients with complex problems can be referred for more specialized services. Administrative direction of both regional health services and regional hospital services can be combined at this level under a regional medical officer. The central administration of the ministry of health provides policies and guidance for an entire health service and, in some instances, also provides a central planning unit.

Problems of transportation and communication over great distances, shortages of staff and other resources, and inadequacies in staff preparation and motivation often lead to malfunctions in the system. Nonetheless, the public health services developed in African and Asian countries have generally provided a sound basis for future development within the framework of national development.

Variations. The organization of public health services in Latin American countries differs substantially from those of Africa and Asia; these differences are largely an expression of their different historical backgrounds. The Latin American countries are generally more affluent than are the countries of Asia and Africa. Private practice is more widespread, and private or voluntary agencies are more prominent. Health services are provided largely by local and national governments. Many Latin American countries also have systems of clinics and hospitals for workers financed by employers and workers. The distribution of health services, with health centres, hospitals, and preventive services, is roughly as described in the previous section. The Latin American countries. however, have used auxiliaries less than African and Asian countries. Latin America has pioneered in the development of health-planning methods. Chile has one of the most advanced approaches to health planning in the world.

Differences between Latin America and Africa

Thailand is an example of a country that was never colonized and therefore has no historical influence favouring any particular pattern of health services. The Thai Ministry of Health has a well-developed system of hospitals and health centres distributed across the country to serve rural as well as urban people. It differs substantially from the pattern described in the previous section in that, despite the extreme shortages of physicians and nurses in rural areas, the nation has been reluctant to use auxiliaries for medical care. It does however, use auxiliary midwives and sanitarians. Hospital services and public health services have entirely separate administration. Within the public-health services, there are a number of separate divisions—*e.g.*, for tuberculosis, venereal disease, and nutrition — each with its own staff, budget, and facilities. The trend elsewhere has been away from relatively independent, disease-oriented approaches and toward integrated systems in which the same network of health services handles all or almost all problems.

***Education of** health personnel.* In the consideration of the education of health personnel, a particular set of problems emerges. Educational programs for auxiliaries are often well suited to the local situation, perhaps because they were not established in the more developed nations. Medical and nursing education, on the other hand, is similar to that of the more advanced countries, and it prepares students better for working in industrialized nations than in their own. This misfit between education and the jobs to be done has probably contributed substantially both to the lack of effectiveness of health services systems and to the migration of professional personnel to the more developed countries.

RECENT PROGRESS IN PUBLIC HEALTH

More developed nations. Among the more developed nations the following trends are apparent.

Increasing interest of national governments. An important trend in public health has been the increasing interest of national governments in the organization of medical care. Formerly, governments were chiefly concerned with basic health problems, such as environmental sanitation, medical care of the poor, quarantine, and the control of communicable diseases. Gradually, they have extended their activities into the field of medical care services in the home, clinic, and hospital, so as to provide comprehensive health care for entire communities. Three factors have influenced this trend: (1) the nongovernmental voluntary agencies have been unable to meet the rising cost of medical care; (2) there is an increasing appreciation of the economic loss to a country from sickness; and (3) there is an increasing public interest in social services.

Factors in the increase of government involvement

*The **broadening outlook on social welfare.*** Health and social welfare are now recognized as complementary, and social legislation tends to cover both areas. There is an administrative trend toward a close cooperation between health and social welfare services.

*Changing **concepts of preventable disease.*** Until recently, the term preventable disease referred to a circumscribed group of infectious diseases. The term is acquiring a broader meaning, however, as epidemiological methods are applied to other conditions. Preventive health services now deal with a wide range of health hazards, such as malignant tumours, rheumatism, cardiovascular diseases, other chronic and degenerative diseases, and even accidents.

Integration of preventive and medical care services. Medical care had its origin in the humanitarian motive of caring for the sick, while preventive services sprang from the need to protect a healthy environment from epidemic diseases. They grew apart from each other, but recently the trend has been to integrate them within a comprehensive health service. Such an integration is the fundamental principle of public health in the U.S.S.R. and in some other eastern European countries in which all local health services are centred in the district hospital under one administration. In other European countries, especially in rural areas, the two branches are brought together by the local medical practitioner. The focal point of many discussions is the role that the hospital should play in health services. Many authorities feel that its influence at present is too restricted and that it should spread beyond its walls to health centres and homes.

Provisions directed toward better mental health. A new orientation in mental health has allowed it to take a place in the preventive services. Improvements in arrangements for mental health include the provision of outpatient clinics and inpatient accommodations at general hospitals for early mental cases, an increase in child-guidance and marriage-guidance clinics, and schemes for the care of alcoholics and drug addicts. There have also been significant developments in the treatment of maladjusted members of society. Gains in understanding of psychoneuroses by general practitioners and the development of research facilities are also noteworthy.

Growing emphasis on health education. Many countries have expanded their work in health education, usually in cooperation with voluntary agencies. The most effective work is carried out at the local level, especially in schools. The trend is toward an expansion of health education as an essential preventive health service.

The biostatistical, epidemiological approach. A statistical service is essential in planning, administering, and evaluating health services. The interest of public authorities in medical-care schemes has vastly increased the importance of statistics on the incidence of diseases and other problems, as well as the epidemiology necessary to combat them. Both are vital in the planning, organization, and evaluation of medical care schemes. Traditionally, the epidemiological method was used for infectious diseases, but it is now being used increasingly for noninfectious diseases and the problems of medical care.

Changes resulting from an aging population. In more affluent nations, an increase in older age groups brings about the need for public health facilities to provide special services for them. Health care of the elderly includes measures to prevent premature aging and the chronic and degenerative diseases and to deal with the psychological problems resulting from loneliness and inactivity. Geriatric clinics have been set up to meet these needs and to conduct research into the process of senescence.

New hazards to health

Concern regarding the quality of the environment. There is widespread concern about environmental deterioration. The advent of controlled atomic radiation has created new hazards to health, such as the potential pollution of air or water by radioactive discharges, the possible effects from radioactive fallout on the public generally, and the dangers to workers in atomic installations in industry. An increasing population requires an increase in industrial and commercial activities, which add to the volume of pollutants that threaten the atmosphere, rivers, lakes, and oceans and have destructive effects on natural ecology. Individual countries have taken steps toward the control of environmental deterioration, and means of international regulation have also been proposed.

Developing nations. In view of the large numbers of serious health problems in the developing nations and their limited resources for dealing with them, it is understandable that along with substantial progress there would be some stagnation, or even regression.

Communicable-disease control. Smallpox and malaria are examples of diseases that have been brought under closer control throughout the world. For other diseases, such as hepatitis (liver inflammation), rabies, leprosy, and sleeping sickness, there have been important additions to understanding that can contribute to their eventual control.

Disease problems that await solution. El Tor cholera, which has appeared in epidemic form in previously uninvolved areas, represents one of the most serious challenges to public health. Venereal disease, an old problem, has increased at a disturbing rate. Certain parasitic diseases have spread as man has brought about changes in his environment—the increase in schistosomiasis (infestation with blood fluke by means of snails as the intermediate hosts) in irrigation and man-made lake areas is an example. Widespread malnutrition, particularly protein–calorie malnutrition in small children, remains a vitally important problem. Protein-rich food supplements and more effective educational programs are being developed to combat it.

Family health. The problems of rapidly growing populations have important consequences at both the family and the national level. Problems of maternal and child health, human reproduction, and human genetics, including family planning, are now seen as aspects of the greater problem of the health of the whole family as a single and fundamental social unit. Accordingly, family health is being viewed as a matter deserving high priority among the tasks of national public health services.

Health manpower. There is widespread recognition of inadequacies in both number and education of health personnel. The trend is in the direction of coordinating the education of all levels of health personnel with the particular health service in which the graduates will function. This trend, in turn, requires close relationships between educational institutions and the agencies responsible for health services.

Comprehensive community health services. The fragmentation of earlier health service organizations, such as single-disease-oriented programs and the separation of curative and preventive services, is giving way to more comprehensive organizational patterns. Health promotion, disease prevention, curing of the ill and their rehabilitation are brought together into one network of integrated services that reaches to the community level.

National health planning. Decisions of great complexity are involved in allocating limited resources to provide health services for large numbers of people. In order to achieve optimum results, there has been an increasing emphasis on the health-planning process and on the design of more effective health-services systems. A number of countries have established health-planning units in the ministry of health or the national planning organization. An important aspect of national health planning is the close coordination between planning, budgeting, implementing, and evaluating programs.

BIBLIOGRAPHY. C. FRASER BROCKINGTON, ***World Health,*** 2nd ed. (1968), a comprehensive treatment of public health concepts and international health organization; WORLD HEALTH ORGANIZATION, ***Health Services in Europe*** (1965), official WHO description of health services in Europe; ***Official Records of the World Health organization,*** annual report of the director general to the World Health Assembly and to the United Nations, reviewing major issues of world health; PHILIP E. SARTWELL (ed.), ***Preventive Medicine and Public Health,*** 9th ed. (1965), definitive textbook on preventive medicine and public health; ALFRED H. KATZ and JEAN SPENCER FELTON (eds.), ***Health and the Community*** (1965), a compilation of writings by experts on various aspects of community health; JOHN H. BRYANT, ***Health and the Developing World*** (1969), the product of a comprehensive study sponsored by the Rockefeller Foundation of health problems, health care systems, and the education of health personnel in Africa, Asia, and Latin America; JESSIE GARRAD and SIR MAX ROSENHEIM, ***Social Aspects of Clinical Medicine*** (1970), a consideration of how social medicine might be practiced in the United Kingdom; GEORGE ROSEN, ***A History of Public Health*** (1958), a classic writing on the history of public health throughout the ages.

(J.H.Br.)

Public Opinion

There are many difficulties in the way of defining public opinion, the most important of which are discussed below. A simple definition is that public opinion is an aggregate of the individual views, attitudes, and beliefs about a particular topic, expressed by a significant proportion of a community.

HISTORICAL BACKGROUND

Although the term public opinion was not used until the 18th century, phenomena that closely resembled public opinion seem to have occurred in many historical epochs. One of the oldest written records from ancient Egypt, a poem entitled "The Dispute With His Soul of One Who is Tired of Life," refers to an upheaval that apparently involved a complete reorientation of mass opinion:

> To whom shall I speak today?
> People are greedy. . . .
> Gentleness of spirit has perished.
> All the people are impudent. . . .
> People laugh at crimes of him who before
> Would have enraged the righteous. . . .
> There are no just men.
> The earth has been given over to evil doers.

Similar references to popular attitudes can be found in the history of Babylonia and Assyria. The prophets of ancient Israel sometimes justified the policies of the government to the people and sometimes appealed to the people to oppose the government. In both cases they were concerned with swaying opinion. And in classical Greece it was observed by many that everything depended on the people, and the people were dependent on the word. Wealth, fame, and respect all could be given or taken away by persuading the populace.

Wide dissemination of news, which is usually necessary for the formation of public opinion, could be observed in classical Rome. Much of this took place through person-to-person channels. When the Roman statesman Cicero was in Cilicia in the year 51 BC, he asked his friend Caelius to keep him informed of what was happening in the capital. Caelius promised to do so: "If anything important of a political nature should occur . . . I will diligently describe to you its origin, the general opinion about it, and the prospects of future action that it opens up." Rome also had its wall newspapers, composed by Roman officials and posted in public places to inform the public about acts of the government and principal local events.

Public opinion in the Middle Ages

During the Middle Ages in western Europe, the masses were encased in a rural, traditional society in which most activities and attitudes were dictated by a person's station in life; but phenomena much like public opinion could be observed among the religious, intellectual, and political elite. Religious disputations, the struggle between popes and the Holy Roman Empire, and the dynastic ambitions of princes all involved efforts to persuade, to create a following, and to line up the opinions of those who counted. In 1191 the English bishop William of Ely was attacked by his political opponents for hiring troubadours to extol his merits in public places, so that "people spoke of him as though his equal did not exist on earth." The propaganda battle between emperors and popes was waged largely through sermons, but handwritten literature also played a part.

From the end of the 13th century, the ranks of those who could be drawn into controversy regarding current affairs grew steadily. 'There was an increasing spread of education among the lay population. The rise of humanism in Italy saw the emergence of a group of writers and publicists whose services were eagerly sought by the princes who were consolidating national states. Some of these writers were used as advisers and diplomats; others were employed as publicists because of their ability to sway opinion. Such a one was the Italian Pietro Aretino (1492–1556), of whom it was said that he knew how to defame, to threaten, and to flatter better than all others and whose services were sought by both Charles V of Spain and Francis I of France. The Italian political philosopher Niccolb Machiavelli, a contemporary of Aretino, wrote that princes should not ignore popular opinion, particularly in regard to such matters as the distribution of offices.

Influence of printing on the growth of public opinion

The invention of printing from movable type in the 15th century and the Protestant Reformation in the 16th increased still further the numbers of people able to form opinions on contemporary issues. Martin Luther broke with the humanists by abandoning the use of classical Latin, intelligible only to the educated, and turned directly to the masses. "I will gladly leave to others the honour of doing great things," he wrote, "and will not be ashamed of preaching and writing in German for the unschooled layman." Luther's 95 theses, which were printed against his will and widely spread throughout Europe, were of a theological nature, but he also wrote on such subjects as the war against the Turks, the Peasants' Revolt, and the evils of usury. His vigorous expressions and the counterblasts from his many opponents, both lay and clerical, led to the formation of larger and larger groups holding opinions on important matters of the day.

Extensive attempts to create and influence public opinion were made during the Thirty Years' War (1618–48). A flood of propaganda tracts, many of them illustrated with woodcuts, emanated from both sides. Opinions were also swayed by means of speeches, sermons, and face-to-face discussions. Not surprisingly, both civil and religious authorities attempted to control the dissemination of unwelcome ideas by increasingly strict censorship. The first pope to have an Index of Prohibited Books drawn up was Paul IV in 1559. Charles IX of France decreed in 1563 that nothing could be printed without the special permission of the king.

More quietly, but more significantly, newspapers and news services were developing. Rudimentary private news services had been maintained by political authorities and wealthy merchants ever since classical times, but they were not available to the public. By 1500, however, it was possible to buy specialized news sheets in many of the principal cities of Europe. One of these, printed in 1514 or 1515, contains an extract of a merchant's letter telling of the Portuguese discovery of Brazil. The first regularly printed newspapers appeared about 1600 and multiplied rapidly thereafter, although they were frequently bedevilled by censorship regulations. Regular postal services, started in France in 1464 and in the Austrian Empire in 1490, facilitated the spread of information enormously.

Bourses as sources of informed opinion

The great news centres of early modern times were the financial exchanges. With the introduction of a paid civil service and the employment of paid soldiers in the place of vassals, princes found it necessary to borrow money. The bankers, in turn, had to know a great deal about the credit of the princes, the state of their political fortunes, and their reputations with their subjects. All kinds of political and economic information flowed to the money-lending centres at Antwerp, Lyon, and Niirnberg, and this information gave rise to generally held opinions in the banking community. The ditta di borsa—the opinion on the bourse—is often referred to in documents of the period. Queen Elizabeth of England was regarded as especially well informed because Sir Thomas Gresham, the finance agent of the English crown, kept in constant touch with the Antwerp bourse.

Significantly, it was another financial official who first popularized the term public opinion in modern times. Jacques Necker, finance minister of Louis XVI on the eve of the French Revolution, noted repeatedly in his writings that public credit depended upon the opinions of holders and buyers of government securities about the viability of the royal administration. He, too, was vitally concerned with the ditta di *borsa*. But he also remarked on the power of public opinion in other areas. "This public opinion," Necker wrote, "strengthens or weakens all human institutions." As he saw it, public opinion should be taken into account in political undertakings. Necker was not, however, concerned with the opinions of all Frenchmen. For him, the people who collectively shaped public opinion were those who could read and write, who lived in cities, who kept up with the day's news, and who had money to buy government securities —in short, the bourgeoisie.

A public opinion that extended beyond the middle classes and embraced the urban masses took shape during the French Revolution. Observers of the Revolution were mystified, and often terrified, by this new phenomenon of public opinion, which seemed able to sweep aside the entrenched institutions of the time: the monarchy, the church, and the feudal system. Thinkers of the late 18th and early 19th centuries advanced a variety of definitions as to what public opinion actually was. One of the most detailed descriptions was given in 1799 by the German poet Christoph Wieland, who closely followed the stormy events in France and the rest of western Europe:

> I, for my part, understand by it an opinion that gradually takes root among a whole people, especially among those who have the most influence when they work together as a group. In this way it wins the upper hand to such an extent that one meets it everywhere. It is an opinion that without being noticed takes possession of most heads, and even in situations where it does not dare to express itself out loud can be recognized by a louder and louder muffled murmur. It then requires only some small opening that will allow it air, and it will break out with force. Then it can change all nations in a brief time and give whole parts of the world a new configuration.

A German philosopher of the time, Christian Garve, gave much more emphasis to the rational component:

> Public opinion as interpreted . . . by those French writers who are clearest on the subject is the agreement of many or of the majority of the citizens of a state with respect to judgments which every single individual has arrived at as a result of his own reflection or of his practical knowledge of a given matter.

The English philosopher Jeremy Bentham, who advanced the first detailed discussion of public opinion in English, was troubled by the difficulty of defining it and advised that the term be employed only in deference to common usage.

RECENT DEFINITIONS AND CONCEPTIONS OF PUBLIC OPINION

In spite of voluminous discussions of the subject, scholars still do not agree on a definition of public opinion. Members of a roundtable of the American Political Science Association that met in 1925 divided into three groups: those who did not believe that there was such a thing as public opinion; those who accepted its existence but doubted their ability to define it precisely; and those who could offer a definition. This last group could not, however, agree on the definition to be adopted. Although few scholars now question the existence of such a phenomenon as public opinion, differences in defining it have persisted to the present day.

These differences stem in part from the varying perspectives with which scholars have approached the study of public opinion and in part from the fact that the phenomenon is still not completely understood. Political scientists and some historians have tended to emphasize public opinion's role in the governmental process, paying particular attention to its influence on government policy. Some political scientists have regarded public opinion as equivalent to the national will. In this sense, there can be only one public opinion on an issue at any one time.

Sociological conceptions of public opinion

Sociologists usually give more emphasis to public opinion as a product of social interaction and communication. According to the sociological view, there can be no public opinion without communication among members of the public who are interested in a given issue. A large number of persons may hold quite similar views, but these will not coalesce into public opinion as long as each person remains ignorant of the opinions of the others. Communication may take place by means of the mass media of the press, radio, and television or through face-to-face discussions. Either way, people learn how others think about a given issue and may take the opinions of others into account in making up their own minds.

Sociologists suggest that there may be many different public opinions existing on a given issue at the same time. One body of opinion may be dominant or may be reflected in governmental policy, but this does not mean that other organized bodies of opinion do not exist. The sociological approach also sees the public-opinion phenomenon as extending to areas that are of little or no concern to government. Thus, fads and fashions are appropriate subject matter for students of public opinion, as are public attitudes toward movie stars or corporations.

It is often the case that opinions expressed in public may differ from those expressed in private and that only the former contribute to public opinion. Similarly, some attitudes--even though widely shared—may not be expressed at all. Thus, in a totalitarian state, a great many people may be opposed to the government but may fear to express their attitudes even to their families and friends. In such cases, an antigovernment public opinion fails to develop.

Private opinions, if expressed in public, may become a basis for public opinions. Until the 1930s, for example, there was an unwritten prohibition in the United States against discussions of venereal disease, although many indivduals had private opinions about it. Then, when the subject began to be treated in the mass media and public opinion researchers began to ask questions about it, opinions that had formerly been private were expressed in public, and sentiment in favour of government action to stamp out venereal disease developed.

Some public-opinion survey specialists have preferred a definition that links public opinion directly to their polling procedures. Public opinion is therefore defined as being identical to what people's responses to a survey questionnaire would be. Other similar definitions have been to the effect that public opinion is whatever is discovered by public-opinion polls. This definition, while widely used in practice, has the disadvantage of implying that public opinion does not exist in places and times in which there are no opinion polls. A more generally applicable approach that embodies much the same reasoning is that public opinion on any matter may be conceived as the hypothetical result of some imaginary survey or vote.

Those who are primarily engaged in the manipulation of public opinion, notably professional politicians and public relations men, rarely stop to define it. The American journalist and political scientist Walter Lippmann has observed that there has been a tendency in democracies to make a mystery out of public opinion but that "there have been skilled organizers of opinion who understood the mystery well enough to create majorities on election day." (**Public *Opinion,*** 1922.) Public relations practitioners have concerned themselves less with public opinion in general than with the opinions of specified "publics" that may affect the fortunes of a client: employees, stockholders, government officials, suppliers, and potential buyers, for example. Both politicians and public relations men are interested in influencing behaviour and thus in determining any attitudes and opinions that may affect that behaviour, whatever they may be called.

Nearly all scholars and manipulators of public opinion, regardless of the way they may define it, agree that at least four factors are involved in public opinion: there must be an issue; there must be a significant number of individuals who express opinions on the issue; there must be some kind of a consensus among at least some of these opinions; and this consensus must directly or indirectly exert influence.

THE FORMATION AND CHANGE OF PUBLIC OPINION

The democratic system itself defines a number of issues on which citizens are under pressure to form opinions. They are called upon to decide among various candidates in elections and, on occasion, to vote on constitutional amendments and various other propositions. Almost any matter on which the executive or legislature has to decide may become a public issue, if a significant number of persons wish to make it one. The attitudes of these persons are often stimulated or reinforced by outside agencies—a crusading newspaper, a pressure group, or a governmental agency or official. Even matters that are not within the purview of any governmental agency may become public issues. No agency, for example, has the authority to determine how many children a family may have or how long a man's hair should be, but discussion on these subjects has been sufficiently intense to generate widespread opinion about them.

Elements involved in the formation of attitudes

Once a public issue is identified, a certain number of people will begin to form attitudes about it. If the attitude is expressed to others by sufficient numbers of people, a public opinion on the topic begins to emerge. Not all people develop attitudes on public issues; some may not be interested, and others simply may not hear about them. The attitudes that are formed may be held for various reasons. Thus, four men may all be opposed to higher property taxes but for very different reasons. One man may not be against higher taxes in principle, but he opposes them because he is having trouble paying the mortgage on his house. This attitude serves an adjustment, or utilitarian, function in that it helps its holder to accommodate the immediate financial situation in which he finds himself. A second man may fight the tax because he does not want a certain social group, such as the poor or the unemployed, to derive any benefit from tax revenues. Such an attitude may be the result of a psychological insecurity, of a desire to keep the poor "in their place" in order to bolster his own sense of superiority toward underprivileged groups. For such a person, the atitude serves an ego-defensive function. A third man may esist the tax increase because he believes that governmental activities should be severely restricted. His attitude has a value-expressive function in that it reflects his overall philosophy. A fourth man opposes the increased tax because he is familiar with instances of governmental waste and is convinced that all necessary services could be rendered if officials spent the already available funds more rationally. His attitude is thus determined by knowledge or experience in that it is a reflection of what he has learned in the past. A fifth man, of course, might fight the tax for all four reasons. A seemingly homogeneous body of public opinion may thus be comprised of individual opinions that are rooted in very different interests and values. If an attitude does not serve a function such as one of the above, it is unlikely to be formed: an attitude must be useful in some way to the person who holds it.

How many people will actually form opinions on a given issue, as well as what sort of opinions they form, depends partly on their own pre-existing knowledge, attitudes, and values; partly on the personal situations in which they find themselves; and partly on a number of environmental factors. As far as pre-existing knowledge and attitudes are concerned, it is often surprising to discover how many people are not informed about major issues and therefore have no attitudes toward them. In 1964, for instance, one in four Americans did not know that the government in China was Communist, and about the same proportion at any given time is ignorant of any major issue of public policy. Substantially the same situation has been found to exist in western European countries, and ignorance is probably even more widespread in nations with lower levels of education. Values are of considerable importance in determining whether people will form opinions on a particular topic. If people feel that their moral principles or personal philosophies are involved in an issue, they are more likely to take a favourable or opposing stand.

Environmental factors play an extremely important part in the formation of attitudes and opinions. Most pervasive is the influence of the immediate social environment: family, friends, neighbourhood, place of work, church, or school. People usually adjust their attitudes to conform with those that are most prevalent in the social groups to which they belong. If a person who considers himself a liberal is surrounded in his home or at his place of work by people who profess conservatism, he is more likely to switch his vote than is a liberal whose family and friends share his political views. Similarly, it was found in World War II that men transferred from one unit to another often adjusted their opinions to conform more closely with those in the unit to which they were transferred.

The influence of mass media

Other important environmental factors involved in the formation of public opinion include the existence of pressure groups, advertising, and public relations and the attitudes of opinion leaders inside and outside government. The press, radio, and television are usually less important than the immediate social environment when it comes to the formation of attitudes, but they are still significant. They focus the attention on certain personalities and issues, and many people subsequently form opinions about these issues. Government officials have noted that their mail from the public tends to "follow the headlines"; whatever is featured in the press at a particular moment is likely to be the subject that most people write about. The mass media can also activate and reinforce latent attitudes. Political attitudes, for example, are likely to be activated and reinforced just before an election. Voters who may have only a mild preference for one party or candidate before the election campaign starts are often worked up by the mass media to a point where they not only take the trouble to vote but may contribute money or help a party organization in some other way.

The mass media play another extremely important role in letting individuals know what other people think and in giving leaders large audiences. In this way they make it possible for public opinion to include a large number of individuals and to spread over wider geographic areas. It appears in fact that in some European countries the growth of broadcasting, and especially television, has affected the operation of the parliamentary system. Before television, national elections were seen largely as contests between a number of candidates or parties for parliamentary seats. More recently, elections in such countries as West Germany and Great Britain have appeared more as a personal struggle between the leaders of the principal parties concerned, since these leaders were featured on television and came to personify their parties. Television in France and the United States has been regarded as a powerful force strengthening the presidential system, since the President can easily appeal to a national audience over the heads of elected legislative representatives.

Even when the mass media are thinly spread, as in developing countries or in nations where the media are strictly controlled, word-of-mouth can sometimes perform the same functions as the press and broadcasting, although on a more limited scale. In developing countries, it is common for those who are literate to read from newspapers to those who are not, or for large numbers of persons to gather around the one village radio. Word of mouth in the marketplace or neighbourhood then carries the information further. In countries where important news is suppressed by the government, a great deal of information is transmitted by rumour. The official newspaper of the Chinese Communist Party has denounced "little broadcasting stations," as rumour carriers are sometimes called in the People's Republic, criticizing them for distributing unauthorized news about the party and the state. Word of mouth thus helps public opinion to form in developing countries and encourages "underground" opinion in totalitarian countries, even though these processes are slower and usually involve fewer people than in countries where the media network is dense and uncontrolled.

The role of pressure groups

Pressure groups, or interest groups, also play an important part in the formation and spread of public opinion on issues of relevance to themselves. These groups may be concerned with political, economic, or ideological issues and often work through the mass media as well as by word of mouth in trying to influence attitudes. Some of the larger or more affluent interest groups in the U.S., western Europe, and elsewhere make use of advertising and public relations to influence opinion. During the late 1960s and early 1970s, for example, hundreds of thousands of dollars were spent in the U.S. on advertising by opponents and proponents of United States policy in Southeast Asia, and lesser amounts were devoted to interest group advertising on other public issues. In Britain, much advertising space was purchased by opponents of Britain's proposed entry into the Common Market.

Opinion leaders play a major role in defining important issues and in influencing individual opinions regarding them. Political leaders, in particular, can turn a hitherto relatively unknown problem into a national issue if they decide to call attention to it. One of the ways in which opinion leaders rally opinion and smooth out the differences among those who are in basic agreement on a subject is by coining or popularizing symbols or slogans: Sir Winston Churchill popularized the phrase cold war, and the Allies in World War I were fighting "a war to end all wars." Slogans are perhaps among the most useful tools that are available to the political leader. Once enunciated, symbols and slogans are frequently kept alive and communicated to large audiences by the mass media and may become the cornerstone of public opinion on any given issue.

Opinion leaders are not confined only to prominent figures in public life. There are likely to be persons in every social group to whom others in the immediate environment look for guidance on certain subjects. Thus, one person may be thought of by those in his own social group as especially qualified in the realm of local politics, another as a reliable guide in foreign affairs, and a third as an expert when it comes to buying a house. These local opinion leaders are generally unknown outside their own circle of friends and acquaintances, but their cumulative influence in the formation of public opinion is substantial.

Although a person's psychological makeup, his personal circumstances, and external factors such as pressure groups and opinion leaders all play a role in the formation of opinions, it is still not known exactly how public opinion on an issue takes shape. Many aspects of the public opinion process have yet to be explored. The same is true with regard to changes in public opinion. Some of these can be accounted for by changing events and circumstances, but others are more difficult to explain. It is known that public opinion on some subjects tends to follow events. Public attitudes toward other nations, for example, seem to depend largely on the relations between the governments of the two nations. Hostile attitudes do not cause poor relations; they are the result of them. People presumably change their attitudes when these attitudes do not correspond with their perception of prevailing circumstances and hence are not useful as guides to action. It is also frequently the case that an issue ceases to be important and simply fades from public attention, while new issues arise as the basis for new bodies of opinion. There are still, nevertheless, major changes in public opinion that are difficult to explain. During the second half of the 20th century in many parts of the world, attitudes toward religion, family, sex, international relations, social welfare, and the economy have undergone major shifts. There have been important issues claiming public attention in all these areas, but the changes in public opinion are difficult to relate to any major event or even to any complex of events.

PUBLIC OPINION AND GOVERNMENT

Many early thinkers saw public opinion as a powerful force that rulers must learn how to control. The 18th-century French philosopher Jean-Jacques Rousseau believed that all laws were based upon it but that this did not necessarily diminish the powers of government. It was Rousseau's opinion that "Whoever makes it his business to give laws to a people must know how to sway opinions and through them govern the passions of men." The 19th-century German philosopher G.W.F. Hegel described public opinion as containing both truth and falsehood together and added that it was the task of the great man to distinguish the one from the other. Jeremy Bentham saw the greatest difficulty of the legislator as being "in conciliating the public opinion, in correcting it when erroneous, and in giving it that bent which shall be most favourable to produce obedience to his mandates."

Public opinion as a safeguard against abuse of authority

At the same time, Bentham and some other thinkers believed that public opinion was a useful check on the authority of rulers. Bentham demanded that all official acts be given publicity, so that an enlightened public opinion could pass on them, as would a tribunal: "To the pernicious exercise of the power of government it is the only check." The British jurist and historian James Bryce, writing a century later, maintained that if government was based on popular consent, this would give a nation great stability and strength: "It has no need to fear discussion and agitation. It can bend all its resources to the accomplishment of its collective ends." Bryce did not, however, believe that mass opinion could or should dominate details of governmental policy, since most people did not have the leisure or inclination to arrive at a position on every question. Rather, the masses would set the general tone for policy, their sentiments leading them to take a stand on the side of justice, honour, and peace.

Those who worked to advance international understanding were particularly likely to invoke the power of public opinion. Both Bentham's "Plan for an Universal and Perpetual Peace" (1789) and Immanuel Kant's proposals in his essay "On Perpetual Peace" (1795) were based on the belief that public opinion is peace loving and that international peace can be sustained by it. Those who advocated establishment of the League of Nations after World War I also looked to world public opinion as the principal force that would sustain the League. Drafters of the charter of the United Nations Educational, Scientific, and Cultural Organization (UNESCO) apparently had somewhat the same idea, noting that because the origins of war were to be found in the minds of the masses of men it was in the minds of men that the defenses of peace should be constructed.

Some scholars, while acknowledging the power of public opinion, warned that it could be a dangerous force. The 19th-century French writer Alexis de Tocqueville was concerned about the possible "tyranny of the majority" if government was in fact to be an expression of mass attitudes. Many other writers have expressed concern, often in a more extreme form, about the dangers of allowing government policy to be influenced too much by public opinion, which may well be uninformed, unthinking, and unstable. But whether public opinion is regarded as a constructive or a baneful force in a democracy, there are few politicians who are prepared to deny in public that government should follow public opinion.

An assessment of the significance of public opinion

In recent years, political scientists have been less concerned with what part public opinion should play in a democratic polity and have given more attention to establishing what part it does play in actuality. From the examination of numerous histories of policy formation, it is clear that no sweeping generalization can be made that will hold in all cases. The role of public opinion appears to vary from issue to issue, and the way it asserts itself differs from one democracy to another. The safest generalization that can be made is that public opinion does not influence the details of most policies but that it does set limits within which the policy maker must operate. That is, public officials will usually seek to satisfy a widespread demand, or will at least take it into account in their deliberations, and they will also try to avoid decisions that they believe will fly in the face of popular opinion. In addition, it has been observed that the relation between public opinion and public policy is two way; policy influences opinion, as well as the reverse, and there is usually at least an initial tendency for the public to accept a decision once it is made. Public opinion seems to be particularly effective in influencing policy making at the local level, as officials appear to feel themselves constrained to yield to popular pressures for better roads, better schools, or more hospitals.

Public opinion at the national level seems to play a more limited role—partly because of the inability of most people to understand the complexities of most issues faced by government and partly because of the growth of executive power and the development of large governmental bureaucracies that serve as screens between the policy maker and the public. Representative government itself also tends to limit the power of public opinion to influence specific decisions, since ordinarily the public is given the choice only of approving or disapproving the election of a given official.

PUBLIC-OPINION POLLING

Public-opinion polling can provide a fairly exact analysis of the distribution of opinions on almost any issue within a given population. Assuming that the proper questions are asked, polling can also reveal something about the intensity with which opinions are held, about the reasons for these opinions, and about whether or not the issues have been discussed with others. Polling does not usually reveal whether or not the people holding an opinion can be thought of as constituting a cohesive group, and it is unlikely to provide very much information about the elites who may have played an important part in developing the opinion. But in spite of these deficiencies, polling is a valuable tool for estimating the state of public opinion on almost any subject.

Origins of opinion research

Opinion research developed from market research. Early market researchers picked small samples of the population and used these to obtain information on such questions as how many people read a given magazine or listen to the radio and what the public likes and dislikes in regard to various consumer goods. About 1930, both commercial researchers and scholars began to experiment with the use of these market research techniques to obtain information on opinions about political issues. In 1935 the American psychologist George Gallup began conducting nationwide surveys of opinions on current political and social issues in the United States. One of the first questions asked by the Gallup Poll (its full name is the American Institute of Public Opinion) was "are Federal expenditures for relief and recovery too great, too little, or about right?" To this, 60 percent of the sample replied that they were too great, only 9 percent thought they were too little, and 31 percent regarded them as about right.

From the 1930s on, the spread of opinion polls conducted by both commercial and academic practitioners continued at an accelerated pace in the U.S. and, to a more limited extent, elsewhere. In 1937, the Public *Opinion Quarterly*, which later became the official organ of the American Association for Public Opinion Research, began publication. State and local polls, some sponsored by newspapers, were started in many parts of the country, and opinion research centres were organized at several universities. Before and during World War II, opinion polls were extensively used by U.S. government agencies, notably the Department of Agriculture, the Treasury Department, and the War Department, and some government-sponsored polls continued into the postwar period.

At the same time, opinion research was increasingly used in other parts of the world. Several affiliates of the American Institute of Public Opinion were organized in Europe and Australia in the late 1930s, and following World War II polling organizations appeared in numerous countries of Asia and Latin America. By 1970 the World Association for Public Opinion Research had over 300 members in 41 nations, including several eastern European countries. The association does not include any members from the Soviet Union, although numerous public-opinion surveys have been conducted there by Soviet scholars.

Polls have been successful in forecasting election results in nearly every case in which they have been used for this purpose. The two most notable failures were in the United States in 1948, when nearly all polls forecast a Republican victory and the Democrats won by a narrow margin, and in Great Britain in 1970, when all but one of the major polls predicted a Labour Party victory and the Conservative Party achieved a majority. In both cases the major polls did not show large deviations from the actual results but nevertheless picked the wrong winner. Professional opinion researchers point out that predicting elections will always be chancy because of the possibility of last-minute shifts and turnout problems; nevertheless, their record has been good over the years.

Although popular attention has been focussed on polls taken before major elections, most polling is devoted to other subjects, and university-based opinion researchers usually do not make election forecasts at all. Support for opinion studies comes largely from public agencies, foundations, and commercial firms, which are interested in such questions as how well people feel that their health, educational, and other needs are satisfied; how such problems as racial prejudice and drug addiction can be attacked; or how well a given industry is meeting public demands. Polls that are regularly published in newspapers or magazines usually have to do with some lively social issue—and elections are included only as one of many subjects of interest.

Methodology of opinion polling. The principal steps in opinion polling are the following: defining the "universe," choosing a sample, framing a questionnaire, interviewing persons in the sample, tabulating the results, and analyzing or interpreting the results.

Universe is the term used to denote whatever body of people is being studied. This is not always easy to define precisely. If, for example, one is making a study of college-student opinion, it is necessary to decide whether the universe should include full-time students only or whether it should include those who are not candidates for established degrees. The way in which these questions are answered will have an important bearing on the outcome of the survey and possibly on its usefulness.

Once the universe has been carefully defined a sample of the universe has to be picked. If possible, a "probability sample" should be chosen. Ideally, the best way to do this would be to assign a number to each person in the universe––or write his name on a slip of paper—place all the numbered or named slips in a container, mix thoroughly, and then pick a sample without looking at the names or numbers. In this way, each slip will have the same probability of being chosen. If each person is numbered, the same effect can be achieved by using tables of random numbers, which can often be purchased in book form. The random numbers are matched up with the numbered members of the universe, until a sample of the desired size is drawn. The numbering procedure is often not practicable, but a few universes are already assigned numbers —all the workers in a given factory, for instance, or all members of the armed forces. It is not surprising that some of the most reliable opinion surveys have been conducted by the military.

Another method, not quite so reliable statistically, is to include every "nth" member of the universe in the sample. Thus, if one wishes to study the attitudes of the subscribers to a certain magazine and the magazine has 10,000 subscribers, one could take every tenth name from the subscription list and end up with a sample of 1,000

Quota sampling

Neither of these methods is likely to be useful when the universe consists of a large population that has not been numbered and when the names of members of the universe are not listed in a card file somewhere. This is the situation that was faced by market and opinion researchers when they first started conducting large-scale surveys. They therefore adopted the simpler device of picking a "quota sample." In quota sampling, an effort is made to match up the characteristics of the sample with those of the universe, so that a small replica of the universe is achieved. If one knows, possibly on the basis of the most recent census, that there are 51 women to every 49 men in the universe, then the sample should reflect these proportions. The same principle should be applied with respect to age, income, education, occupation, religious preference, national origin, area of residence, and indeed every characteristic that might be relevant to the range of opinions being studied. Each interviewer is told the characteristics of the people that he must locate and interview.

Most survey organizations prior to 1948 used quota samples, and some still do. The British organization that correctly forecast the outcome of the 1970 parliamentary election used a quota sample, but it is, nevertheless, a risky technique. In many countries census data are poor or nonexistent. Even when there are reliable census data, some characteristics that may affect the opinions being studied cannot be taken into account. It is not known, for example, how many vegetarians there are in most populations or how many extroverts and introverts. Yet these characteristics may be related to opinions on certain subjects. Statisticians point out that in a quota sample it is impossible to give each member of the universe a known chance of being selected, and one cannot therefore calculate the range of error in the results that could be due to chance. In this type of sample, furthermore, interviewers have to use their judgment in selecting respondents, and their standards may vary. The great advantage of a quota sample is that it is rather inexpensive to design and interview. By contrast, selecting and interviewing a probability sample from a large population can be very expensive.

How large a sample should be depends on the precision that is desired. For many purposes, a sample of a few hundred is adequate—if it is properly *chosen.* A magazine, for instance, might poll a random sample of 200 of its subscribers and find that 18 percent wanted more fiction and 62 percent wanted more articles on current social issues. Even if each of these figures was wrong by as much as 10 percentage points, the poll would probably still be of value since it would give fairly accurate information about the way the subscribers ranked the types of content. An election poll, on the other hand, would have to be much more accurate than this, since leading candidates often split the vote rather evenly. For most purposes, a national sample of about 1,500 cases is adequate *unless* it is desired to make comparisons among rather small subgroups in the population or to compare one small group with a much larger one. In such cases a larger sample is required in order that a significant number of members of the minority group may be included.

Allowances for chance and errors

There are no hard and fast rules for interpreting poll results, since there are many possible sources for bias or error. Nevertheless, for a well-conducted poll the following "rule of thumb" allowances for error are helpful. When any group of people is compared with another and the sample size of the smaller group is about 100, the difference between the percentages based on the two groups should be greater than 14 if this difference is to be regarded as significant. If the smaller group is larger than 100, the allowance for error caused by chance decreases approximately as follows: for a group comprising 200 cases, allow 10 percentage points; for 400 cases, allow 7 percentage points; for 800, allow 5; for 1,000, allow 4; for 2,000, allow 3. Thus, if a national sample survey shows that 27 percent of college students favour a volunteer army, while 35 percent of adults who are not in college do, and there are only 200 students in the sample, the difference between the two groups may well be due to chance. If the difference were greater than 10 percentage points, then it is much more likely that the opinions of students actually differ from those who are not in college. Similar allowances have to be made when election polls are interpreted. The larger the sample and the larger the difference between the number of preferences expressed for each candidate, the greater the certainty with which the election result can be predicted.

Even larger variations than those due to chance may be caused by the way the questions are worded. If one poll asks "Are you in favour of increasing government aid to higher education?" while another poll asks: "Are you in favour of the president's (or premier's) recommendation that government aid to higher education be increased?" the second question is likely to receive many more affirmative answers than the first. Similarly, the distribution of replies will often vary if an alternative is stated. The question might be phrased: "Are you in favour of increasing government aid to higher education, or do you think enough tax money is being spent on higher education now?" In this form, it is probable that the question would receive fewer affirmative responses than if only the first half were used. As a rule, relatively slight changes in question wording do not cause great variations in response when people hold very strong opinions, but if their opinions are not firm then slight differences may sway them one way or another. Opinion researchers therefore frequently ask exactly the same question over a period of years. In this way, the results from an earlier survey can more safely be compared with the results from a later one.

Questionnaire construction, as with sampling, requires a high degree of skill. The question must be clear to people of varying educational levels and backgrounds. Questions must not embarrass respondents. They must be arranged in a logical order and so on. Even experienced researchers find it necessary to pretest their questionnaires. They send out interviewers to interview a small sample with the preliminary questions, and they may revise these questions to ensure that they are unambiguous and are actually obtaining the information sought.

Poll questions may be of the "forced-choice" or "free-answer" type. In the former, a respondent is asked to reply "yes" or "no" or else may be given a list of prepared alternatives. Even so, many respondents are likely to reply "don't know" or to prefer an alternative that the researcher had not listed in advance. A free-answer question allows the respondent to state his opinion in his own words. For instance, "What do you think are the most important problems facing the country today?"

Importance of careful interviewing

Interviewing is also a skilled operation. An inexperienced interviewer may bias a respondent's answer by the way he asks a question. He may antagonize some respondents so that they refuse to go on with the interview. He may nor record the replies to free-answer questions accurately, or he may not be sufficiently persistent in locating designated respondents. Most large polling organizations give interviewers special training before they are sent out on surveys or else contract with an interviewing service that has trained and experienced interviewers available. A good sample and a well-tested questionnaire are not sufficient to guarantee an accurate survey if interviewing is slipshod.

Tabulation is usually done by machine. To simplify this process, most questionnaires are "precoded," which is to say that numbers appear beside each question and each possible response. The answers given by respondents can thus be translated rapidly into a numerical form that can be used in a computer. In the case of free-answer questions, responses must usually be grouped into categories, each of which is also assigned a number. How the categories are defined may make a large difference in the way the results are presented. If a respondent mentions narcotics addiction as a major problem facing the country, for instance, this answer might be coded as a health problem or a crime problem or might be grouped with other replies dealing with drug abuse or alcoholism.

The final steps in a survey are the analysis and presentation of results. Some reports present only what are termed marginals—the proportion of respondents giving certain answers to each question. If 40 percent favour one candidate, 50 percent another, and 10 percent are undecided, these figures are marginals. Usually, however, a number of cross tabulations are also given. These may show, for instance, that Candidate A's support comes disproportionately from Jewish groups, and Candidate B's, from Irish groups. Sometimes a cross tabulation will substantially change the meaning of survey results. A poll may seem to show that one candidate is the favourite of black voters and another of white voters. But if the preferences of poor respondents and well-to-do respondents are analyzed separately, it may turn out that Candidate A is actually supported by most poor people, and Candidate B, by most well-to-do people. The most important factor in determining voting intention may thus not be whether a respondent is white or black but whether he is well-to-do or poor.

Assessing a survey's reliability

In judging the overall reliability of a survey, it is advisable to scrutinize at least eight factors: (1) the identity of the sponsor and the past record of reliability of the organization conducting the poll; (2) the exact wording of the questions used; (3) the care with which the population sampled has been defined; (4) the size of the sample and the method by which it was chosen; (5) the "completion rate," or proportion of the sample that actually responded (this is especially important in mail polls, in which frequently fewer than half of those in the sample respond); *(6)* the degree to which particular results are based on the whole sample or on small parts of it; (7) the way in which the interviewing was done (whether by telephone, mail, or face-to-face); and (8) the time that the survey was taken (intervening events frequently make people change their opinions).

Criticisms and justifications of opinion surveys. There have been many criticisms of public-opinion polling. Among these are that people are asked to give opinions on matters about which they are not competent to judge, that polling interferes with the democratic process, and that survey research causes annoyance and invasion of privacy.

It is often pointed out that most members of the public are not familiar with the details of complex policies such as those governing tariffs or missile defense systems. Therefore, it is argued, opinion researchers should not ask questions about such subjects. The results at best could be meaningless and at worst misleading, since respondents may be reluctant to admit that they are ignorant. Critics also refer to the fact that many people hold inconsistent opinions, as shown by the polls themselves. The same person may favour larger government expenditures and at the same time be opposed to higher taxes.

Poll takers usually acknowledge that these problems exist but maintain that they can be overcome by careful survey procedures and by proper interpretation of results. It is common for surveys to include "filter" questions, which help to separate those who are familiar with an issue from those who are not. Thus, the interviewer might first inquire: "Have you heard or read about the government's policy on the tariff?" Then the interviewer would ask only those who answered "yes" whether they were or were not in favour of the policy advocated by the government. Sometimes polls include factual questions that help to assess knowledge, such as "Can you tell me how the veto power in the United Nations Security Council works?" Furthermore, argue the researchers, if people are ignorant, or if they hold inconsistent opinions, this should be known. It is not possible to raise the level of information if areas of ignorance or inconsistency are not identified.

Critics allege also that election polls create a "bandwagon effect"—that people want to be on the winning side and therefore switch their votes to the candidates whom the polls show to be ahead. They complain that surveys undermine representative democracy, since issues should be decided by elected representatives on the basis of the best judgment and expert testimony—not on the basis of popularity contests. They point out that some well-qualified can dates may decide not to run for office because the polls indicate that they have little chance of winning and that a candidate who is far behind in the polls has difficulty in raising funds for campaign expenditures since few contributors want to waste money on a lost cause. Other candidates find out from polls what the public wants and merely pander to popular preferences rather than run on their convictions about what is best for the country.

Those engaged in election research usually concede that polls may dissuade some candidates and also may inhibit campaign contributions. But they also point out that candidates and contributors would have to make their decisions on some basis anyway. If there were no polls, other and less accurate methods would be used to test public sentiment; columnists and political pundits would still make forecasts. As far as the bandwagon effect is concerned, careful studies have failed to show that it exists.

An abuse that is recognized by both critics and poll takers is the practice of leaking to the press partial or distorted results from private polls. A politician may contract privately with a research organization and may then release only results for those areas in which he is ahead, or he may release old results without stating the time when the poll was taken, or he may conceal the fact that a very small sample was used and that the results may have a large margin of error.

Finally, critics aver that the proliferation of opinion polls and market research surveys places an unfair burden on the public. People may be asked to respond to questionnaires that take an hour or more of their time. Pollers may tie up their telephones or occupy their doorsteps for long periods, sometimes asking questions about private matters that are not suitable subjects for public inquiry. But insofar as public resistance to polling is concerned, researchers point out that the "refusal" rate in most surveys is rather low. Most people, in fact, seem to enjoy answering the questions. They also note that, with the use of small samples, it is unlikely that any one individual will be approached very often.

Legislation to deal with these and other problems of poll taking has been proposed in the United States, Britain, and a number of other countries; however, little legislation has been adopted. Survey researchers maintain that such abuses as exist should be dealt with by the profession and by educating the public to evaluate and criticize poll results, rather than by government regulation.

BIBLIOGRAPHY. There is no adequate history of public opinion in English. The best account is W. BAUER, *Die Öffentliche Meinung in der Weltgeschichte* (1929); see also H. SPEIER, "Historical Development of Public Opinion," *American Journal of Sociology*, 55:376–388 (1950); and P.A. PALMER, "The Concept of Public Opinion in Political Theory," in *Essays in History and Political Theory in Honor of Charles Howard McIlwain* (1936). General texts include: H.L. CHILDS, *Public Opinion* (1965); B.C. HENNESSEY, *Public Opinion*, 2nd ed. (1970); J.W. ALBIG, *Modern Public Opinion* (1956); N.J. POWELL, *Anatomy of Public Opinion* (1951); and L.W. DOOB, *Public Opinion and Propaganda* (1948). Classic treatments of public opinion include: J. BRYCE, *The American Commonwealth*, vol. 2 (1900); A.L. LOWELL, *Public Opinion and Popular Government*, rev. ed. (1926); A.V. DICEY, *Lectures on the Relation of Law and Public Opinion in England During the Nineteenth Century* (1914); and W. LIPPMANN, *Public Opiniorz* (1922). Historical approaches to the study of public opinion are explored in a special issue of the *Public Opinion Quarterly*, vol. 31, no. 4 (1967–68); a discussion of attitudes and attitude change, *ibid.*, vol. 24, no. 1 (1960). The nature and formation of public opinion in the U.S. are dealt with in G.A. ALMOND, *The American People and Foreign Policy* (1950); excellent use of poll data is made in V.O. KEY, JR., *Public Opinion and American Democracy* (1961). S. KELLEY, JR., *Professional Public Relations and Political Power* (1956), presents fascinating case histories of the manipulation of public opinion. Principal election studies, which reveal a great deal about public opinion, include: P.F. LAZARSFELD, B. BERELSON, and H. GAUDET, *The People's Choice* (1944); B. BERELSON, P.F. LAZARSFELD, and W.H. MCPHEE, *Voting* (1954); and A. CAMPBELL *et al.*, *The American Voter* (1960). Diffusion of information and opinion leadership is discussed in E. KATZ and P.F. LAZARSFELD, *Personal Influence* (1955). Two excellent expositions of sampling method are F.F. STEPHAN and P.J. McCARTHY, *Sampling Opinions* (1958); and M.H. HANSEN, W.N. HURWITZ, and W.G. MADOW, *Sample Survey Methods and Theory*, 2 vol. (1953). Detailed treatments of the conduct of surveys may be found in H.H. HYMAN, *Interviewing in Social Research* (1954), and *Survey Design and Analysis* (1955).

(W.P.D.)

Public Utilities, U.S.

The main services supplied by public utilities are (1) local and interregional transportation—airlines, buses, motor freight carriers, gas and oil pipelines, railroads, and water carriers; (2) telephone and telegraph; (3) power, heat, and light; and (4) community facilities for water, sanitation, and irrigation.

In nearly every country such enterprises are owned and operated by the state and as such are covered in the article PUBLIC ENTERPRISES. This article is confined to the coverage of the virtually unique situation of the U.S., where public utility services are supplied, by and large, by private firms subject to government regulation.

THE ECONOMIC BASIS OF REGULATION

Reasons for the special treatment of utilities

The tendency toward monopoly. The classic economic explanation of the need for extensive regulation of public utilities is that such businesses are "natural monopolies." The term is misleading. As James R. Nelson has argued:

> "natural monopolies" in fact originated in response to a belief that some goal, or goals, or public policy would be advanced by encouraging or permitting a monopoly to be formed. and discouraging or forbidding future competition with this monopoly. (From "The Role of Competition in the Regulated Industries," *The Antitrust Bulletin*, xi, Jan.–April 1966, p. 3.)

It is more accurate, therefore, to say merely that the technology of production, transmission, and distribution almost inevitably leads to a complete or partial monopoly of the market for the services rendered and that competition in the public utility sector is often less effective and desirable than competition in industry in general.

The tendency toward monopoly arises from the following factors. Some public utility enterprises are subject to "economies of scale"; that is, assuming a given state of technology, one firm of the most efficient size can produce all or more than the market demands at a lower average cost than can two or more competing firms. Or. stated another way, the larger the output of a utility plant, such as a generating station, the lower will be the cost of production per unit of output. Economies of scale, however, are not unique to public utility enterprises; they are found in virtually every business. What is of significance is the fact that a utility's market (in industries other than transportation) is restricted by the necessarily close connection between the utility plant and the consumers' premises. Further, because storage of a utility's service is limited at best and because a utility must have adequate capacity to meet its customers' peak demand requirements, it tends to have unused or surplus capacity most of the time. Competition serves to aggravate this situation; whereas if a utility is free to cultivate its market area intensively. it can attract commercial, industrial, and residential customers with diversified demands and thereby minimize its unused capacity.

Other reasons. These factors making for monopoly are reinforced by four others. First, utilities require a large investment in fixed, and highly specialized, plants. Electric utilities, for example, must invest about **$4.50** to get **$1** of annual gross sales, while manufacturing companies as a group must invest only about **47** cents to get **$1** of sales. This investment, largely bound to its original location, represents a significant percentage of a utility's total cost. Second, utility equipment must be located below, upon, and over public property. (Utilities, in fact, are provided with the governmental power of eminent domain, which makes possible the compulsory sale of private property.) Space limitations do not in themselves lead to monopoly; but as the number of conduits or mains increases, the streets are torn up more frequently, thus creating a public nuisance. Moreover, the number of desirable sites also may be limited, such as those for hydroelectric power plants. Third, for many uses, demand for the product tends to be "inelastic"; that is, as the price of a utility's service increases, consumers increase their expenditures rather than decrease their demand; utilities, therefore, to the extent that they are supplying basic necessities, have some (but far from absolute) control over the rates they charge consumers. Finally, price differentiation or discrimination among customers is both possible and generally attractive to utility enterprises. Fixed costs represent a large percentage of total costs; unused capacity exists much of the time; and there are important differences in the demands of various consumers. A utility company often finds that a single rate low enough to maintain full capacity output fails to yield revenues sufficient to cover costs, while one set high enough to cover costs will result in unused capacity. Under these circumstances, discrimination may increase revenues and minimize unused capacity.

While these are the major economic characteristics of public utilities, they are not confined to public utilities nor do they suffice to explain why such enterprises are subject to detailed regulation. For electric power generation transmission, and distribution, and for the transportation and distribution of natural gas, regulation is based on the existence of significant economies of scale in a given local or regional market. For telephone and telegraph services, regulation also is based on significant economies of scale as well as on the expense and inconvenience to consumers of having parallel competing systems. Few question the need of regulating these industries. Many, however, question the current need for detailed regulation of natural-gas producers and of the surface transportation industry. In these industries, economies of scale are not such an important consideration. Regulation of natural-gas producers is based on the proposition that the concentration of ownership is high enough to raise the possibility of monopolistic pricing practices. Regulation of the railroads was initially based on their oligopolistic market structure (*i.e.*, few competitors) over long-haul routes and their monopolistic market structure over local and short-haul routes. In both cases, discriminatory pricing was a profitable practice. The development of other, competing modes of transport reduced the economic justification for detailed regulation of the railroads, but instead of being abolished, regulation was extended to the competing modes. The regulation of bus, highway freight, and water transport was largely predicated on the need to prevent cutthroat or destructive competition within each of these modes and to prevent undue price discrimination throughout the transportation industry. Airline regulation was primarily based on safety considerations and on the promotion of the industry.

The real distinction between utilities and other businesses is thus one of degree; utilities possess certain economic characteristics to a greater degree than do non-utility industries. Economies of scale, for example, partly explain the need for large plants in many industries. High fixed costs are found in the concrete and steel industries, among others. Further, even though the services of public utilities are considered essential or necessary, society does not regard them as so necessary that they should be provided irrespective of the purchaser's ability to pay for them. And some important manufacturing industries have escaped regulation similar to that imposed upon public utilities not because they are more competitive than the utilities but because of the difficulties of regulating the manufacture of commodities. The desire to enforce competition in such industries led to the enactment of the Sherman Antitrust Act in **1890** and the establishment of the Federal Trade Commission in **1914.** The tendency for many businesses to practice price discrimination caused passage of a law limiting such discrimination as early as **1914** (Clayton Act) and the strengthening of this law in **1936** (Robinson-Patman Act).

THE LEGAL BASIS OF REGULATION

Major judicial decisions

Federal regulation of business proceeds under Article I, section 8, of the Constitution, which gives Congress power "to regulate Commerce ... among the several states." The interstate commerce clause has been broadly interpreted, particularly since the late 1930s, and it seems unlikely to impose significant limits on federal regulatory powers in the future. The Supreme Court held in a **1946** decision that federal power to regulate interstate commerce "is as broad as the economic needs of the nation." (From *American* Power & Light Co. v. Securities & Exchange Commission, **329** U.S. **90.)**

The public interest concept. The Constitution provides that those powers not delegated to the federal government and not specifically prohibited by it to the states may be exercised by the states. States thus have the broad authority to legislate for protection of the health, safety, morals, and general welfare of their citizens. These are known collectively as the "police powers" of the states and also have been interpreted broadly. In the words of the Supreme Court in Nebbia v. New ***York*, 291** U.S. **502 (1934)**:

> . . . a state is free to adopt whatever economic policy may reasonably be deemed to promote the public welfare, and to enforce that policy by legislation adapted to its purpose. . . . Price control, like any other form of regulation, is unconstitutional only if arbitrary, discriminatory, or demonstrably irrelevant to the policy the legislature is free to adopt.

The rights of corporations are protected against invasion by the acts of government in both federal and state constitutions. Corporations, in short, are guaranteed due process of law. Moreover, under the U.S. constitutional system, all acts of legislatures and administrators are subject to judicial review. The final authority is the Supreme Court of the United States.

The public-utility concept. In the well-known case of *Munn* v. *Illinois*, **94** U.S. **113** (1877), the Supreme Court attempted to establish a separate category of businesses, known as public utilities, that required detailed government regulation. In the *Munn* case, Chief Justice Morrison Waite, in upholding the validity of an **1871** Illinois statute fixing maximum rates for storing grain in elevators, noted both the importance of the grain trade and the strategic position of the grain elevators. ("They stand ... in the very 'gateway of commerce,' and take toll from all who pass.") Quoting from Lord Chief Justice Sir Matthew Hale's words of nearly two centuries earlier, he held that "when private property is 'affected with a public interest, it ceases to be *juris privati* only.' " Property becomes affected with a public interest

> when used in a manner to make it of public consequence, and affect the community at large. When, therefore, one devotes his property to a use in which the public has an interest, he, in effect, grants to the public an interest in that use, and must submit to be controlled by the public for the common good, to the extent of the interest he has thus created. He may withdraw his grant by discontinuing the use; but, so long as he maintains the use, he must submit to control.

In 1914 the Supreme Court upheld regulation of the fire insurance business (***German Alliance Insurance Co.*** v. ***Kansas,*** 233 U.S. 389). Said Justice Joseph McKenna: insurance

> is practically a necessity to business activity and enterprise. It is, therefore, essentially different from ordinary commercial transactions, and ... is of the greatest public concern.

In unanimously holding unconstitutional a state of Kansas statute providing for extensive control of commodity prices and wage rates in ***Wolff Packing Co.*** v. ***Court of Industrial Relations of Kansas,*** 262 U.S. 522 (1923), the court concluded:

> The circumstances which clothe a particular kind of business with a public interest, in the sense of Munn v. ***Illinois*** and the other cases, must be such as to create a peculiarly close relation between the public and those engaged in it, and raise implications of an affirmative obligation on their part to be reasonable in dealing with the public.

A majority of the Supreme Court, in 1928, opposed the extension of the public-utility concept to employment agencies, holding that the interest of the public in the matter of employment is not "that ***'public interest'*** which the law contemplates as the basis for legislative price control" (***Ribnik*** v. *McBride,* 277 U.S. 350). And a year later the Court held unconstitutional a Tennessee statute that sought to fix the prices at which gasoline could be sold within the state (***Williams*** v. ***Standard Oil Co.,*** 278 ***U.S.*** 235). The business of dealing in gasoline, said the Court, "does not come within the phrase 'affected with a public interest.' " The Supreme Court's last attempt to clarify the public-utility concept came in the 1934 ***Nebbia*** case cited above. The New York legislature, in 1933, enacted a statute that declared that the milk industry was affected with a public interest. The statute set up a Milk Control Board to control the prices and trade practices of milk producers and distributors. In upholding the statute, Justice Owen Roberts stated that (1) "the dairy industry is not, in the accepted sense of the phrase, a public utility"; (2) "there is in this case no suggestion of any monopoly or monopolistic practice"; and (3) "those engaged in the business are in no way dependent upon public grants or franchises for the privilege of conducting their activities." He added:

> It is clear that there is no closed class or category of businesses affected with a public interest. . . . The phrase "affected with a public interest" can, in the nature of things, mean no more than that an industry, for adequate reason, is subject to control for the public good. . . . If the law-making body within its sphere of government concludes that the conditions or practices in an industry make unrestricted competition an inadequate safeguard of the consumer's interests, produce waste harmful to the public, threaten ultimately to cut off the supply of a commodity needed by the public, or portend the destruction of the industry itself, appropriate statutes passed in an honest effort to correct the threatened consequences may not be set aside because the regulation adopted fixes prices reasonably deemed by the legislature to be fair to those engaged in the industry and to the consuming public.

It seems clear, then, that the public-interest concept is no longer synonymous with the public-utility concept and that the latter concept is included within the former. Nevertheless, the essential elements of the public-utility concept were succinctly outlined by Judge Frederick Vinson in a 1943 decision (***Davies Warehouse Co.*** v. ***Brown,*** 137 F. 2d 201):

> If a business is (1) affected with a public interest, ***and*** (2) bears an intimate connection with the processes of transportation and distribution, ***nnd*** (3) is under an obligation to afford its facilities to the public generally, upon demand, at fair and non-discriminatory rates, ***and*** (4) enjoys, in a large measure an independence and freedom from business competition brought about either (a) by its acquirement of a monopolistic status, or (b) by the grant of a franchise or certificate from the State placing it in this position, it is ... a public utility. ...

The obligations of public utilities

Public utilities have four major obligations imposed on them because of their special status. First, they are obligated to serve all who apply for service. Within a market (service) area, and within the limit of its capacity (ability to serve), a utility must be prepared to serve any customer who is willing and able to pay for the service. Second, they are obligated to render safe and adequate service and must be prepared for foreseeable increases in demand. Third, they have the obligation to serve all customers on equal terms. Unjust or undue discrimination among customers is forbidden. Finally, they are obligated to charge only a "just and reasonable" price for the services rendered.

THE DEVELOPMENT OF REGULATORY COMMISSIONS

Early regulation. Modem regulation is carried out largely by administrative commissions or agencies, but such has not always been the case. The earliest form of regulation was judicial: enforcement of the common law duties through lawsuits brought by individuals who thought themselves injured. Legislative regulation gradually supplanted judicial regulation during the second half of the 19th century, and by 1898 the function of the judiciary had been restricted to the review of legislative and administrative acts.

Legislative regulation by means of corporate charters or special franchise was the next to be tried. Above all else, such regulation proved to be inflexible. Each change in a charter required a special legislative amendment. But since legislatures were in session only a small percentage of the time, and since they found their attention being claimed by many other matters, the required adjustments often were delayed. Likewise, each change in a franchise provision had to be approved by both parties since the Supreme Court, in an 1819 decision (***Trustees of Dartmouth College*** v. ***Woodward, 4*** **Wheaton 518),** held that a franchise had the same status as a contract. Further, it was impossible for the legislature to pay much attention to financial and accounting control or to service and safety aspects of regulation.

Railroad commissions. As the early methods of regulation proved defective and, as the demand for more stringent and continuous control arose, "independent" regulatory commissions, operating under general legislative statutes, were created. The first commissions, generally those created before 1870, were largely advisory bodies, and their chief concern was with the railroads. They made recommendations to state legislatures and railroad managements; appraised property taken by the railroads under the right of eminent domain; enforced railroad safety standards; and, basically, served as fact-finding bodies. They had no control over rates.

Shortly after the beginning of the Granger movement in the Midwest—a populist movement that sought state legislation to control local railroad and grain elevator rates—the first commissions with mandatory powers were established. Between 1871 and 1874, Illinois, Iowa, Minnesota, and Wisconsin established commissions with power to set maximum rates, prevent discrimination, and forbid mergers of competing railroad lines. While the Granger laws, except in Illinois, were repealed by the end of the 1870s, they established a pattern followed by other states. By 1887, when Congress created the Interstate Commerce Commission (partly patterned after the British Railway Commission of 1873) to regulate the nation's railroads, 25 states had regulatory commissions.

State and federal regulation

Public utility commissions. The third, or modern, period began in 1907 with the creation of two powerful commissions: in New York under the leadership of Gov. Charles Evans Hughes, and in Wisconsin under the urging of Sen. Robert M. LaFollette. In both states the legislatures extended the regulatory powers of the commissions to utilities other than railroads: gas, light, power, telephone, and telegraph companies. Both commissions, moreover, were delegated broad powers, including security regulation, examination of accounts and property, the fixing of rates, the requirement of detailed reports in prescribed form, and the right to prescribe uniform systems of accounts. These two commissions became models, and by 1920 more than two-thirds of the states had regulatory commissions. The state commissions were strengthened, their jurisdictions extended, and their powers further increased after the stock market crash of 1929; and several federal commissions were established to regulate interstate commerce. Today, all 50 states, plus the District of Columbia, have commissions known as public-utility or public-service commissions, railroad commissions, corporation commissions, or commerce commissions.

There are five federal commissions with jurisdiction over the interstate activities of public utilities: (1) the Interstate Commerce Commission, with jurisdiction subsequently mtended to cover oil pipelines (1906), motor carriers (1935), interstate and coastal water carriers (1940), and freight forwarders (1942); (2) the Civil Aeronautics Board, created in 1940 (as a reorganization of the Civil Aeronautics Authority of 1938), with jurisdiction over commercial air transportation; (3) the Federal Power Commission, organized as an independent agency in 1930, with regulatory powers over hydroelectric projects and the transportation and sale at wholesale of electric power (since 1935) and natural gas (since 1938); (4) the Federal Communications Commission, established in 1934 (succeeding the Federal Radio Commission of 1927), with jurisdiction over broadcasting and—by a transfer of powers from the Interstate Commerce Commission—interstate telephone and telegraph services; and (5) the Securities and Exchange Commission, organized in 1934 and given power under the Public Utility Holding Company Act of 1935 to regulate the finances and corporate structures of electric and gas utility holding companies.

At the state level, a majority of the commissions are composed of three or five members appointed by the executive or elected either by popular vote or by the legislature with overlapping four-year to six-year terms. At the federal level, the commissions are composed of 5, 7, or 11 members appointed by the President, with Senate approval, for terms ranging from five to seven years. Federal commissions are bipartisan by law.

PRINCIPLES OF REGULATION

Setting a return on capital

It is the basic objective of regulation to prevent a public utility from earning excessive (monopoly) profits and from engaging in unreasonable (inequitable) price discrimination among customers, commodities, and places. Thus rate regulation—*i.e.*, control of the rate level (earnings) and control of the rate structure (rates or prices)—has been the major task of the commissions. But if rate regulation is to be effective, control also must be exercised over accounting procedures, financial matters, service, and safety.

The primary standard of rate regulation is the revenue-requirement or capital-attraction standard. Stated simply, a utility is permitted to set rates that (1) will cover operating costs and (2) will result in a reasonable rate of return (profit) on the property devoted to the business.

The rate level. The first aspect of rate regulation, the rate level, involves the determination of the utility's total revenue requirement and can be expressed as a formula:

$$R = O + (V - D)r$$

where R is the total revenue required, O is the operating costs, V is the gross value of the tangible and intangible property, D is the accrued depreciation of the tangible and reproducible property, and r is the rate of return.

Operating costs include all types of operating expenses (wages, salaries, fuel, maintenance, and research) plus annual charges for depreciation and operating taxes. For a typical electric or telephone company, operating expenses average about 44 percent of revenues, depreciation 12 percent, and taxes 22 percent—a total of 78 percent of revenues. Many of these costs are determined by normal competitive forces (wages, fuel. and maintenance) or bv various levels of government (taxes); others are determined by the individual firms (expenditures on advertising, research and development, and charitable contributions). While a utility legally may spend any amount it chooses for such purposes, all expenditures are subject to review by the relevant commissions as to their reasonableness.

The net or depreciated value of the tangible and intangible property is known as the rate base. The value of the utility's "used and useful" tangible property may be measured in either of two ways: original cost (cost of plant and equipment when built or purchased) or reproduction cost (value of plant and equipment expressed in current dollars). The proper measure has been a continuous source of controversy ever since the early days of regulation, with the utilities generally favouring reproduction cost (especially in times of rising prices) and a majority of the commissions favouring original cost. Regardless of the measure used, an appropriate amount of depreciation must be subtracted in order to reflect the depreciated value of the property. Land, usually separated from other tangible property, is commonly valued either at its original cost or at the value of adjacent property. No depreciation is subtracted since land tends to appreciate in value over time. The rate base also includes an allowance for working capital and, in some circumstances, amounts for interest during construction, water rights, and leaseholds. In former years the utilities argued that several intangibles—particularly franchise value, going concern value, and good will—should be considered, but current commission practices exclude these items.

The fair rate of return is usually expressed as a percentage of the rate base. For example, a rate base of $200,000,000 combined with an 8 percent return results in an annual allowance of $16,000,000 as the fair-return component. (To this figure must be added operating costs to determine the utility's total revenue requirement.) With respect to the fair rate of return, there is no formula that can be used; like other aspects of regulation, a fair rate of return is a matter of judgment. The return allowed, however, should perform two functions. It should be fair to investors so as to avoid the confiscation of their property. It should also preserve the credit standing of the utility to permit it to attract new capital so as to maintain, improve, and expand its services in response to consumer demand. Utilities must compete for investment funds in the capital market with non-utility businesses. Moreover, they are not guaranteed a fair rate of return; they are entitled to a fair return only if it can be earned.

The rate structure. The second aspect of rate regulation, the rate structure, involves the establishment of the rates (prices) the utility is to be permitted to charge its customers so as to earn the required revenue. Utility rate structures are highly differentiated. Customers are classified into various groups (*e.g.*, residential, commercial, industrial), and the rate charged each group is different. And within each group, rates are typically reduced as the quantity purchased increases, with the result that customers may not even pay the same average rate per unit. Such rates are based upon both supply (cost of service) and demand (value of service) considerations. When based upon the latter, discrimination results. As noted earlier, such discrimination is permitted as long as it is "just and reasonable," since often it is beneficial to the utility and to the customer. In recent years, with the growing strength of competitive forces, a new dimension has been added to the rate structure problem. This new dimension involves the complex issue of defining properly the costs relevant to competitive rates.

Accounting and financing. At the turn of the century and continuing into the 1930s, irregular accounting procedures such as the overstatement of property values and financial abuses such as overcapitalization occurred throughout American industry, including the utilities. As the accounting profession and business matured, abuses decreased markedly, and they now are primarily problems of the past. Today a majority of the commissions prescribe uniform systems of accounts and have power to regulate or control the issuance of securities. Most commissions also require competitive bidding on new security issues.

Service and safety. The control of service and safety has two aspects: quality and quantity. Quality refers to such matters as methods of billing, accuracy of meters, continuity of service, deposits and repayments, and treatment of complaints. Quantity refers to such matters as entry restrictions, service abandonment, consolidations, mergers, and acquisitions. Entry, for instance, including both the certification of a new company and the certification of an existing firm to serve a new area or a new route, is rigidly controlled by the commissions. Certificates of public convenience and necessity are required to provide most utility services. Each applicant must show that the proposed service is required by the public and that it is "fit, willing, and able" to perform properly the proposed service and to conform to all relevant regulations. Safety regulation is of particular importance with respect to the transportation industries and includes such matters as driver or pilot qualifications, air-traffic control, inspection and maintenance, and accident reporting and investigation. All jurisdiction over transport safety has been transferred, at the federal level, from the regulatory commissions to the Department of Transportation.

PROBLEMS OF SPECIFIC INDUSTRIES

Rate, service, and safety regulation are common to all public utilities. In other respects the problems of each industry differ.

utilities and the public interest

Transportation. Government policy toward the transportation industry has sought two basic, and often conflicting, objectives—promotion of its use (by extending public aid in many forms) and regulation (by controlling entry, mergers, rates, and service). But since present policy has evolved in a piecemeal fashion over three-quarters of a century, little attention has been given to the development of a coordinated policy. Several postwar studies have concluded that government policy has promoted waste and inefficiency throughout the transportation industry. Serious inconsistencies in policy have developed. Control of entry has worked to protect existing carriers; control of minimum rates has often allocated traffic among competing carriers without consideration of the carriers' relative cost advantages. The position of common carriers, plagued by overcapacity or financial problems, has grown increasingly difficult.

There is growing support for a radical change in government policy and for greater reliance upon the forces of competition. It is generally recognized that investment (both public and private) in transport facilities must be increased significantly for such purposes as improved airports and aviation safety and for long-run solutions to the pressing urban transport problem, including the acute commutation problem in the larger cities.

Electric power. The combined functions of generating, transmitting, and distributing electric energy comprise the nation's largest industry. Private (investor-owned) electric utilities dominate the industry, but public power (municipal and federal) and cooperatives (under the Rural Electrification Act of 1936) are of considerable significance. Indeed, the private versus public issue has been an important one for many years. Technological advances in generation (including nuclear power) and transmission have made feasible a rapid growth in interconnections and power pools but also have raised the issue of reliability (*i.e.*, how best to achieve coordination by all segments of the industry to prevent failures, such as the dramatic Northeast power failure of 1965, when a series of local failures led to a total blackout extending from New York City to Canada). A new wave of consolidations within the industry has resulted in fewer and larger systems, causing much concern as to the future of smaller (both private and public) electric systems. Competition between electric power and the fossil fuels has become stronger as electricity has invaded the winter heating market and natural gas the summer air-conditioning market. Greater emphasis has been placed upon promotional practices and competitive rate making. And in locating and building the new facilities required for the future, the electric utilities have come under increased pressure to consider the conservation of natural resources, the maintenance of scenic or natural beauty, and the preservation of historic sites.

Natural gas. Natural gas was an unimportant source of domestic energy prior to the 1920s. As large oil and gas fields were discovered and as improvements in pipeline construction were made, the industry's importance grew. Gas now supplies nearly one-third of the nation's total energy. It is generally accepted that pipeline and local distribution companies should be regulated; they have the traditional economic characteristics of regulated industries. The regulation of independent producers, however, has caused constant controversy ever since the Supreme Court, in 1954, held that the Federal Power Commission's jurisdiction extended to such producers (Phillips Petroleum Co. v. Wisconsin, 347 U.S. 672). A movement for exemption by Congress failed, and the FPC subsequently adopted an area-rate approach on the basis of Re Phillips Petroleum *Co.*, 35 P.U.R. 3d 199 (1960):

> the determination of fair prices for gas based on reasonable financial requirements of the industry and not on the particular rate base and expense of each natural gas company.

But many issues remain to be resolved before the new approach can be called a success. Among those issues are how high a reserves-production ratio is needed and what price is required to achieve it. Finally, since natural gas is limited in quantity, conservation inevitably enters into the commission's certification decisions.

Communications. The communications (telephone and telegraph) industry is dominated by the American Telephone and Telegraph Company and the Western Union Corporation. Competitive forces, however, have raised new problems and challenges in regulating this expanding industry. Technological advance has tended to blur the traditional distinction between voice and record communication and also has made feasible the entry of the common carriers into data processing and specialized information services. The convergence of communications and computer technologies also has induced the entry of equipment manufacturers and other firms into the supplying of communications systems and services. Recent decisions by the Federal Communications Commission appear to favour more competition for the domestic common carriers. As a result, competitive rate making has assumed greater importance. Other issues include the desirability of a domestic satellite system (and its ownership) and the proper allocation of the scarce radio spectrum. And on the international front, communications satellites and submarine cables offer alternative and competing means of providing telecommunications services.

CONTINUING QUESTIONS

Government regulation of private enterprise raises many complex problems. Some of these involve the regulatory process itself. In the early days of regulation the independent commission was thought to. have a number of advantages over legislative or judicial regulation, including continuity of policy, expertise, impartiality, and flexibility of procedure. In practice, commissions have not been able to escape their political environment. They are subject to pressures from the executive and legislative branches and also from the industries under their jurisdiction. They depend on the executive for their appointments and budget requests, on the legislature for their jurisdiction and appropriations, and on the judiciary for review of their actions. It has been argued that commissions tend to get transformed from vigorous protectors of the public interest into captives of the interests they are set up to regulate, with the regulatory process becoming simply a means of maintaining the status quo. Independence, in short, has not been achieved.

The future of the regulatory commission

Even if independence were possible, there is some question as to its desirability. Some have contended that cornmissions should not be isolated from the political support and leadership that is essential for the success of regulation and for the achievement of national economic goals.

Some of the problems are external to the commissions.

Appointments frequently have been made on the basis of political considerations rather than training or experience. The statutes that commissions seek to enforce often lack definitive standards and may even be out of date. Salaries, especially at the state level, are too low to attract and hold fully qualified personnel; budgets are insufficient to permit adequate staffs.

Some of the problems, finally, are perhaps inherent in regulation itself. As Clair Wilcox has argued in *Public Policies Toward Business:*

> Regulation is static, backward-looking, preoccupied with the problems of the past. It does nothing to stimulate change, seeking to maintain order on the basis of the old technology. It is slow to adapt to change; new problems appear, but regulatory thinking lags. Competition, by contrast, is dynamic. . . .
>
> Regulation is slower than competition. It must satisfy the requirements of due process: investigate, give notice, hold hearings, study the record, make findings, issue orders, permit appeals. All this takes time and delays action. . . .

While not everyone agrees with all of these criticisms, few would deny that there is room for improvement. Toward this goal, several steps already have been taken. The establishment of the Department of Transportation (1966) represents a partial attempt to coordinate domestic transportation policy; the Federal Power Commission's *National Power Survey* (1964) represents a similar attempt in the electric power field. Many commissions have tried to streamline their procedures and reduce regulatory delay: prehearing conferences are used to determine the areas of agreement and of dispute; previously prepared and distributed (so-called canned) testimony often is permitted, reducing the burden of trial examination; informal proceedings are increasingly being used in place of formal procedures. And, in an effort to provide for further improvements in administration, the Administrative Conference of the United States (a permanent, independent government agency) was established by Congress in 1964.

CONCLUSION

The problems of regulation are numerous, but this does not mean that regulation of public utilities has been a failure. The economic performance of many utilities has been impressive. With the exception of transportation, the growth of utilities in the postwar period has far exceeded that of the unregulated industries—nearly double the growth rate of the economy as a whole. Their output per man-hour in the last decade has increased faster than for any other industrial sector. Their rates have risen far less than prices generally (in many cases, their rates have actually declined), while their returns on capital have been considerably below those earned by other industries. They have accounted for a significant percentage of the nation's capital investment. And, finally, they have provided a wide variety and high quality of service. Only in the transportation industry does the economic performance reflect badly upon the regulatory process itself.

The great challenge to regulation is whether it can answer the requirements of the future. Many of the issues that commanded attention in the past have been largely settled. Others have taken their place—problems of pollution and aesthetics, competitive pricing, and promotional practices, to mention only a few. The goals and policies of regulation must be made more explicit; the flexibility of the regulatory process must be increased by eliminating complex, costly, and often confusing procedures; and criteria must be developed for evaluating the performance of the regulatory process itself. In short, public-utility regulation must find tools and concepts appropriate to industries that are technologically dynamic and increasingly subject to competitive forces, and that provide incentives for future technological progress.

BIBLIOGRAPHY

Textbooks: I.R. BARNES, *The Economics of Public Utility Regulation* (1942); E.W. CLEMENS, *Economics and Public Utilities* (1950); M.L. FAIR and E.W. WILLIAMS, JR., *Economics of Transportation,* rev. ed. (1959); P.J. GARFIELD and W.F. LOVEJOY, *Public Utility Economics* (1964); M.G. GLAESER, *Public Utilities in American Capitalism* (1957); D.P. LOCKLIN, *Economics* of *Transportation,* 6th ed. (1966); D.F. PEGRUM, *Transportation: Economics and Public Policy,* rev. ed. (1968); C.F. PHILLIPS, JR., *The Economics of Regulation,* rev. ed. (1969); C.E. TROXEL, *Economics of Public Utilities* (1947), *Economics of Transport* (1955).

Economic studies: J.C. BONBRIGHT, *Principles of Public Utility Rates* (1961); R.E. CAVES, *Air Transport and Its Regulators* (1962); J.R. MEYER *et al., The Economics of Competition in the Transportation Industries* (1959); J.C. NELSON, *Railroad Transportation and Public Policy* (1959); W.G. SHEPHERD and T.G. GIES (eds.), *Utility Regulation* (1966); H.M. TREBING (ed.), *Performance Under Regulation* (1968), and ed. with R.H. HOWARD, *Rate of Return Under Regulation* (1969).

Case studies: W.K. JONES, *Cases and Materials on Regulated Industries* (1967); F.X. WELCH, *Cases and Text on Public Utility Regulation,* rev. ed. (1968).

Studies of the regulatory process: M.H. BERNSTEIN, *Regulating Business by Independent Commission* (1955); W.L. CARY, *Politics and the Regulatory Agencies* (1967); H.J. FRIENDLY, *The Federal Administrative Agencies: The Need for Better Definition of Standards* (1962); L.M. KOHLMEIER, JR., *The Regulators: Watchdog Agencies and the Public Interest* (1969).

(C.F.P.)

Publishing, History of

Publishing is the activity that involves selection, preparation, and marketing of printed matter. It has grown from small and ancient beginnings into a vast and complex industry responsible for the dissemination of all manner of cultural material, from the most lofty to the most trivial; its impact upon civilization is impossible to calculate.

The history of publishing is characterized by a close interplay of technical innovation and social change, each promoting the other. Publishing as it is known today depends on a series of three major inventions—writing, paper, and printing—and one crucial social development—the spread of literacy. Before the invention of writing, perhaps by the Sumerians in the 4th millennium BC, information could be spread only by word of mouth, with all the accompanying limitations of place and time. Writing was originally regarded not as a means of disseminating information but as a way to fix religious formulations or to secure codes of law, genealogies, and other socially important matters, which had previously been stored in a succession of human memories. Publishing could begin only after the monopoly of letters, often held by a priestly caste, had been broken, probably in connection with the development of the value of writing in commerce. Scripts of various kinds came to be used over most of the ancient world for proclamations, correspondence, transactions, and records; but book production was confined largely to religious centres of learning, as it would be again later in medieval Europe. Only in Hellenistic Greece, in Rome, and in China, where there were essentially nontheocratic societies, does there seem to have been any publishing in the modern sense—*i.e.*, a copying industry supplying a lay readership.

The invention and original function of writing

The invention of printing transformed the possibilities of the written word but not without opposition from socially conservative forces. Printing seems to have been first invented in China in the 6th century AD in the form of block printing. The Chinese progressed to printing with movable type but abandoned it because it was unsuitable for their language and script. Other Chinese inventions, including paper (AD 105), were passed on to Europe by the Arabs but not, it seems, printing. The reason may well lie in Arab insistence on hand copying of the Qur'ān (Arabic printing of the Qur'ān does not appear to have been officially sanctioned until 1825). The invention of printing in Europe is usually attributed to Johannes Gutenberg in Germany about 1440–50, after a period of block printing from about 1400. Gutenberg's achievement was not a single invention but a whole new craft involving movable metal type, ink, paper, and press. In less than 50 years it had been carried over most of Europe, largely by German printers.

Gutenberg's achievement

Printing in Europe is inseparable from the Renaissance and Reformation. It grew from the climate and needs of the first, and it fought in the battles of the second. It has been at the heart of the whole expanding movement of the past 500 years. Although printing was thought of at first merely as a means of avoiding copying errors, its possibilities for mass-producing written matter soon became evident. In 1498, for instance, 18,000 letters of indulgence were printed at Barcelona. The market for books was still small, but literacy had spread beyond the clergy and had reached the emerging middle classes. The church, the state, the universities, the reformers and radicals were all quick to use the press. Not surprisingly, every kind of attempt was made to control and regulate such a "dangerous" new mode of communication. Freedom of the press was pursued and attacked for the next three centuries; but by the end of the 18th century a large measure of freedom had been won in most countries, and a wide range of printed matter was in circulation. The mechanization of printing in the 19th century and its further great development in the 20th, which went hand in hand with an ever-increasing spread of literacy and ever-rising standards of education, finally brought the printed word to its present powerful position as a means of influencing minds and, hence, societies. But even in the early 1970s, one-half of the world's people could not read or write.

The functions peculiar to the publisher—*i.e.*, the selecting, editing, and designing of the material, arranging its production and distribution, and bearing the financial risk or the responsibility for the whole operation—often merged in the past with those of the author, the printer, or the bookseller. With increasing specialization, however, publishing became, certainly by the 19th century, an increasingly distinct occupation on its own. Most modern publishers purchase printing services in the open market, solicit manuscripts from authors, and distribute their wares to shops for final sale. In some cases, the publisher may sell direct to the final user either by mail or with the help of a sales force. Encyclopaedias, for example, are often distributed in this way.

The great variety of published matter that has grown up over the centuries falls into two main categories, periodical and nonperiodical; *i.e.*, publications that appear at more or less regular intervals and are members of a series and those that [illegible] on single occasions (except for r[illegible]es of essentially the same material).

Of th[illegible] nonperiodical publications, [illegible]k[illegible]te by far the largest class; they are also, in one form or another, by far the oldest of all types of publication and go back to the earliest civilizations. In their main function, of giving permanence to man's thoughts and records of his achievements, they answer a deep human need. Not every published book is of lasting value; but a nation's books, taken as a whole and winnowed out by the passing years, can be said to be its main cultural storehouse. Conquerors or usurpers wishing to destroy a people's heritage have often burned its books, as did Ch'in Shih Huang Ti in China in 213 BC, the Spaniards in Mexico in 1520, and, a more futile gesture, the Nazis in modern times.

There is no wholly satisfactory definition of a book, as the word covers such a variety of publications (some books, such as *The* World Almanac, appear periodically). For statistical purposes, however, the United Nations Educational, Social and Cultural Organization defines it as "a non-periodical printed publication of at least 49 pages excluding covers."

Periodical publications divide further into two main classes, newspapers and magazines. Though the boundary between them is not sharp—there are magazines devoted to news, and many newspapers have magazine features—their differences of format, tempo, and function are sufficiently marked: the newspaper (daily or weekly) usually has large, loose pages, a high degree of immediacy, and highly miscellaneous contents; while the magazine (weekly, monthly, or quarterly) has smaller pages, is usually fastened together and sometimes bound, and is less urgent in tone and more specialized in content. Both sprang up only after the invention of printing, but both have shown a phenomenal rate of growth to meet the demand for quick information and regular entertainment. Newspapers have long been by far the most widely read published matter; the democratizing process of the past 200 years would be unthinkable without them. Magazines, close behind them, historically and in terms of readership, rapidly branched out from their learned origins into "periodicals of amusement." Today there is probably not a single interest, frivolous or serious, of man, woman, or child, that is not catered to by a magazine.

There are, of course, many other types of publications besides books, newspapers, and magazines that may be mentioned, though for various reasons they fall outside the scope of this article and reference to them must be sought under other headings. In many cases the same principles of publishing apply, and it is only the nature of the product and the technicalities of its manufacture that are different. There is, for instance, the important business of map publishing (see MAPS AND MAPPING; HYDROGRAPHIC CHARTING). Another important field is music publishing, of a great variety of material, from complete symphonic scores to sheet music of the latest popular hit. A further range of activities might be grouped under the term "utility publishing"; *i.e.*, the issuing of calendars, diaries, timetables, ready reckoners, guide books, and all manner of informational or directional material, not to mention postcards and greeting cards. A great deal of occasional publishing, of pamphlets and booklets, is done by organizations to further particular aims or to spread particular views; *e.g.*, by churches, religious groups, societies, and political parties. This kind of publishing is sometimes subsidized and not on a purely commercial basis.

This article is divided into the following sections:

I. Book publishing

THE ANCIENT WORLD

The first books. For 4,000 to 5,000 years books have been produced in many different forms. The oldest, the papyrus roll, goes back to about 3000 BC in Egypt, being made from the papyrus plant, which grew freely by the Nile. The roll itself derives from inscribed banners of papyrus that were hung in Egyptian temples; rolled up they became portable inscriptions. The Greeks and Romans also used it, and because the Greeks got their papyrus from the port of Byblos, they called it by that name, from which comes the word bible. To the Romans, a papyrus roll was a *charta* or *volumen.* In Babylonia and Assyria, which had no papyrus, writing was done on clay tablets (from about 3000 BC) in cuneiform script (wedge writing). A "book" then consisted of a number of cuneiform tablets in a labelled container. In China, tablets of bamboo were used up to about 200 BC and then scrolls made from silk waste and, finally, paper (after AD 105). In north India, birch bark was a popular material for manuscripts and in the south, palm leaves, held together by threads. Another material for scrolls in the Middle East was leather; from this, parchment or vellum is said to have been invented by Eumenes II of Pergamum (197–159 BC) when his rival book collector in Egypt cut off supplies of papyrus. In actuality, parchment is known from elsewhere at the same date.

Books began to approach their modern form with the

The codex

codex, for which sheets of papyrus were folded vertically to make leaves. The codex form was first used for Christian literature in the 1st or 2nd century AD, and it may have been devised in order to make up a book large enough to hold more than one gospel or epistle. By the 4th century its use had spread to pagan works, while at about the same time vellum took over from papyrus; for several centuries the vellum codex was then the standard form. The idea of the codex seems to have been suggested by the Roman notebook of waxed wooden tablets fastened with string (pugillares), and these may also have been the origin of the word book, which derives from the Germanic bdc, meaning "beech." Hinged, waxed tablets were also used by the Assyrians in the 8th century BC, however.

The contents of the earliest books, whatever their origin, are almost always religious or semireligious—hymns, prayers, and rites that put man in touch with the divine, and myths, legends, and epics that give an account of his origins. Books with such contents are usually followed by semi-sacred codes of law, collections of proverbs and precepts, and priestly profesaional texts on divination, medicine and magic, history, astronomy and astrology, and other instructional matter. Finally, a purely secular literature is set down, stories, poems, and love songs. All of these categories are well represented in the oldest civilizations, those of Egypt, Babylonia-Assyria, and India. Though there may have been some trade in texts—*e.g.*, of the Egyptian Book of the Dead, copies of which were required at every funeral—books were produced and preserved mainly in libraries attached to palaces and temples, each of the latter specializing in whatever was appropriate to its deity. Babylonia was particularly well endowed; education, including tablet writing, was compulsory for freemen, and some of the surviving tablets bear directions for getting a "book" from the library. Insofar as the contents of early books were made generally known, this was still done largely by word of mouth, by priests and sages, and by storytellers and singers. Many of the earliest works go back, of course, far beyond the invention of writing. Such sacred scriptures as the Indian Vedas, several times the length of the Bible, and such legendary works as the great Indian epic, *Mahābhārata*, were handed down orally for hundreds of years before being recorded in writing.

Libraries of Alexandria and Pergamum

Beginnings of publishing. *Ancient* Greece. The first signs of publishing in the modern sense appear in ancient Greece, but they are signs only; details are lacking. In the pre-classical period, books seem to have been used in the traditional way by reciters, actors, and singers as aids to the memory. By the 5th century BC, however, a small reading public was developing, mocked by Aristophanes as "highbrows." Books were copied by individuals (or their slaves) for private use and there are references in Plato and Eupolis to the possibility of buying them. In the Hellenistic period, after the founding of the great library at Alexandria (3rd century BC) and its rival at Pergamum (2nd century BC), a considerable traffic in books must have grown up. Other libraries, public and private, became common and the reading habit well established. By the 2nd century AD, Lucian was satirizing the ignorance of booksellers and their rich patrons. Throughout this period, except for the local experiment with parchment at Pergamum, the usual form of book was the papyrus roll, about eight or nine inches (20 to 23 centimetres) high (less for poetry or epigrams) and up to 35 feet (11 metres) long (more in Egypt), made by sticking sheets of papyrus together. The writing was arranged in columns according to fairly strict conventions but without much help in the way of punctuation or even word spacing. One 15-foot (five-metre) roll would have accommodated one of the books of Plato's Republic; the "books" into which longer works were divided were merely such rolls made in a convenient size for handling.

Ancient Rome. In ancient Rome, there is little trace of publishing before the 3rd century BC and none of libraries before the 1st century. Liber, the Latin word for "book," means "bark," which suggests that this was once used as writing material; but nothing has survived from the earliest period. After Greek models had been absorbed in its pre-classical period, Latin literature came into full flower, and with it a flourishing book trade grew up. At the time of Cicero (106–43 BC), it was well established and no longer dependent on Alexandria. Its best known figure was Titus Pomponius Atticus (109–32 BC), a rich patron of the arts who was a close friend of Cicero as well as his publisher. Atticus employed a large number of trained slaves to copy and produce books and eventually had retail branches of his business in the provinces as well as in Rome. The form was still the papyrus roll, though the better ones now had rollers, often with decorated knobs. Editions are estimated to have been of from 500 to 1,000 copies, and overproduction was not unknown; unwanted books ended up as wrapping paper. By the 1st century AD, according to Seneca, a library was considered as essential to a house as a bath; and in literature references to professional booksellers become more frequent. A bookshop is described by Martial as having its pillars covered with advertisements of books. Many of the familiar problems of the book trade already appear in Rome: censorship, plagiarism, piracy, and the author's lack of a just reward. There was no law of copyright, but early in the 2nd century AD publishers appear to have taken care of themselves by forming an association, "for the better protection of their interests in literary property."

Chinese inventions

Ancient China. At about the same time an equally vigorous publishing activity was going on in China. The classical period of Chinese literature is usually placed between 600 and 200 BC, the time of the Chinese philosophers Lao-tzu, Confucius, and Mencius, among many others. Although the first Chinese emperor, Ch'in Shih Huang Ti, tried to blot this out in 213 BC by burning books and executing hundreds of scholars, the tradition was rescued under the Han dynasty (206 BC to AD 220) and firmly consolidated by a system of state examinations in the classic texts, through which the civil service was recruited. It was this "literocracy," as it has been called, that provided fertile ground for the rich and humane literary culture of China. By 100 BC, the Imperial library is said to have contained 13,000 volumes. No less remarkable is a series of completely original Chinese inventions in writing technique: the bristle brush, a soft writing material made from silk waste, and, finally, paper, which greatly increased the availability of books (in the form of scrolls). In AD 175, Confucian texts began to be carved into stone tablets from which rubbings were taken. This process seems to have been the very beginning of printing. The invention of lampblack ink came in AD 400 and printing from wooden blocks in the 6th century—for religious pictures at first, then for complete texts. Much of this publishing was done under Imperial or local government direction, but private publishing also throve. A publisher's account from 1176 for a book of 1,300 pages indicates a 3 to 1 ratio of selling price to production costs, a useful profit margin in the absence of modern overhead costs. As one might expect, the oldest printed book in the world is Chinese: a paper scroll from AD 868, 16 feet long and one foot wide (five by three-tenths metres) bearing a Chinese translation of the Indian Diamond Sutra.

Scholarship in monasteries

Developments after the fall of the Roman Empire. With the decline of Rome after the 5th century, publishing in Europe fell into what may be called the primordial pattern: the production and distribution of books was confined largely to religious centres of learning; literacy was maintained henceforth, and spread to a limited degree, by the Christian Church. Elsewhere, this pattern continued to be general. At temples and retreats, Hindu and Buddhist, Zoroastrian, Jewish, and Manichaean texts, and those of other Christian sects and of the Muslims after the 7th century were copied, preserved, and carried abroad by priests, monks, and missionaries. Centres of study sprang up around distinguished scholars or under enlightened rulers; monasteries were founded through saintly inspiration. Though they were subject to decay from within and to destruction from without, these scattered foundations maintained a tenuous but effective web of learning through the centuries. The Venerable Bede (died 735) "published" his great history of the

English people from his cell at Jarrow, and it was copied as far away as Rome. Largely because of St. Benedict (died 543/547), who stressed work as one of the monastic virtues, scriptoria, or writing rooms, came to be regular features of Christian religious houses. In these, some of the most beautiful books in the world were produced, notably in Celtic monasteries, in the form of illuminated manuscripts, besides more everyday editions, not only of devotional and theological works but also of secular literature, especially chronicles. Books were inclined to be large and precious and were lent only under stringent conditions; they were often chained to the desks. An important practical development of the 9th century was the type of script known as Carolingian minuscule, which grew out of Charlemagne's drive to civilize the Franks. Besides other advantages, it regularized the use of small and capital letters and of spaces between words.

Spread of the use of paper. Though paper had been invented by the Chinese as early as AD 105, the Chinese kept the secret of its manufacture jealously. It reached Korea and Japan by about 600 but did not reach the West until much later. The Arabs in 751 captured some Chinese papermakers, extorted their secret, and set up a paper mill at Samarkand. From there, paper travelled via Baghdad, Damascus, and Egypt to Morocco and entered Europe through Spain and Sicily in the 11th to 12th centuries. For books, paper took the place of parchment increasingly between the 13th and 15th centuries. Its great advantage was, of course, its cheapness and abundance (vegetable fibre instead of skin), which reduced the cost of books and so made them more widely available.

The *stationarii*

Secularization of publishing. At the end of the 12th century, with the founding of the first universities, book publishing in Europe began to acquire a broader base. The schools of theology, law, and medicine at Paris, Bologna, and Salerno outgrew their origins and became independent of the church. Other universities soon followed and codices of all kinds were in greater demand. This led to the rise of a new class of book trader, the *stationarii*, or stationers, with fixed premises, as opposed to the itinerant peddlers of books. The latter continued to provide a private, sometimes clandestine, means of distribution, as appears, for instance, from the wide circulation in England, in spite of its proscription, of John Wycliffe's translation of the Bible (*c.* 1382 and 1388). The *stationarii,* however, were under the control of the universities, who authorized them to deal in approved texts only. The main function was to supply to students learned codices, both new (from their own copiers) and secondhand and to act as a kind of circulating library. They also sold writing materials; hence the modern word stationery. The script of the codices during the 12th to 13th centuries underwent a further change: the Carolingian minuscule broke into the Gothic style with its angles and flourishes—more decorative but less clear. As the Middle Ages advanced and monastic life decayed, book production became increasingly secular and, as the Renaissance got under way, more diverse. In the main centres, Paris above all, the scriveners and limners (illuminators) grew sufficiently numerous to form guilds. There was a growing demand for copies of the classics, not to mention all the fresh works—the fables, romances, and poetry—that were beginning to appear more abundantly. These were copied at will; and, because there was no title page as yet, the author often went unrecognized. By the middle of the 15th century, bookcopying in Europe had become quite an industry. A bookseller in Florence, Vespasiano da Bisticci, was employing as many as 50 scribes at a time; and in Haguenau in Alsace, Diebolt Lauber was running something like a book factory for the open market. Block printing had finally reached the west (*c.* 1400) and was being used for illustration and even for whole books. The stage was set for Gutenberg.

SPREAD OF THE ART OF PRINTING: 1450–1550

Before the invention of printing, the number of manuscript books in Europe could be counted in scores of thousands. By 1500, after only 50 years of printing, there were more than 9,000,000 books. These bare figures indicate the explosive impact of the press, the rapidity with which it spread, the great need that had arisen for an artificial script—and the tenuous nature of written culture up to that time.

The *incunabula*

The printed books of this initial period, up to 1500, are known as *incunabula; i.e.,* "swaddling clothes," "cradle," from a Latin phrase used in 1639 to describe the beginnings of typography. The dividing line, however, is artificial. The initial period of printing, a restless, highly competitive free-for-all, runs well into the 16th century. Printing only began to settle down, to become regulated from within and controlled from without, after about 1550. In this first 100 years, the printer naturally dominated the book trade with his new craft. He was often his own typefounder, printer, editor, publisher, and bookseller; only papermaking and, usually, bookbinding were outside his province. Yet even from the beginning, other arrangements cropped up that foreshadowed the future.

Gutenberg Bible

Printing has been called the great German contribution to civilization; in its early days it was known as the German art. After its invention (*c.* 1440–50) by a goldsmith of Mainz, Johannes Gutenberg, it was disseminated with missionary zeal—and a keen commercial sense—largely by Germans and largely along the trade routes of German merchants. Gutenberg himself is usually credited with what is known as the 42-line Bible (1456); the 36-line Bible; and a popular encyclopaedia called the *Catholicon* (1460). His employee and successor, Peter Schoffer, and his son Johann continued in business until 1459; but Mainz itself never became one of the main centres of the book trade. It was soon challenged by Strassburg where, in 1460–61, Johann Mentelin, with an eye for the lay market, brought out a Bible compressed into fewer pages and followed this with the first printed Bible in German or any other vernacular. A few years later, Cologne had its first press (1464) and became an important centre of printing in the northwest. There William Caxton, the first English printer, learned the art in 1471–72. Cologne's early production was almost entirely in Latin because of the heavy bias of its university toward orthodox Thomist theology. In the south, printing quickly spread to the other great trading centres, Basel (1466), Niirnberg (1470), and Augsburg (1472). Basel became famous for the scholarly editions of Johann Amerbach (died 1513) and Johann Froben, who had the benefit of distinguished advisers, including the Dutch Humanist scholar Erasmus. In Augsburg, the first press was set up alongside the renowned scriptorium of the Abbey of SS. Ulrich and Afra; and the tradition of the illuminated manuscript was carried over into equally sumptuous editions of illustrated printed books. At Niirnberg, which soon took the lead in the book trade, Anton Koberger (died 1513) operated on a large, international scale. At his peak, he ran 24 presses and had links with Basel, Strassburg, Lyons, Paris, and many other cities. He could be called the first great businessman publisher and the first publisher to rise socially—to membership of the town council. By 1500 there were presses in some 60 German towns, including Liibeck (1475), the head of the Hanseatic League. From there, printing spread to Denmark, Sweden, Rostock, Danzig, and Russia, though the first printer who went to Russia was apparently murdered before he could achieve anything. Russia began to print in 1552, with the help of a printer from Copenhagen. It could be said that book printing, after its birth in medieval Germany, was carried to maturity in Humanistic Italy.

Italy. The printing press reached Italy very early (1463), via the Benedictine monastery of Subiaco, near Rome, which had strong German connections and a famous scriptorium. Two German printers, Konrad Sweynheim and Arnold Pannartz, who had settled there, soon moved to Rome (1467), where the church encouraged the production of cheap books. In Italy as in Germany, however, it was the great commercial centres that became centres of printing and publishing. By 1500, Venice had no fewer than 150 presses; and two Venetian printers exercised a decisive influence on the form of the

First cheap pocket editions

book: an outstanding typographer, Nicolas Jenson, who perfected the roman type face in 1470, and Aldus Manutius, the greatest printer-publisher of his time. Aldus began printing in 1490 with a series of Greek authors. He then hit on the idea of bringing out cheap "pocket editions" for the new readers produced by the Humanist movement. Beginning in 1501 and continuing with six titles a year for the next five years, he issued a series of Latin texts in octavo format that were models of scholarship and elegance. To keep down the cost, Aldus printed editions of 1,000, instead of the more usual 250; and to fill the page economically, he used an "italic" type designed for him by Francesco Griffo, another innovation of far-reaching effect. The Aldine editions were widely copied, piratically and otherwise, and their dolphin and anchor was the first instance of a publisher's imprint that became a hallmark of excellence. Venice was also important as the source of printing for the south Slavs. About 75 percent of the incunabula in Yugoslavian libraries originated in Venice: printing began in Yugoslavia at Cetinje in 1493.

Patronage of Francis I

France. The way in which printing came to France is of special interest because it shows a publisher in command from the start. In 1470, the rector and librarian of the Sorbonne invited three German printers to set up a press on university premises. The scholars chose the books and supervised the printing, even to specifying the type. By deciding on roman, they greatly helped the eventual defeat of Gothic. Among the early French printers were Jean Dupré, a businessman publisher of ***editions de luxe*** ("luxury editions"), who set up in 1481; and Antoine Vérard, who began printing in 1485. Vérard was the first to print the Book of Hours, a book containing the prayers or offices appointed to be said at canonical hours, and set a standard for French elegance. After 1500, when the Renaissance had begun to be felt in France, a brilliant group of scholarly printers, including Josse Bade, Geoffrop Tory, and the Estienne (Stephanus) family, who published without a break for five generations (1502–1674), carried France into the lead in European book production and consolidated the Aldine type of book—compact, cheap, and printed in roman and italic types. The golden age of French typography is usually placed in the reign of Francis I (1515–47), one of the few monarchs ever to take a keen personal interest in printing. He was the patron and friend of Robert Estienne. In 1538 he ordered Estienne to give a copy of every Greek book he printed to the royal library, thus founding the first copyright library. In 1539 he laid down a code for printers, which included a prohibition on the use of any imprint that could be confused with another. Outside Paris, the only significant centre of printing in France was Lyons, which had close links with Basel. While Paris was under the watchful eye of the Sorbonne (University of Paris) theologians, Lyons was able to publish Humanist and Protestant works more freely. Among its foremost printers were Johann Trechsel and his sons, Melchior and Caspar; Sebastian Greyff of Gryphius; and a fine typographer, Robert Granjon. By about 1600, however, the pressure of the Inquisition and the competition of Paris had put an end to printing in Lyons. Thereafter, the French book trade was centred entirely in Paris.

Complutensian Polyglot Bible

Other European printers. The rest of Europe had presses almost as quickly; *e.g.*, Utrecht (1470), Budapest (1473), and Cracow (1474), in each case through Germans. Spain in particular shows the direct connection. There the first press was set up in 1473 at Valencia, where the German trading company of Ravensburg had an important base. Though Madrid eventually became dominant (after 1566), publishing flourished in the early period at Barcelona, Burgos, Zaragoza, Seville, and the university towns of Salamanca and Alcalá. In Lisbon, the first printed book was apparently produced in 1489 by a Jew; he was reinforced in 1495 by two printers sent for by the Queen of Portugal. Spain quickly evolved its own distinctive style of book, full of dignity and printed largely in Gothic types. The most remarkable production of the period was the magnificent Complutensian Polyglot Bible, sponsored by Cardinal Ximenes "to revive the hitherto dormant study of the scriptures," which it effectively did. It was printed at Alcalá, in Hebrew, Chaldee, Syriac, Greek, and Latin, by Arnaldo Guillermo de Brocar, the first great Spanish printer. Editorial work was begun in 1502, the six volumes were printed in 1514–17, and the book finally issued in 1522. From Spain, printing crossed the Atlantic during this early period. Juan Cromberger of Seville, whose father, Jacob, had set up a press there in 1502, secured the privilege for printing in Mexico and sent over one of his men, Juan Pablos. In 1539, Pablos published the first printed book in the New World, ***Doctrina christiana en la lengua mexicana e castellana*** ("Christian Doctrine in the Mexican and Castilian Language").

England. Comvared with the Continent. England in the early days of printing was in a state of backwardness, intellectually and economically, partly because of the Wars of the Roses in which the Lancaster and York families fought for the throne. Printing reached England fairly late (1476), and in 1500 there were still only five printers in England, all in London and all foreigners. Type seems to have been largely imported from the Continent until about 1567, and paper until about 1589 (except for a brief spell 1495–98). In an Act of 1484 to restrict aliens trading in England, Richard III deliberately exempted all aliens connected with the book trade, in order to encourage its domestic development. In the following year, Henry VII appointed a foreigner, Peter Actors of Savoy, as royal stationer, with complete freedom to import books. For about 40 years, England was a profitable field for Continental printers and their agents. This necessary free trade was brought to an end and native stationers protected under Henry VIII, whose acts of 1523, 1529 and 1534 regulated foreign craftsmen and finally prohibited the free importation of books. It has been estimated that up to 1535 two-thirds of those employed in the book trade in England were foreigners.

William Caxton and his press

It is all the more remarkable that England's first printer, William Caxton, was in fact an Englishman and the only native to introduce printing to his country. After learning to print at Cologne (1471–72), Caxton set up a press at Bruges (1474) and printed his first book, ***Recuyell of the Historyes of Troye,*** which was his own translation from the French and probably the main reason why this semi-retired merchant gentleman took to printing at the age of 50. He had returned to England through the encouragement of Edward IV and continued to receive royal patronage under Richard III and Henry VII. Caxton is important not so much as a printer (he was not a very good one) but because from the first he published in English instead of Latin and so helped to shape the language at a time when it was still in flux. Of the 90 odd books he printed, 74 were in English, of which 22 were his own translations. Some, such as the ***Ordre of Chyvalry*** and the ***Fayttes of Armes,*** were for the pleasure of his royal patrons; but his range was wide and included ***Dicts or Sayings of the Philosophres*** (1477; his first book in England); two editions of Chaucer's ***Canterbury Tales*** (the second because a better manuscript came to hand); ***The Fables of Aesop*** (in his own translation from the French); Sir Thomas Malory's ***Kyng Arthur;*** and his largest work, ***The Golden Legend,*** a compilation of such ecclesiastical lore as lives of the saints, homilies, and commentaries on church services, a considerable editorial labour apart from the printing.

Caxton's press was carried on after his death by his assistant, Wynkyn de Worde of Alsace. In the absence of court connections and also because he was a shrewd businessman, he relied more on bread-and-butter lines, liturgies and school books (educational reform was just beginning). He published some 800 titles, mostly small volumes for the ordinary citizen, and continued Caxton's standardizing of the language, a solid contribution to the native book trade. His contemporary, the best of the early printers, was Richard Pynson of Normandy, who began printing in 1492 and became printer to the king in 1508. Pynson, the first to use roman type in England (1509), published the first English book on arithmetic (1522). After his early liturgies and some fine illustrated books, he concentrated mainly on legal works. In 1521 he published Henry VIII's answer to Luther in defense of the papacy, for which the King received the title of *fidei defensor* ("defender of the faith) from the Pope.

The book trade. The book trade during this early period showed enormous vitality and variety. Competition was fierce and unscrupulous. A printer of Parma in 1473, apologizing for careless work, explained that others were bringing out the same text, and so he had to rush it through the press "more quickly than asparagus could be cooked." Though most of the early firms were small printer-publishers, many different arrangements were made and there was at least one individual, Johann Rynmann of Augsburg (died 1522), who published nearly 200 books but printed none of them. Publishing companies, which both financed and guided the printing enterprise, were also tried, as at Milan in 1472 and at Perugia in 1475. Publishers were not slow to develop their publicity. The medieval scribes had placed their names, the date when they finished their labours, and perhaps a prayer or a note on the book, at the end of their codices. From this grew the printer's colophon or tailpieces, which gave, besides the title of the book, the date and place of printing, the name of the printer, his device, and a bit of self-advertisement, often composed by a professional writer. By about 1480, the information of the colophon began to appear at the front of the book as a title page, along with the title itself and the name of the author. Advertisements for books, in the form of handbills or broadsheets, are known from about 1466 onward, including Caxton's of 1477, ending with a polite request not to tear it down, *Supplico stet cedula* ("Please let the poster stand"). Publisher's lists and catalogues occur almost as early. Distribution of books along the trade routes, with their courier services, appears to have been highly effective. In 1467, for instance, a bookseller in Riga had a stock of books issued by Schoffer in Mainz. There were also the regular trade fairs, especially those at Frankfurt, where a major annual book fair is still held today, and at Stourbridge in England, which was associated with the book trade up to about 1725. Besides the stationers, who may sometimes have functioned as wholesalers, there were also retailers known as "book-carriers."

Printing's effect on languages

Early printing had a profound effect on national languages and literatures. It began at once to create, standardize, and preserve them. Caxton, in the preface to his translation of the *Aeneid,* after telling a story of confused dialects, ended up "Lo! what should a man in these days now write, eggs or eyren?" By choosing words "understood of common people" and by printing all he could of English literature, he steered the English language along its main line of development. The early printing of great vernacular works, such as those of Dante, Petrarch, and Boccaccio in Italy, or the publishing of a vernacular Bible, such as that of Luther in Germany, gave many languages their standard form. The French language owes much to an early printer-publisher, the scholarly Robert Estienne, who is known as the father of French lexicography for his dictionaries as well as for his typographical innovations in the 1530s. Up to 1500, about three-quarters of all printing was in Latin, but thereafter it steadily declined as books appeared in the vernaculars and reached an ever-widening public.

Controls over printing. The church at first had every reason to welcome printing. Bibles (preferably in Latin), missals, breviaries, and ecclesiastical literature poured from the early presses of Europe; and the first printed best seller was a devotional work of universal acceptance, Thomas à Kempis's *De imitatione Christi,* which went through 99 editions between 1471 and 1500. Such sales were matched, however, between 1500 and 1520 by the works of a Humanist, Desiderius Erasmus, and, after 1517, by those of a heretic, Martin Luther. The church had always exercised a censorship over written matter, especially through the universities in the latter part of the Middle Ages. As the works of the reformers swelled in volume and tone, this censorship became increasingly harsh. The Inquisition was restored, and it was decreed in 1543 that no book might be printed or sold without permission from the church. Lists of banned books were drawn up, and the first general *Index Librorum Prohibitorum* was issued in 1559. Dutch printers in particular suffered under the Inquisition and a number went to the stake for publishing Protestant books. To avoid such a fate and to beat the censor, some resorted to the fake imprint, putting a fictitious printer or place of publication (or both, or neither) on the title page.

The Index of Forbidden Books

Censorship also began to be exercised in varying degrees by individual rulers, especially in England, where church and state had been united under Henry VIII after his defection from Rome. The Tudors, with little right under common law, arrogated to themselves authority to control the press. After about 1525, endless proclamations were issued against heretical or seditious books. The most important was that of 1538 against "naughty printed books," which made it necessary to secure a license from the Privy Council or other royal nominees for the printing or distribution of any book in English.

The Stationers' Company

In this attempt at control, an increasingly prominent part came to be played by the Stationers' Company. Since its formation in 1403 from the old fraternities of scriveners, limners, bookbinders, and stationers, it had sought to protect its members and regulate competition. Its first application for a royal charter in 1542 seems to have gone unheeded; but in 1557, an important date in the English book trade, the interests of the crown (then the Catholic Mary Tudor), which wanted a ready instrument of control, coincided with those of the company (under a Catholic first Master), and it was granted a charter that gave it a virtual monopoly. Thereafter, only those who were members of the company or who otherwise had special privileges or patents might print matter for sale in the kingdom. Under the system of royal privileges begun by Henry VIII, a printer was sometimes given the sole right to print and sell a particular book or class of books for a specified number of years, to enable him to recoup his outlay. This type of regulation now came into the hands of the Stationers' Company. After licensing by the authorities, all books had to be entered in the company's register, on payment of a small fee. The first stationer to enter a book acquired a right to the title or "copy" of it, which could then be transferred as might any other property. As the beginning of a system of copyright, this procedure was an admirable development; but the grip that the company obtained and its self-interested subservience to authority were to stunt the free growth of the English book trade for the next 100 years.

THE FLOURISHING BOOK TRADE: 1550-1800

During this period, there were virtually no technical changes in methods of book production, but the organization of the trade moved gradually toward its modern form. The key functions of publishing, selecting the material and bearing the risk, shifted from the printer to the bookseller and from him to the publisher in his own right; the author, too, at last came into his own. The battle with the censor became increasingly fierce before enlightenment triumphed, at least after a fashion. Literacy grew steadily and the book trade expanded, both within and beyond national boundaries. All this took place in outline by about 1800, when the mechanical inventions of the 19th century, among other factors, began to lift the trade to fresh levels.

The Netherlands. In Europe between 1550 and 1700, the lead in book publishing passed for a time to the Netherlands. The business founded at Antwerp in 1549 by Christophe Plantin, a Frenchman by birth, came to dominate the Catholic south of the country, both in quantity and in quality. Its finest production was probably the eight-volume Polyglot Bible (1568–73), the *Biblia regia.* The firm was carried on for generations by the descendants of Plantin's son-in-law, Balthasar Moretus, and today its premises are a museum of printing. In the Protestant north, the house of Elzevir occupied a similar position. After its founding by Louis Elzevir, who issued his first book in 1593, it was extended by succeeding generations to The Hague, Utrecht, and Amsterdam, with varying fortunes. A duodecimo (small-format) series of classi-

The Polyglot Bible

cal texts that the Elzevirs began issuing in 1629 more than matched the earlier Aldine editions in excellence at a reasonable standard price. The Dutch, as great seafarers, were pre-eminent publishers of atlases, a word that was first used when the maps of Gerardus Mercator were published by his son, Rumold, in 1595. The high skill of Dutch engravers also went into their emblem books (books of symbolic pictures with accompanying verse), for which there was a considerable demand between 1580 and 1650.

France. In France, as the monarchy reasserted itself after the wars of religion, publishing became increasingly centralized. In 1620, Louis XIII set up a private press in the Louvre, the Imprimerie Royale, which the Cardinal de Richelieu turned into a state establishment in 1640. Since then, this national press, through all its changes of name through changes of regime, has provided a continuous standard of excellence to which book production in France could refer. Louis XIII also tried to regulate the trade. By an ordinance of 1618, a body was established similar to the English Stationers' Company, the Chambre des Syndicats; because it contained two royal nominees, however, its control was even more absolute than that in England. Censorship, though it remained for a time with the Sorbonne, also passed eventually to officials of the crown. Under these cramping conditions, publishers were inclined to play safe; as elsewhere, more controversial works first appeared outside the country (often in Holland or Geneva) or under a false imprint. But French books fully upheld the influence of French taste in Europe. A remarkable publishing feat of the 18th century was the first collected edition of Voltaire's works (1785–89). It was produced at Kehl in Sud-Baden by Pierre-Augustin-Caron de Beaumarchais, the author of ***The Barber of*** *Seville* and ***The Marriage of*** *Figaro*, who bought the copyrights from Panckoucke in Paris and the printing equipment (especially for the purpose) from the widow of the great English typographer John Baskerville.

Publishing the works of Voltaire

Germany. After the Reformation, the intellectual life of Germany was predominantly Protestant and the book trade almost entirely so. Through its book fairs, Frankfurt had become the centre of German publishing and even a kind of European clearinghouse. In 1579, however, the fair came under the supervision of the imperial censorship commission (Frankfurt being a free imperial city), and this action gradually killed it. After about 1650, though Frankfurt continued to be important for the production of type and illustrated books, the centre of the trade shifted decisively to Leipzig. There, an enlightened government and a celebrated university favoured cultural life and books not least. Two Leipzig firms dating from the 17th century have come down to the present day: that founded by Johann Friedrich Gleditsch in 1694, which was taken over by F.A. Brockhaus in 1830, and that founded by Moritz Georg Weidmann in 1682. A Weidmann partner, Philipp Erasmus Reich, was known in the 18th century as "the prince of the German book trade." He could be said to have invented the net price principle (see below) and the idea of a booksellers' association (1765), which in 1825 became the Borsenverein der Deutschen Buchhandler, a unique organization of publishers, wholesalers, and retailers. At the culmination of the classical period of German literature toward the end of the 18th century, three publishers were outstanding—Georg Joachim Göschen in Leipzig; Johann Friedrich Cotta in Tiibingen and Stuttgart; and Johann Friedrich Unger in Berlin, all of whom had a share in publishing Schiller and Goethe. Unger also published the magnificent translation of Shakespeare by A.W. Schlegel (8 vols., 1797–1810).

England. In the golden age of Elizabeth I, publishing in England was probably at its most turbulent. Through her Injunctions of 1559, Elizabeth confirmed the charter of the Stationers' Company and the system of licensing by the crown or its nominees, which now included church dignitaries. Controls were tightened in 1586 by a decree of Star Chamber, which confined printing to London, except for one press each in the universities of Oxford and Cambridge. The Stationers' Company was given powers to inspect printing offices and to seize and destroy offending material or presses, which it zealously did, as much in defense of its monopoly as in support of the crown. But despite stern measures, including execution, the great religious question, in which Elizabeth steered a precarious course between Papists and Puritans, continued to be fought out with secret presses on both sides.

Within the legitimate trade, the booksellers had begun to get the upper hand. The incorporation of the Stationers' Company, like that of other London companies, was in itself an indication of the ascendancy of the trader over the craft's man. Under Elizabeth, as part of a developing system of monopolies, the former short-term privileges for particular works or classes of works were granted, for a consideration, as life patents with rights of reversion, such as those enjoyed by Richard Tottel for law books or John Day for ABCs and catechisms. The printers had already been driven by their costs to make arrangements with the booksellers, to their own disadvantage. Gradually, the best "copies" came into the hands of a rich few, who ruled the company and who, in the words of a report of 1582, "keepe no printing howse, neither beare any charge of letter, or other furniture but onlie paye for the workmanship." In 1577, an abortive revolt was led by John Wolfe, a former apprentice of John Day, who maintained his right to print what he pleased. He was twice imprisoned but was finally bought off by admission to the company. In 1584, to still the discontent, some of the rich patentees surrendered a number of "copies" to the company for the benefit of its poorer members. These were added to in 1603, when James I withdrew some patents from individuals and sold them to the company, again for "the poore of the same." In this way the company itself became a publishing organization; and having tasted the advantages, it bought up more and more "copies" on its own account. These came to be divided into "stocks," the English Stock, Bible Stock, Irish Stock, Latin Stock, and Ballad Stock, with shares allocated among its members. By 1640, through leasing the patents at its discretion, the company controlled most of the printing offices in London. The benefit to the poor stationers was somewhat marginal and the monopoly and lack of foreign stimulus caused England to lag behind the Continent in standards of production.

For all that, the privileged men did some good publishing; a few even supported their authors during their labours. Some landmarks of the period were John Lyly's ***Euphues,*** published by Gabriel Cawood (1578); Sir Thomas North's translation of Plutarch's ***Lives,*** so important for Shakespeare, by Thomas Vautroullier (1579); and Spenser's ***Faerie Queene,*** by William Ponsonbie (1589–96), not to mention the Authorised Version of the Bible (1611), which was completed in a room at Stationers' Hall and printed at the expense of Robert Barker, the king's printer. The city, however, had no love for the theatre, so the wealthy publishers missed the real glory of the age, its drama. Publication of drama was left, along with much of the poetry and the popular literature, to the unprivileged and to the outright pirates, who scrambled for what they could get and but for whom much of it would never have been printed. To join this fringe, the would-be publisher had only to get hold of a manuscript, by fair means or foul, enter it as his "copy" (or dispense with the formality), and have it printed. Just such a man was Thomas Thorpe, the publisher of Shakespeare's sonnets (1609); the mysterious "Mr. W.H." in the dedication is thought by many to be the person who procured him his "copy." The first Shakespeare play to be published (*Titus Andronicus*, 1594) was printed by a notorious pirate, John Danter, who also brought out, anonymously, a defective ***Romeo and Juliet*** (1597), largely from shorthand notes made during performance. Eighteen of the plays appeared in "good" and "bad" quartos before the great First Folio in 1623. A typical imprint of the time. of the "good" second quarto of ***Hamlet*** (1604), reads: "Printed by I.R. for N.L. and are to be sold at his shoppe under Saint Dunston's Church in Fleetstreet"; *i.e.*, printed by James Roberts for Nicholas Ling. For the First Folio, a large undertaking of over 900 pages, a syndicate of five was formed,

Printing of Shakespeare's sonnets

headed by Edward Blount and William Jaggard; the Folio was printed, none too well, by William's son, Isaac.

The struggle to govern and limit the press continued through most of the 17th century. In 1637, at the height of episcopal tyranny and Puritan pamphleteering in Britain, Star Chamber issued its most drastic decree, which confirmed previous enactments, laid down detailed licensing procedures, reduced the total number of printers to 23, and prescribed severe penalties for offenses. Four years later, however, Star Chamber itself was swept away by Parliament, and in the ensuing uncertainty the trade had a taste of freedom. This new situation quickly alarmed not only the Stationers' Company, who saw their privileges vanishing, but also Parliament, which proved to be as reactionary as the royalists. In 1643, it passed an ordinance restoring both licensing and the powers of the company. It was this Act that prompted Milton to write his *Areopagitica,* a noble and powerful plea for freedom of the press, which demolishes every argument for censorship. After the Restoration, the Licensing Act of 1662 was ruthlessly enforced until after the Plague (1665), when its rigours were mitigated; it lapsed in 1679. James II revived licensing in 1685, but Parliament refused to renew it in 1694. Thereafter, restraint, harassment, and persecution continued, but by other means, under a broad interpretation of the meaning of libel. With the end of licensing and the gradual breakdown of the whole guild system, the Stationers' Company declined in importance; but it remained useful in connection with copyright.

In the latter part of the 17th century, the trade gathered momentum rapidly, partly through the rise of the periodical press (see below), with its growing body of writers and readers. For successful books the rewards became considerable and the author's right to a proper share more widely recognized. Dryden, according to Pope, received a total of £1,200 for his *Virgil* (1697), at a time when a shopkeeper might receive £50 a year and a labourer £15. Patronage continued, with all its political implications; but dedications became increasingly cut-and-dried, costing five guineas for a poem, perhaps, or 20 for a play; royalty was naturally expected to pay more. By the 1750s it was virtually at an end; "We have done with patronage," said Dr. Johnson. In its place came the public at large, to whom Henry Fielding dedicated his *Historical Register* in 1737. In the expanding literary market, the enterprising publisher no longer built up monopolies in bread-and-butter lines (though he has never despised them) but tried instead to collect all the most promising authors. Through his personal inclinations, his sense of public taste, and his readiness to risk novelty, he began to play a part of his own in the course of literary development. As this side of the business absorbed more and more of his energies, the final separation of publisher and bookseller came about, though never so decisively as that between bookseller and printer. The bookseller, for his part, also had his hands full as the retail trade became better organized. The first advertisement claiming that his books were "available in every bookshop" seems to have been made by a Leipzig publisher in 1717.

Payment to authors

Copyright Act of 1709

In England this whole transition was marked—and fostered—by the passing of the Copyright Act of 1709, the first of its kind in any country. It was "An Act for the encouragement of Learning, by vesting of the copies of printed books in the authors or purchasers of such copies during the times therein mentioned." For books already printed, the time was 21 years, "and no longer" (from April 10, 1710, when the Act came into force). For unpublished works, it was 14 years, "and no longer," though if the author was still living at the end of that time, the copyright returned to him for a further period of 14 years. Penalties were laid down, and registration at Stationers' Hall was made a condition for their enforcement.

The Copyright Act of 1709, like all subsequent measures, tried to strike a balance between the needs of those who make a living from books—writers, printers, and publishers—and the interests of the reading public, which are far from identical; it tried, in other words, to limit privilege as well as piracy. The terms it set have since come to be regarded as too short; but in setting any term at all, and in focussing attention on the author as prime producer, it was revolutionary.

The fathers of modern publishing in England, which may be said to date from this time, were Jacob Tonson (died 1736), who acquired the copyright of *Paradise Losi* and published for Dryden, Addison, Steele, and Pope, among others; and Barnaby Bernard Lintot, who also published Pope, paying him some £5,300 in all for his *Iliad.* Two modern British firms go back to this period, Rivingtons and Longmans, Green & Co., Charles Rivington starting in 1711 and Thomas Longman in 1724, by buying the business of William Taylor, the publisher of *Robinson Crusoe.* At mid-century, the best known figure in the trade, and one of the best liked, was Robert Dodsley (1703–64), the footman-poet who was befriended by Pope. Among "his" authors were Pope himself, Goldsmith, Sterne, and Johnson. He is credited with suggesting the idea of the *Dictionary* to Dr. Johnson, and his name heads the list of "gentlemen partners" who financed it. Such co-operative associations were popular for longer works. They were known as congers and developed into a system of shares in individual books which could be bought and sold at will.

United States. During the 18th century, the book trade in the United States began to flourish. Printing had begun there in 1638, when the first printers, Stephen Daye and his two sons (Joseph Glover, who arranged the venture, died on the voyage), went out from Cambridge, England, to Cambridge, Massachusetts. After printing *The Oath of a Free-Mart* (1638) and *An Almanack for the Year of Our Lord 1639,* they produced their first book, *The Whole Booke of Psalmes,* in 1640. Cambridge, Massachusetts, had the sole privilege of printing, but the monopoly was broken in 1674, when Marmaduke Johnson, who had come over to print an Indian Bible (1663), moved his press to Boston. Gradually others followed; Philadelphia had a press in 1685; New York, in 1693. It was difficult for the colonial printer, as for any small printer, to produce large works because of shortage of type; but his patronage by the government helped to give his products a dignified style. Almanacs, primers, and law books were the staples; theology, as ever, was the leading category. Up to 1769, America bought its presses from England but thereafter became independent, as also for ink and paper. Books were sold in various ways—by subscription, by the printer himself, by hawkers, and through shopkeepers. Though Massachusetts passed a law against hawkers in 1713, it carefully excluded book peddlers, who had a valuable function in rural areas. The first bookseller seems to have been Hezekiah Usher of Boston, *c.* 1647, who added books to his general merchandise.

Spread of education and literacy. The great increase in available reading matter after about 1650 resulted from the spread of education to the middle classes, especially to women, with a fringe of literacy extending a lot further. The grammar school foundations in the wake of the Renaissance had been followed by a more practical approach to the needs of a child, stemming from Comenius (1592–1671), who produced the first instructional picture book, *Orbis Sensualium Pictus* (1658). Compulsory education seems to have begun in the little dukedom of Weimar in 1619, followed by Prussia in 1717, and France after the Revolution. England was comparatively late, 1870; but charity schools organized by various societies and religious bodies, such as the Society for Promoting Christian Knowledge (1698), and the Sunday School movement, started by Robert Raikes (1780), helped to compensate. The wider readership is reflected among the middle classes by the rich development of the prose novel in the 18th century and, among the less well-to-do, by the great sale for almanacs and chapbooks. The almanacs, such as Benjamin Franklin's *Poor Richard's Almanack* (Philadelphia, 1732–64), usually consisted of miscellaneous information and homilitic matter (religious and moral sayings), while the chapbooks, a few pages cheaply produced, contained a popular story or ballad illustrated by a crude woodcut; a well-known one is *The famous and remarkable History of Sir Richard Whittington, three times Lord Mayor of London,* 1656.

National copyright libraries

Growth of libraries. Growth in the book trade led naturally to growth in the scale, number, and types of libraries. Some of the oldest collections of books developed into national "copyright libraries," of immense value for bibliographical purposes. Sir Thomas Bodley opened his famous library at Oxford in 1602, and in 1610 the Stationers' Company undertook to give it a copy of every book printed in England. Later, Acts of Parliament required the delivery of copies of every book to a varying number of libraries; at present it stands at six, the most important being the library of the British Museum, founded in 1759. This idea of a definitive collection has been adopted elsewhere; *e.g.*, in the United States, where the Librarian of the Library of Congress (founded in 1800) was appointed copyright officer in 1870.

In the 18th century, a characteristic development was the commercial lending library, and in the 19th the free public libiary. Despite the fears of publishers and booksellers, circulating libraries have promoted rather than diminished the sale of books, besides being a steady market in themselves.

Decline of censorship. The general spread of more rational notions in the 18th century led to the waning of censorship in most western countries. It was abolished in Sweden in 1766, in Denmark in 1770, and in Germany in 1848. The clearest statement, to which lip service, at least, is now almost universally paid, came from the French National Assembly in 1789: "The free communication of thought and opinion is one of the most precious rights of man; every citizen may therefore speak, write and print freely." In the United States, no formal censorship was ever established; control over printed matter has always been exercised through the courts under the law of libel. This was also the case in England after the lapsing of the Licensing Act in 1694; but two important steps had yet to be taken: in 1766, Parliament put an end to general warrants; *i.e.*, for the arrest of unnamed persons and for the seizure of unspecified papers; and in 1792, Charles Fox's Libel Act finally gave the jury the right to decide the whole issue, which had previously depended mainly on the judge. In both of these fundamental reforms, a fighting part was played by John Wilkes (1725–97), Member of Parliament and writer. The removal of censorship has never, of course, prevented persecution; but the issues at any rate become more open.

MODERN PUBLISHING: FROM THE 19TH CENTURY TO THE PRESENT

Revolutionary inventions

The 19th century. In the 19th century, a whole new era began. A series of technical developments, in the book trade as in other industries, dramatically raised output and lowered costs. Stereotyping, the iron press, the application of steam power, mechanical rypecasting and typesetting, new methods of reproducing illustrations—these inventions, spaced out through the century and often resisted by the printer, amounted to a revolution in book production. Paper, handmade up to 1800, formed over 20 percent of the cost of a book in 1740; by 1910, it had fallen to a little over 7 percent. Bindings, too, became less expensive. After 1820, cloth cases began to be used in place of leather, and increasingly the publisher issued his books ready bound. Previously, he had done so only with cheaper books; the bindings of others had been left to the bookseller or private buyer. Improved means of communication led to wider distribution, and railway travel itself became a spur to reading. For Europe and America, expansion and competition were the essence of the century, and the book trade had a full share of both. While the population of Europe doubled, that of the United States increased fifteenfold. A thirst for improvement and entertainment greatly expanded readership, leading to a rapid growth in every category of book from the scholarly to the juvenile. The interplay of technical innovation and social change was never more close. The only victim in the book trade was design, part of the price that was paid almost universally in the first phase of machine production.

Publishing was now well established, with its characteristic blend of commerce and idealism. Though some older methods lingered, European publishing houses were, in function, what they are today. The stability of the pattern is reflected in the large number of firms that are still in business after more than 150 years. Some early English foundations have already been mentioned. Among others are: Thomas Nelson & Sons, Ltd. (1798), William Blackwood & Sons, Ltd. (1804), A. & C. Black Ltd. (1807), Blackie & Sons, Ltd. (1809), William Collins, Sons & Co., Ltd. (1819), and W. & R. Chambers, Ltd. (1820). For Germany, one might name Vandenhoeck & Ruprecht (1735), C.H. Beck (1763), and F.A. Brockhaus (1805); and for France, slightly later, Garnier (1833) and Plon (1854). Their tendency to specialize made French and German publishers more vulnerable to change than their English colleagues, who aimed as a rule at greater flexibility. Literary and intellectual currents were flowing strongly and the number of new books rose by leaps and bounds. Rough figures for England indicate 100 new titles per year up to about 1750, rising to 600 by 1825, and to 6,000 before the end of the century. Equally characteristic was the appearance of popular series at low prices, "literature for the millions," as Archibald Constable (1774–1827) was the first to call it. The forerunner was the publisher John Bell's ***The Poets of Great Britain*** (rivalling Dr. Johnson's), which appeared in 1777–83, in 109 volumes at six shillings each, when even a slim volume usually cost a guinea or more. By the 1850s, the application of the new techniques of mass production had brought down the price of a cheap reprint to one shilling, as in the Railway Library of novels (George Routledge, 1,300 volumes, 1848–98), for instance, or in the three series of classics issued by H.G. Bohn in 1846, 1850, and 1853. Later reprints were cheaper still. Cheapest of all was Cassell's National Library (209 vols., 1886–90), bound in paper for three pence and in cloth for six pence, that is, a twelfth the price of the Bell set. On the Continent, two German series were outstanding. The Tauchnitz Collection of British and American Authors (1841–1939) became known to thousands of travellers. Tauchnitz voluntarily paid royalties and forbade the sale of his editions in Britain. Even more successful was Reclam's Universal-Bibliothek, begun in 1867 and still going strong, with total sales running into hundreds of millions. An important factor in this series, as in others later, was the release of works through the expiration of copyright.

The low-priced popular series in England

Book piracy. In the United States, publishing gradually became centralized in a few cities—Philadelphia, Boston, and New York. As in Europe, several modern firms go back to the early 19th century and even further: J. B. Lippincott Co. (1792), John Wiley & Sons, Inc. (1807), Harper & Row (1817), G.P. Putnam's Sons (the first Putnam, 1840), and Charles Scribner's Sons (1846). Although American literature put down strong roots during this century, piracy from England rose to great heights. There was sharp competition to be the first to secure proofs of any important new book. Publishers would send to the dockside and produce an American edition almost within hours, as they did in 1823 with Sir Walter Scott's ***Peveril of the Peak.*** In the absence of international copyright agreements, the English author usually received nothing, but there were honorable exceptions; Harper Brothers, for instance, paid considerable royalties to Charles Dickens and Thomas Macaulay, among others. There was also at least one famous case of piracy in reverse. When Harriet Beecher Stowe's slavery novel ***Uncle Tom's Cabin*** came out in the United States in 1852, 1,500,000 copies rapidly appeared in England, some editions selling for sixpence. Though it can be argued with some justice that piracy is not only inevitable but possibly even desirable for the sake of cultural diffusion in some circumstances, the availability of cheap foreign books, if overprolonged as it almost certainly was in the United States, can damage the prospects for home-produced literature. Though there were some household names, such as Washington Irving, James Fenimore Cooper, Ralph Waldo Emerson, and Henry Wadsworth Longfellow, U.S. writers in general had a lean time; and the strong development of the magazine short story and the lecture tour in the United States has been attributed in part to their difficulties. Toward the end of the century U.S. publishing was further enriched by translations of many foreign works, as a result of the flood of immigrants into New York.

Price regulation. While 19th-century publishing was competitive and individualistic, its growing volume pointed increasingly to the need for greater organization. A major problem, once booksellers had become distinct from publishers, was suicidal price-cutting in the retail trade. Though price regulation ran counter to accepted notions of free competition and met with fierce opposition, in the general interest of the industry it was inevitable. Like copyright, it helped to provide a firm structure within which fair prices could be calculated. The "net price" principle, first raised in the previous century by the German publisher Reich (see above), was adopted in Germany in 1887 through the work of the Börsenverein, the trade organization founded in 1825. Under this principle, the publisher allows trade discount to the bookseller only on condition that the book is sold to the public at not less than its "net published price" as fixed by the publisher. In England a first attempt to introduce it by the booksellers in the 1850s was condemned to failure by the Free Traders; but toward the end of the century some publishers, led by Macmillan, began to replace the variable discounts by fixed prices. To press for the new system, the Associated Booksellers of Great Britain and Ireland was formed in 1895, and the Publishers Association was created in 1896. These two organizations then worked out the Net Book Agreement (1901), primarily through the efforts of Frederick (later Sir Frederick) Macmillan. The principle has since been generally adopted, although only to a limited extent in the United States. At roughly the same time, the founding of the Society of Authors (1884) in England and the Authors' League (1912) in the United States helped to standardize fair dealing over contracts and the payment of royalties to authors.

Trade catalogues. The trade also became better organized in the provision of comprehensive catalogues of current books. These began as early as the twice-yearly book fairs at Frankfurt (first catalogue 1564) and Leipzig (first catalogue 1594). So great was the value of the Frankfurt catalogue that an English edition was published in 1617–28. Eventually, all such semiprivate ventures, as ***A Catalogue of all the Books Printed in the United States*** (1804) or English catalogues deriving from ***The Publishers' Circular*** (1837) or ***Whitaker's*** **(1874),** became national lists, such as the ***Biblionravhie de la France*** (1811 ff.), the U.S. ***Cumulative Book List*** (1898 ff.), the ***Deutsche Nationalbibliographie*** (1931 ff.) and the ***British National Bibliography*** **(1950 ff.).**

Development of copyright law. Copyright, too, underwent considerable development. By the end of the century, most countries had some provision, and various terms of protection were tried, running from publication or from the date of the author's death. The United States first legislated in 1790, France in 1793, and Germany in 1839. Moves toward an international code began in 1828 with Denmark. They took the form of reciprocal treaty arrangements between individual countries by which foreign authors received the same protection as did native authors. Britain joined the movement in several arrangements between 1844 and 1886. In 1885 a uniform international system of copyright was initiated by the Berne Convention, which is still in force. The customary term of protection is the author's lifetime plus 50 years. Most countries subscribed to the Convention, but not the United States or Russia. The United States continued to protect its domestic printing industry up to 1955, when it joined the Universal Copyright Convention (UNESCO 1952). While the Berne Convention prescribed a minimum level of protection, the Universal Convention was based on the concept of "national treatment"—each member country treating works by citizens of other member countries as it would those of its own citizens. Thus the United States was able to enter into an international agreement without the necessity of immediately revising its own copyright law. Since the Universal Convention contained a provision that the Convention would not be applicable between any two countries that belonged to the Berne Union, it served primarily as a treaty between the United States and the countries that recognized international copyright. The Soviet Union became a party to the Berne Convention in 1973.

International copyright code

The 20th century. In the 20th century the effects of state education in the more advanced countries became increasingly apparent. Standards of living rose, and as in earlier times, these two conditions brought increased use and publication of books. During the late 1890s and the early years of the new century many new publishing houses were founded. In the industrialized countries, though wages were rising, a small business could be staffed economically, and printing costs were such that it was economically feasible to print as few as 1,000 copies of a new book. It was thus comparatively easy to make a start, especially because the long-term credit that printers were prepared to grant made a minimum of capital necessary.

Book publishing grew to a substantial industry, consisting mostly of small units in the Western world but also embracing a number of large concerns, many of which were public companies employing staffs of 1,000 or more. Specialization became frequent, particularly in educational books, as the potentialities of the new school populations were realized. Some, such as Macmillan, in both their British and U.S. houses, had begun to issue schoolbooks almost by chance; then, as their sales grew most profitably, they developed separate departments for lower school and college textbooks. Others, such as The American Book Company, and Methuen in London, had begun specifically with educational books in mind. For more than one leading London firm, India, despite its high illiteracy rate, began to grow strongly as a market and to repay the care and expense involved in setting up separate Indian branches.

Textbook publishing

The first literary agents. A new factor at this time, which was to change the financial climate for fiction publishers in particular, was the advent of the literary agent. The first agent began business in 1875, and between 1900 and 1914 many more appeared. Reasonable though it was that authors who were unable themselves to handle their business with publishers satisfactorily should employ a professional to bargain for them, the higher rates of royalty and larger advance payments thus secured cut seriously into a publisher's profit, making it considerably more difficult to finance the most speculative part of the business, the encouragement of new talent. The system of literary agents began in England but spread rapidly to the United States and also to Continental countries, though in the latter it did not assume so great an importance. Keenly resented at first, the literary agent, by pressing for higher payment to authors he represented, may have been indirectly responsible for the greater selling efforts that some publishers began to make early in the century.

Sales methods. The discreet sales methods of the 19th century, whereby the traveller merely showed his samples and the publisher took small spaces in newspapers for the bare announcement of title and author of his new books, were replaced by more forceful techniques. In this United States publishers took a prominent part. Less hampered by inhibitions over the more blatant forms of salesmanship than their European colleagues, New York City houses began to take large advertisements, make extravagant claims for the qualities of their books, and thus build up bigger sales for new books than was customary in other countries. As in Britain, the existence of a prosperous middle class with fast-growing incomes was one factor; the vast spread of the population across the continent was another. These factors, combined with the development of the railroad, led to the successful development of mail-order advertising and selling. The sale of books such as works of reference by subscription was another technique that rapidly developed and grew into business worth millions of dollars in the United States and elsewhere. It involved securing an undertaking to buy

Installment selling of books

on installments over many months an already published set of books; it could also be used to secure advance orders for an expensive work, probably in several volumes, that the publisher was planning to issue, as was sometimes done in the 18th century. Continental countries, too, exploited the method; and considerable use was made of the door-to-door canvasser.

Effects of World War I. The coming of war in 1914 naturally had a disrupting, though not wholly destructive, effect upon book publishing in European countries. Shortage of paper necessitated rationing to two-thirds of prewar consumption in the case of Britain, while from hundreds of thousands of those in the armed forces came a tremendous demand for light reading. Although at one time the cost of paper rose to eight times its prewar level, sales of books increased sharply. The extra quantities could be supplied only at the expense of quality, and the standards of paper and binding were appalling. It would have been disastrous for a publisher to be left with large stocks of these books since paper supplies quickly returned to normal after the war, and the poorly produced books became unsalable. Of Continental countries, Germany suffered the worst shortages, though the principal publishers were able to stay in business; a worse ordeal in many respects awaited them with their postwar inflation. In England, by contrast with the situation in World War II, there was reluctance to recognize books as of any special importance to the national effort; virtually no direct use of them was made by the government, and it was not until the last four months of the war that a small proportion of publishers' staffs were granted any relief from compulsory national service.

Loss of purchasing power by the middle class

An immediate aftereffect of the war in Europe was a sharp reduction in the purchasing power of the middle class. Whereas before, in most European countries, a proportion of the educated and professional classes could be relied upon to spend regularly upon new books, high taxation, inflation, and trade depression in the postwar years cut down on spare money. Those publishers who continued to cater only to that public found it increasingly difficult to trade profitably, and many went out of business or were absorbed into larger firms. In the United States, on the other hand, boom conditions in the postwar years produced a still more prosperous and enlarged middle class ready to absorb an increasing supply of books. The number of publishing houses grew; and more and more U.S. authors, such as Sinclair Lewis and Ernest Herningway, found a world market. British and Continental publishers turned more readily than before to New York City in search of fresh talent. Universities also increased in number more rapidly in the United States than elsewhere, producing a larger demand for college textbooks. Publishing them became an immensely important part of the business for many U.S. firms, which in some cases depended upon their college departments to carry other parts of their operation, such as the fiction side.

The book club. A new development of vast potential at this time was the book club, an association of members who undertook to purchase, usually each month, a book selected for them by a committee, the advantage being that the book in question was supplied at a lower price than that at which it could be bought in a bookshop. The scheme, of which an early forerunner was the Swiss Cooperative Movement in about 1900, had obvious attractions for that part of the reading public that had no direct access to a bookseller. The pioneer Book-of-the-Month Club in America (1926) developed a membership that ran into hundreds of thousands, followed by The Literary Guild, its great rival, and specialized book clubs that covered a variety of special reader interests. These clubs were strongly opposed at first by both publishers and booksellers, who disliked the additional emphasis placed upon the potential best seller, but they came to supply a genuine need. They also helped to offset the enormous amount of book borrowing from libraries. From the 1950s onward, however, their popularity was somewhat affected by the availability of inexpensive paperbound books sold in thousands of outlets outside the regular book channels.

Design standards. As noted above, machine production had lowered standards of design. William Morris and his Kelmscott Press, however, had begun to work for better typography and book design in the 1890s; and his example had led to the establishment of other private presses, such as The Doves Press and the Ashendene Press, which produced editions (usually limited) of exceptional beauty, printed on handmade paper. Though aimed essentially at the collector and issued at high prices, such books began to influence the more discerning publisher; and by the 1920s a few firms, such as Alfred Knopf in New York City, Chatto and Windus and Jonathan Cape in London, and the Insel Verlag in Leipzig, were seen to be far ahead of their competitors in their standards of design. With careful planning, skillful selection of type face, and provision of layouts to guide the printer, more and more publishers managed to achieve typographically handsome books at a commercial price. It was all part of the Design in Industry movement, which sought to demonstrate that mass production need not preclude beauty. It should be noted, however, that responsibility for design was passing from the printer to the publisher; as the former, with the growth of his business, became more the industrialist and less the craftsman, the latter realized that he must himself take charge of this aspect of the book.

The Great Depression. The great trade slump that began in October 1929 brought a swift decline in the prosperity of U.S. publishing. By 1931 British publishers could no longer depend upon selling a high proportion of their books to the United States, either in the form of physical copies or by way of a contract conceding the U.S. rights. Though the book trade of Europe proved a little more resilient than some other industries, it passed through a difficult period. Sales declined, profits were negligible, and there were many bankruptcies. Attempts were made to find new outlets for books and fresh ways to attract the public to them. In London an annual Book Exhibition was run by ***The Sunday Times*** from 1933 to 1938; and ***The New York Times*** tried a similar venture in its city. The Germans continued to hold their annual Book Fair in Leipzig, but this was primarily a trade function. Some British newspapers, striving for higher circulations, approached publishers to supply them with huge numbers of their popular books, specially printed, to be given away or sold very cheaply, in exchange for coupons from the papers. Booksellers resented the practice, but for hard-pressed publishers it was financially attractive. In the rather desperate climate of the times, some publishers also spent inordinate amounts on newspaper advertising. Reprint book clubs proliferated too, again to the benefit of the few publishers and authors fortunate enough to secure a choice. In 1932 a valuable innovation that stimulated sales was the Book Token, a form of gift certificate. The invention of an English publisher, Harold Raymond, it was at first opposed by many booksellers; but it went on to become a major factor in Christmas sales, and the system was adopted in other countries and by other trades. It was the complete solution to the problem of a donor who wished to make a gift without knowing the recipient's taste.

Appearance of paperbacks

Even in the depressed conditions, publishers still dreamed of tapping a wider readership. This began to become a reality in 1935, when Allen Lane launched his pioneer Penguin series of paperbacks. It was a risky operation, involving speculatively high initial printings to keep down the unit cost. But despite the strongly held belief that paperbacks would not appeal outside the Continent, where they had sold freely, and the resistance of booksellers, who feared a sharp reduction in their receipts, the new series quickly caught on. They represented unprecedented value at the original price of sixpence, equivalent to the cost of a small item in a variety store. Though printed on cheap paper the books employed good typography—far superior to that of any earlier attempts at paperbacks—and the original cover design was attractive in the bold simplicity of its orange and white stripes. A United States agency was arranged shortly before the war and was later taken over by Victor Weybright, who subsequently established the highly successful New American Library for the mass promotion of paperbacks in the world market.

Nazi persecution of the Jews in the immediate prewar years and the impact of the war itself caused a wave of emigration, from Germany and Austria in particular, which brought fresh publishing talent to both Britain and the United States, as well as to other countries, including Australia. Some of the striking developments in the production of art books, with beautiful coloured illustrations, were a direct result of this movement, which bore its fullest fruit after the war. A curious by-product of the Nazi era was the final disappearance of the Gothic typeface in Germany. In order to take the sensible step of abolishing it, the Nazis decreed in 1941 that ***Fraktur*** was a Jewish invention.

Effects of World War II. The war in 1939 that European publishers had feared would utterly destroy their business proved in many respects less terrible in its effects on books than had been imagined. While the destruction of buildings, plant, and vast stocks of books, most notably in London and later in Leipzig, brought publishing to a standstill for individual firms, the activity as a whole continued. As in 1914, but to an even greater extent, the demand for reading matter, both for instruction and entertainment, grew enormously. The nature of the war, with its long periods of waiting alternating with intense bouts of frenzied activity, both induced the need and provided the opportunity for reading. As a result, book sales in the "free" countries rose to fresh heights. The occupied countries of Europe had to endure censorship and a tight control of materials; but most publishers survived and were swift to renew contacts with London and New York City colleagues immediately after the war.

Wartime paper shortages

In the United States, though they were subject to some shortages and inconvenience, publishers were comparatively untouched by the war, and their business expanded rapidly. In Britain, however, because of the acute pressure on shipping, the import of esparto grass, that essential ingredient for good book papers, was strictly limited; and a publisher's paper ration was reduced to 37½ percent of his prewar annual consumption. By closer setting of type and the use of much thinner paper, the ration was stretched to produce the maximum number of copies; but the final appearance of British books inevitably suffered, and they began to compare very unfavourably with those of the United States.

In countries that suffered severe paper shortage there was, of course, a sharp reduction in the number of new books and in size of editions; consequently, with the increase in demand, the available books were rapidly sold out. The result was an enormous, if illusory, increase of profitability for publishers; and despite heavy wartime taxation they found themselves in far better shape financially than ever before. Instead of holding large, and often very slow-selling, stocks with insufficient cash resources, publishers had little stock but ample cash. Thus "birth control" for new books, so often preached but never practiced, had come about and had undoubted commercial benefits in the short term. There was, too, the marginal advantage that those new authors who were able to secure publication in the war years could be virtually certain that their books would be quickly sold out. In these artificial conditions, many publishers were more prepared to risk the work of an untried author. Against this, however, was the very serious shortage of standard works of every kind, classics and educational and reference books; at one time the cry went up that "Shakespeare is out of print!" While a small extra tonnage of paper was released in England in 1942 for the reprinting of books that were considered "nationally important" in wartime, no one could possibly pretend that there was not a real book famine in most European countries. After the war it took about five years for paper to become reasonably plentiful again. Despite the disruption of the war, however, interest in books had increased enormously; and sales were furthered by the total disappearance or severe rationing in most of the warring countries of so many consumer articles that normally compete with books. Contrary to the fears of many publishers, a new reading public was emerging, and it was not lost in the postwar world.

The postwar period. After the end of the war, there was an awkward year or so of reorganization and anticlimax, when many wartime publications suddenly became unsalable; but then publishing, in almost every country, once more expanded rapidly. People who had been cut off entirely from the rest of the world displayed an immense hunger for the books in English that had appeared during the previous six years. Much new business developed in the sale of the actual books and in translation rights. Such conditions continued at a higher level than they had attained in the 1930s, and they were to be further stimulated with the rise of the Frankfurt Book Fair. Social change came to many countries, bringing a broader spread of purchasing power and above all wider educational opportunity for much of the population. The change was to set book publishing upon a bolder and more adventurous course, turning it from a minor industry into one of sufficient growth and profitability to attract even the professional investors. The steady expansion of publishing in Great Britain after the war is illustrated by the following figures for new books and reprints, which may be compared with the 14,904 produced in 1939: 1947, 13,046; 1957, 20,719; 1967, 29,619; 1971, 32,538; and 1980, 48,158. The trend in the United States has been similar.

Revival of German publishing

A feature of the early postwar years was the remarkable phoenixlike rise of the German book trade, literally from the ashes of the Allied air raids, which had destroyed the principal cities with their publishing offices and printing works. Because Leipzig was in the Russian-controlled zone of Germany, however, the centre of the trade moved to Frankfurt for the first time since about 1650. As part of its drive to become the commercial capital of West Germany, Frankfurt developed its exhibition facilities rapidly. Thus, the book trade fair had ideal conditions in which to thrive. Before 1939 it had been largely a domestic affair at which German publishers displayed their new works to booksellers, with only a small number of foreign publishers participating and those almost entirely Continental; but it has steadily grown to be the greatest meeting place for publishers from all over the world. Each year more firms take stands at the fair, and an enormous amount of reciprocal business is done. It provides unique opportunities for the planning of the lavishly illustrated trade, or general, book aimed at an international market. Such books, known as coproductions, depend upon the printing of a large edition, often 100,000 or more, of the coloured illustrations, which are allocated to the participating firms; each firm then arranges for the translation of the text into its particular language.

Iron Curtain publishing. Behind the Iron Curtain, which was not wholly impenetrable to the book trade, publishing was subjected to a state control similar to that initiated in Soviet Russia in 1917. Very few of the famous publishing houses of Poland and Czechoslovakia remain, and in every case they have become state owned and controlled. The normal pattern is for all books upon a particular group of subjects to be issued from one publishing house. Thus in Hungary of the 18 publishers listed in 1981, the principal ones dealt respectively with science, political history, agriculture, music, belles-lettres, military, and technical subjects. The organization in Romania is similar; but in East Germany it is significant that many of the prewar firms remain, 10 dating back more than 100 years, though all are subject to government control. From such statistics as are available, it would appear that there are many fewer individual publishing houses in the Iron Curtain countries; but their output is large, and the number of copies of each title is probably greater than in Western countries. It may be also that more books are bought annually in Eastern Europe, for there are fewer rival consumer goods there. In 1980 in the Soviet Union 80,700 new books and reprints were issued in 145 languages, almost 90 of them being languages of the Soviet Union. The number of copies of books and pamphlets printed in the Soviet Union in 1980 amounted to 1,760,000,000.

Rise of photocomposition

The Paperback Revolution. Besides the economic and social changes that favoured publishing after 1945, an outburst of knowledge, particularly in science and technology, produced many new subjects, many of them subdivided into the highest degree of specialization, all of which called for new books. The many new universities and colleges of technology that sprang up all over the world formed a strong market for the thousands of college books, which came to make up such a large part of many a publisher's list. At the same time there was a major advance in printing, a break away from the traditional letterpress system dependent upon lead type. Photocomposition (composing of printed matter by photographic means rather than by hand), coupled with offset printing technique, obviated much of the hand work of the earlier methods, improved working speeds, and prevented costs from rising as steeply as they would otherwise have done. The trend was toward giant machines for mass production, giving a favourable price for cases in which 100,000 or more copies were needed. Such giant machines became essential for the printing of paperbacks, but the problem remained of printing economically those "short runs" of 3,000 or so in which new authors, from whom many of the important books of the future must come, are normally tried out. Many find grounds for believing that computer typesetting will help cut costs.

By the early 1950s the Paperback Revolution was well under way. Growing from the prewar Penguins and spreading to many other firms, paperbacks began to proliferate into well-printed, inexpensive books on every conceivable subject. Generally known as Pocket Books on the Continent, they swept the world, converting book borrowers into buyers and creating new book readers on a scale never known before. Their use has been particularly widespread in the developing countries, notably those of Africa. Besides their cheapness, putting books for the first time into the area of impulse buying, and the wide range of first class literature available, the new paperbacks had remarkable ubiquity, being found not only in bookshops but in drugstores, street kiosks, and newsstands in railway stations, airports, and hotel lobbies.

By far the greater number of paperbacks have been reprints of books that have had some success in their original clothbound form. Normally the paperback publisher makes an offer for the rights of the hardcover edition and the royalties are shared between the author and the publisher. The cheapness of the paperback is due essentially to the large number printed, seldom fewer than 30,000 and frequently far more, and not, as is often supposed, to the use of paper instead of a hard cover for the binding. While many of the big paperback houses have produced a certain number of new, hitherto unpublished books, the paperback operation is dependent in the main upon books originating with the conventional publishers. It is a fallacy therefore to suppose that, for all their seeming dominance, the paperback is likely to oust the hardcover book.

Scholarly paperbacks

A smaller selling type of paperback has sprung from the enormous growth in the number of university students all over the world. This is the reissue of works of scholarship, science, religion, literature, and art. Many had been out of print for years, and they had often been issued originally in small editions of no more than 2,000 copies by university presses or other specialized publishers. This great extension of the market began in the United States in the 1950s, with prices ranging from 65 cents to $1.95, at that time unheard-of levels for paperbacks; the idea soon spread to Britain and the Continent. This type of operation has usually remained in the hands of the original publishers of the books, who have developed their own series of "University Paperback Books." It became customary for many new academic books to be issued simultaneously in both cloth (hardcover) editions and as paperbacks, the usual price of the latter being a little more than half that of the cloth edition.

University and government presses. The increase in the number of universities was accompanied, as might be expected, by an increase in the number of university presses. The purpose of the presses, like that of the older ones, is to serve the needs of scholarship, to publish specialized material that a purely commercial firm would find impracticable to handle. Their freedom from the more acute profit-making pressures, often a result of direct subsidies, coupled with their assured, if limited, market, enables many to reach high standards of production and commercial viability. Some of the older establishments, such as the Oxford University Press, are, of course, enormous and profitable organizations with worldwide connections and a large list of more general publications.

Another type of publishing house not usually in direct competition with ordinary firms is the state printing office, which is responsible in many countries for issuing all manner of public and official material. In England, Her Majesty's Stationery Office, originally created in 1786 to coordinate office supplies for government departments, has come to issue a wide range of excellent books and pamphlets in connection with museums, galleries, and the advisory function of ministries, besides official papers. In the United States, the Government Printing Office in Washington, D.C., was set up by Congress in 1860 for similar purposes; it too has steadily widened its field of operations. The Soviet Union and China have similar organizations to issue their publications.

The censorship problem. Since 1900 the more overt forms of censorship have continued to wane in the West, though they have never wholly disappeared. The last vestiges of formal religious censorship came to an end in 1966, when the Second Vatican Council decreed that the Roman Catholic Index was not to be renewed. There has been, however, a fierce recrudescence of political or ideological censorship, especially in Germany under the Nazis and in Communist-dominated areas. An intractable problem in the mid-20th century is censorship on moral grounds, with reference in particular to "obscene publications." The pressures tend to be covert and dependent on the prevailing moral climate. This situation leaves a publisher in the invidious position of being forced, on occasion, to risk the outcome of a prosecution. Short of complete freedom, for which the case remains as strong as ever, there seems to be no satisfactory solution.

Despite some pessimistic prophets, books seem unlikely to be supplanted in the foreseeable future. They have already survived several threats—*e.g.*, from radio and television—and have in fact gone from strength to strength. A new threat comes from alternative, electronic methods of storing and presenting information. Some publishers view this rival seriously enough to acquire active interests in the cassette business, for instance. Yet this action is usually taken more as a diversifying operation, to broaden the publisher's base, than from the conviction that books are dying. For convenient, portable availability, nothing compares with a book. Nor is there any pleasure, for a large proportion of mankind, that is capable of replacing the reflective, interior communion that only a book can offer.

Role of the editor

Current publishing practice. Every publishing house has its manufacturing, marketing, and accounts departments, but the heart of the business lies in the editorial function. This has changed in its mode of operation over the years and still varies from one country to another and between firms, but not in essentials. The editor, or sponsor as he is sometimes called, who is often a director, selects the books to be published, deals with the author, and is responsible for the critical reading of the typescript (and its revision if necessary) and for seeing the book through the press, in consultation with the manufacturing and marketing departments. So vital can the editor's part be that his presence in a firm, or his transference to another, can be a major factor in attracting authors to it. In a firm of some size, the editor travels extensively, with regular visits to London, New York City, and other principal cities of the world being as much a part of the work as the quieter but equally demanding desk work upon which the quality of the published book may ultimately depend. A particular branch of editorial work that has grown to be of cardinal importance since World War II concerns the conception, planning, and publication of the hundreds of books needed for the educational schemes, at every level, that have been initiated in the developing countries. From Britain in particular, but also from other former colonial powers and the United States, editors specializing in school and college books have made long journeys to visit teachers and lecturers to promote the writing of the required textbooks. The educational editor has to concentrate almost wholly upon the commissioning of books to fit a particular syllabus in a school or university. Rarely, if ever, does the editor receive an unsolicited typescript that can be accepted at once. The editor must seek material by regular visits, either personally or by one of his assistants, to schools or colleges to find the teachers who have the makings of authorship. Capable teachers often fail to produce a usable book at the first attempt, and much time must then be spent on its revision. The school book that is widely adopted may sell for a generation and reward author and publisher on a scale beyond the dreams of those concerned only with general books. Equally, nothing can fail so completely as the school book that gets no adoptions. Besides the editor as described, there is also the editorial department, which is responsible for the detailed preparation of the typescript before it is printed. This receives more attention today than it used to. Facts, figures, and references are checked, and inelegancies of style are polished where necessary. Many authors owe much more to the work of a good editor at this level than is generally acknowledged.

Subsidiary rights

Broadcast and television rights

Remote copying

Forms of copyright. Book publishing depends fundamentally on copyright (see above), which is the sole right to copy or to produce a work, conceded to the publisher by the author in their mutual agreement. Without this element of monopoly it would be impossible for a publisher to trade. It is also the guarantee for an author that he has legal rights to prevent the use of his material without fair compensation. On the expiration of copyright, anyone is free to publish the work in question without payment to the author or his heirs. Copyright used to be simple and indivisible; nowadays, however, there are many alternative forms in which the text may be reproduced. Their exploitation is governed by individual clauses in the agreement. These subsidiary rights may be briefly summarized. American rights for a British publisher and British rights for a book of American origin can prove to be exceptionally profitable. Though a book normally has its greatest sale in its country of origin, there are cases in which it does even better abroad. The richness of the U.S. market gives it a particular attraction for publishers and authors of almost every other country. Translation rights have become a valuable source of additional revenue, particularly since the establishment of the Berne Convention. All the signatory countries agreed to copyright protection for the unpublished works of nationals of other member countries and for all the work first published in the Convention countries. While many books may earn no more than a few hundred dollars from the rights of translation in a single country, some world best-sellers, by authors of international stature, have a demand in almost every country, new or old, for a translation, and the aggregate earnings are then immense. Paperback rights for the more salable books, whether fiction or fact, are customarily offered to one of the major paperback houses, which flourish in most larger countries. For a best-seller there can be keen competition between the paperback houses, and advances well into seven figures may be offered to the original publisher, who normally controls the reprint rights. The original publisher also stipulates the earliest date at which the paperback may appear; as a rule, this is not less than 12 months after first publication. Serial rights may be sold in several divisions: first serial rights, for which the best price can be obtained probably from a large circulation newspaper or magazine in the capital city, can be in a number of installments appearing several weeks ahead of the issue of the book, or such publication may "straddle" the appearance of the book, some installments before, the rest after. Second serial rights, for which much less is paid, can still yield useful sums: after first serialization has taken place, lesser papers in other parts of the country, or in other countries where the same language is spoken, can use the book. While no publisher would lightly set aside a rich offer for first serial rights, it is often a question whether such publication may not damage the sales of the book. On the other hand, the sum paid can be the equivalent of royalty earnings on many, many thousands of copies, which might in fact never have sold. Digest rights, and their allied condensed book rights, represent another lucrative subsidiary use for books of wide general appeal. ***The Reader's Digest*** in the United States, with its great circulation and numerous foreign language editions, is the most important market for such rights; large sums are paid for the books thus used. Again, a publisher must face the likelihood that this use will seriously affect sales, but the handsome payments made are full compensation for both the publisher and the author, who divide the receipts more or less equally between them. Book club rights are also among those the publisher can exploit; here again, fees received from the clubs are shared with the author. Broadcast and television rights in books interest a publisher primarily for the possibility of bringing a book and its author to the attention of a large segment of the public, rather than for the amounts paid. As a rule, there must be direct quotation from the text if a broadcasting company is to pay anything to the publisher. A television interview with the author, including sight of a copy of the book, is of great publicity value, and the author may even receive a fee for the appearance, but this is not part of the book's earnings. If the author can show a film relating to the book, it would be paid for at the appropriate rates for television use. In sound broadcasting, the reading of a book as a serial is one most remunerative possibility; the other is its full dramatization as a serial. The latter is, of course, still more valuable on television, but such use of new books is rare; more commonly this treatment is accorded to works of classic status. Dramatic and film rights can have importance for fiction, biography, and other general books, but only a small fraction of one percent of those published can be exploited by these means. From the publisher's standpoint, it is reasonable to share in the proceeds from the sale of these rights, for they result from the publisher's efforts. The last group of subsidiary rights, rights for mechanical reproduction by film micrography, xerography, tape or disc recording, or any other technique of sight or sound, are of increasing concern to publishers. Dry-copying machines, easily operated, are to be found all over the world in public, university, and school libraries, and while ordinarily only single copies can legally be made solely for the purposes of private study, it is a simple matter, though illegal, to run off a number of copies of long extracts, which then make it unnecessary to buy more than one copy of the book. Similarly, microfilm enables a single copy to satisfy many users and reduces the number of copies of the book that must be kept available in a library. Remote copying appears likely to make certain types of scholarly and reference publications unnecessary. A perfect copy of the material in question has merely to be held at a central station, from which it can be reproduced on a copying machine at other and distant points; ***e.g.,*** in public libraries. The situation could even arise in which a primary copyright resided in an electronic store of a computer, and the publisher would receive "subsidiary" volume rights. Wherever material originates in the form of a book, however, the publisher must retain an interest in all forms of reproduction as part of his resources for promoting experimental and imaginative work.

Publisher's agreement. A publisher's agreement with an author normally specifies that in consideration of certain payments the former shall, during the legal term of copyright, have the exclusive right to produce or reproduce the said work in any material legible form throughout the world. In many cases, however, this agreement is modified to exclude some of the subsidiary rights named above, depending on the bargaining power of the author or his agent. After clauses specifying the extent of the rights conferred, the basic clause of a royalty agreement is that which states the rate of royalty to be paid. A typical wording is as follows: "On all copies of the said work sold on the normal terms a royalty of 10 percent shall be paid on the published price rising to 12 percent after the sale of 5,000 copies and to 15 percent after 10,000 copies." Other clauses provide for somewhat lower royalty rates on export sales and on cheap editions, on which the publisher's margin of profit is considerably less. Provision is also made for division between author and publisher of any payments received for such subsidiary rights as are included in the agreement. A publisher can fairly claim a share in them if they arise from the fact of book publication. Proofreading is another important matter covered by the agreement, the author being responsible for this. If the cost of making his corrections exceeds a stated figure, he has to pay for the excess. Lastly, in the majority of publishing agreements there is an option clause under which the author undertakes to give the publisher the first offer of his "next literary work suitable for publication in book form," usually with the addition that if, after a stipulated time, no terms shall have been agreed on for its publication the author is free to submit it elsewhere. The exact form of the legal instrument varies in detail; it is possibly drawn up in the greatest detail by U.S. firms because of the complexities of their system of selling: *e.g.*, by mail order, subscription, and similar means, in which the publisher must incur abnormal costs in order to secure the business. The vital condition for this publisher–author relationship, in the past often conducted with complete informality, is that there must be a legal document, a contract, setting out the rights and obligations of the two parties.

Literary agents and scouts. Several references have already been made to literary agents. They have become increasingly important and prominent as publishing has grown more complex. A high proportion of the more successful authors of novels and general books employ literary agents to place their books with publishers and to handle negotiations with them, the author being charged a commission of 10 percent. Besides negotiating and drawing up the contract with the firm, the good agency is equipped to handle the many subsidiary rights. Because an important element in the value of agents to authors is their capacity to extract better terms than the authors would themselves, it is not surprising that publishers have resented their intrusion into the personal, and often very friendly, relationships between themselves and their authors. There can be no doubt, however, that agents do perform a valuable service in relieving authors of the considerable amount of routine work that their literary affairs may involve. Advice on possible new books to be written and occasionally, for the author of exceptional promise, an advance on anticipated earnings are also part of the assistance that the agent may offer. It must be emphasized, however, that agents are interested mainly in general books; they are seldom equipped to handle specialized and technical works.

Literary scouts

Another publishing auxiliary who became significant in the 1950s and 1960s is the literary scout. Though a few had been employed earlier, mainly by U.S. publishers, who had their "lookouts" in one or two European cities, the practice has become more widespread. Many Continental publishers, including Scandinavian firms, employ men or women resident in London, Paris, and New York City to alert them at once to any promising new book, either written or just published. The scouts, who may be connected with a newspaper or literary agency, are usually paid some modest amount as a retainer, probably with a commission of 1 or 2 percent on the published price of the books they recommend, in effect a small royalty on sales. On occasion a valuable find can be quite lucrative to the lucky scout; frequently everything depends upon the speed with which a copy of the work can be got into the hands of the publisher.

Selling and promotion. The publisher's techniques for the promotion of his wares have become increasingly sophisticated in all advanced countries. Typical travellers or booksellers are likely to have college educations, certainly in the United States; they set out only after careful briefing at their home office, with elaborate samples and sales aids, often in a car provided, or partly provided, by the firm. The itinerary for their calls on bookshops (or in the case of educational books to schools and colleges) is prescribed by a supervisor, who usually checks the resulting orders against a quota. A well-run publishing house issues two or three seasonal announcement lists with details of its forthcoming books, as well as an annual catalogue of its present and past books still in print, which are sent to the principal booksellers and librarians. For many books a prospectus may be issued, both for the use of booksellers and for direct mailing by the publisher. The distribution of review copies to the press is the last item in the normal program. These three steps, travelling, catalogues, and reviews, are the vital elements in the machinery of book distribution, which it is virtually impossible to accomplish without the professional work of a publisher. The capacity of some authors to produce a quite presentable book with the help of a printer still leaves them far from their objective unless they can find a publisher to undertake its distribution.

Book advertising

Newspaper and periodical advertising is the publisher's principal means of reaching the public, and standards here have also risen considerably since World War II. Originally handled entirely by the publisher's own staff, it is now not uncommon for the larger houses, especially in the United States and in some Continental countries, to employ advertising agencies to prepare the copy and the general details of the campaign for any important book. While few authors consider that their books are advertised adequately and most publishers are highly doubtful whether press advertising does in fact sell books, the amounts spent in relation to sales revenue are much higher than for most other commodities, seldom less than 5 percent for new books. Without their receipts from publishers' advertising, some periodicals would find it impossible to devote so much space to book reviews, which are in themselves a most valuable aid to sales. The news value of many new books also enables them to secure free publicity through references in the general, as distinct from the literary, pages of a newspaper. A publisher with imagination, or the press officer if there is one, can often suggest aspects of a book susceptible to such treatment. Broadcasting and television services, too, can sometimes be interested in books and their authors, and the resultant publicity may then be extremely effective.

Over the whole field of sales promotion, as publishing houses have grown in size and profitability, there has been a marked tendency for the more commercial methods of general business to be applied to books, which are aggressively promoted to retailers and the public in the same manner as are many other commodities. Though this may increase sales, at least in the short term, it may be doubted whether it is in the interests of the public and to the long-term advantage of good publishing.

II. Newspaper publishing

ORIGINS AND EARLY EVIDENCES

If the essence of a newspaper is the regular publishing of information about recent events, then something of the kind must be as old as civilization. "A community needs news," said the distinguished English author Dame Rebecca West, "for the same reason that a man needs eyes. It has to see where it is going." But there is a gulf between all the early methods of spreading news, such as public announcements by a town crier, the posting of proclamations, or the private circulation of newsletters, and a modern daily paper. The variety and amount of the information, the regularity and speed with which it is published, and the number of people it reaches have grown in ways undreamed of a century ago. This has come about largely through technical developments in printing and communications, but it has involved also a radical change of outlook. Those to whom news is especially important (that is, governments and commercial interests) have always had their own ways of gathering intelligence; but they use it for their own ends and only make public what they choose and in the manner they choose. Until the invention of printing, the public at large had to take what little it was given or make what it could of rumour and hearsay. Printing, however, by offering a powerful new

The public right to know

means of spreading news, eventually gave rise to a powerful new idea: that the public was in a fundamental sense entitled to know—to know, not merely from official or interest& sources but from a free and voluntary source with an ethic of its own, responsible to the public itself, both for standards of truth and for completeness of information, and to know, not merely to satisfy curiosity but also as a means of participating in public affairs. The growth of this idea, the attempts to realize it through newspapers, and the limitations imposed by the forces of authority, ideology, and economics (not to mention human frailty) make up the history of the press. It has been closely bound up with the growth or suppression of democratic ideas. Though the small Athenian democracy could rely on word of mouth to create public awareness, a large modem state depends on its press; and so the freedom and quality of its press have become sure indications of the freedom and quality of a state. Nor has this situation been greatly changed since the advent of radio and television.

The *Acta Diurna* of ancient Rome

Rome. Whatever methods of spreading news there may have been in the oldest civilizations, the earliest of which there are definite facts date from ancient Rome, where a form of gazette was published daily from 59 BC called the ***Acta Diurna.*** Copies of it, written in manuscript, appear to have been hung in prominent places in Rome as well as in the provinces; its origin is attributed to Julius Caesar. The "gazette" recorded important social and political events: plebiscites, public appointments, edicts, treaties, trials and executions, naval and military news, births, marriages, and deaths; anything of special interest, such as the fall of a meteorite or similar portents; and even sports news—*i.e.*, the outcome of gladiatorial contests. The ***Acta Diurna*** were, of course, official notifications from the various authorities and not news gathered freely and independently. They were complementary, for the public at large, to the ***Acta Senatus,*** which recorded the proceedings of the Roman Senate and could be consulted by senators. In addition, the Romans were the first to cultivate the art of general correspondence; and the newsletter, as opposed to the family letter or letter of condolence, was a recognized type.

China. The only other really early approach to a newspaper seems to have existed in China. This too was an authorized gazette, a kind of court circular, ***pao,*** or "report," which began to be issued among officials in the T'ang dynasty (AD 618–907). It appeared in various forms and under various names more or less continually to the end of the Ch'ing dynasty in 1911. Under the Sung dynasty (960–1279), it was known as ***Ti-pao,*** "Palace Report," or ***Ti-chan,*** "Court-Reading-Matter." During the Ming dynasty (1368–1644), there was a government bureau called ***T'ung-cheng-ssu,*** responsible for circulating official news, and this was also the name of its bulletin. Finally, under the Ch'ing (Manchu; 1644–1911), the circular was known as ***King-pao,*** and as the ***Peking Gazette.*** It was handwritten or printed from blocks up to the 17th century, when it began to be printed from wooden type. Though essentially a means of communication among officials, it reflected the accomplishments of the mandarin class.

BEGINNINGS IN EUROPE, AMERICA, AND ASIA: 1400–1800

The development of the newspaper tends to fall into three phases: first, sporadic forerunners, gradually moving toward regularity of appearance; second, more or less regular journals, liable to suppression and subject to censorship and licensing; and, third, a phase in which direct censorship is abandoned but attempts at control continue through taxation, bribery, and prosecution. Thereafter, some degree of independence has followed in favourable circumstances.

Manuscript newsletters

Forerunners of the newspaper. The earliest forerunner was the manuscript newsletter, containing political and commercial information, which was circulated in the late Middle Ages between the various branches of the large trading companies. The newsletters of the financial house of the Fuggers of Augsburg were particularly well known and available even to selected outsiders. Venice, as an important centre of commerce, was also a great centre for newsletters. During the war with Turkey in 1563, the Venetian government issued regular ***fogli d'avvisi,*** or written newssheets which were read aloud in public. Since the cost of admission to a reading was a ***gazeta*** (approximately three-fourths of a penny), this became a common term for such official sheets, even after copies were printed and sold.

Another forerunner was the printed news book or news pamphlet, which usually related some single topical event, political, scandalous, or marvellous. The earliest example known in England is a pamphlet of four leaves printed by Richard Fawkes, known as ***The trew encountre,*** an eyewitness account of the Battle of Flodden Field, with a list of Englishmen who distinguished themselves. It appeared in September 1513, soon after the battle. There were many such pamphlets on the Continent. As early as 1566 in Strassburg and Basel, some were numbered and thus members of a series. In England these pamphlets became more frequent toward the end of the century; about 450 were published between 1590 and 1610. In addition, monthly, half-yearly, and yearly summaries of news began to appear. One of the earliest was the ***Mercurius Gallobelgicus,*** which was issued yearly in Latin from 1594 to 1635. This was the first publication to adopt the name Mercury (the messenger of the gods), which later became popular almost everywhere.

The first true newspapers. Publications usually regarded as the first true newspapers because they combined miscellaneous topical information with some pretensions to regular, periodic appearance began in 1605–10. Possibly the earliest was the ***Nieuwe Tijdinghen,*** published from 1605 onward in Antwerp by Abraham Verhoeven, though the oldest known copy dates only from 1621. It seems to have grown from a commercial bulletin, the ***Courante Bladen,*** which circulated among the merchants of Antwerp and Venice. Three other very early papers, like many of the forerunners, were German: the ***Avisa*** (later ***Aviso***) ***Relation oder Zeitung*** (1609), founded by Duke Heinrich Julius of Brunswick-Wolfenbuttel and printed at Wolfenbüttel; ***The Relation,*** also in 1609, printed by Johann Carolus in Strassburg; and the ***Gedenckwürdige Zeitung*** (1610) of Cologne. The Dutch, as great traders with plenty of foreign correspondents to supply them with intelligence, were not far behind in publishing these "corantos," as they were called ("currents of news"; ***krant*** is still a common name for a Dutch newspaper). The ***Courante uyt Italien*** and ***Courante uyt Duytslandt*** began publication in 1618 and the ***Tijdinghen uyt verscheyde Quartiere*** in 1619, once or twice a week. The first coranto in English was a translation from the Dutch published in Amsterdam in 1620 by Pieter van den Keere. Corantos in French also appeared in Amsterdam in 1620, though not in France until 1631, when the ***Nouvelles ordinaires de divers endroits*** was produced by the Parisian booksellers Louys Vendosme and Jean Martin. In the same year, however, it was superseded by the official ***La Gazette*** (later ***Gazette de France),*** published by Théophraste Renaudot under the patronage of the cardinal de Richelieu. Many other European countries had their first rudimentary newspaper in the 17th century: Switzerland (1610), Austria (1620), Denmark (1634), Italy (1636), Sweden (1645), and Poland (1661).

Dutch "corantos"

England. In England corantos first appeared in 1621. From references in letters of the time, the earliest seem to have been issued by a London stationer, Thomas Archer; but they were unauthorized, and Archer was imprisoned for his pains. Another stationer, Nathaniel Butter, "got licence to print them honestly translated out of the Dutch"; thus he and not Archer is usually regarded as the father of English journalism. The first to appear was a small sheet published in London on September 24, called a ***Corante, or newes from Italy, Germany, Hungarie, Spaine and France, 1621.*** Between 1621 and 1641 Nathaniel Butter, often in conjunction with Archer, Nicholas Boume, and others, was the main publisher of a stream of corantos, relations, and avisos. These included a numbered and dated series of ***Weekley Newes,*** beginning in 1622. This was far from regular, however, since it

Newspaper Statistics

	date of first news-paper	daily news-papers (1979)	total daily circulation (000) (1979)	copies per 1,000 persons (1979)	literacy rate (per-centage) (1980)
Africa					
Algeria	...	4	425.0	22.0	26.4
Angola	...	5	120.0	17.0	30.0
Benin	...	1	1.0	0.3	24.8
Botswana	...	I	17.0	21.0	22.0
Cameroon	...	3	28.0	3.0	12.0
Central African Republic	...	1*	...	...	38.5
Chad	...	1*	1.5*†	11.0*	15.0‡
Congo	...	3*	8.9*†	...	28.8
Egypt	1875	17*	3,012.0*	79.0*	45.7
Ethiopia	...	5	52.0	2.0	10.0‡
Ghana	1890s	5	345.0†	31.O*	34.1‡
Guinea	...	I	20.0	4.0	10.0§
Guinea-Bissau	...	1	6.0	11.0	28.4
Ivory Coast	...	I	53.0	7.0	41.2
Kenya	...	3	156.0	10.0	40.0
Liberia	...	3	11.0	6.0	21.5
Libya	...	3	...	...	52.4
Madagascar	...	5*	40.0*†	4.0*	78.0‖
Malawi	...	2	31.0	5.0	16.5
Mauritius	...	8	121.0	79.0	61.6
Morocco		9	230.0†	12.0	22.2
Mozambiaue		2	42.0†	4.0	7.0
Niger	...	1*	3.0*	1 0*	5.2
Nigeria	1890s	23*	1,324.0*†	19.0*	25.0
Réunion	...	1	23.0	47 0	62.9§
Senegal	...	1	25.0	5.0	45.6
Seychelles	...	2	3.5	56.0	62.0¶
Sierra Leone	...	2	10.0†	3.0	15.0‡
Somalia	...	1*	...	...	5.2
South Africa	1800	23*	1,728.0*	66.0*	89.0
Sudan, The	...	3	18.0	1.0	68.6
Swaziland	...	1	8.0	15.0	22.0
Tanzania	...	2	182.0	10.0	73.5♀
Togo	...	1	7.0	3.0	54.9
Tunisia	...	5	271.0	44.0	50.0
Uganda	...	1	20.0	2.0	44.0
Upper Volta	...	1	1.5	11.0	8.7‖
Zaire	...	6	45.0†	2.0	35.0
Zambia	...	2	109.0	19.0	40.7
Zimbabwe	...	2	111.0	16.0	34.3
America (North)					
Antigua	...	1	6.0	80.0	95.0§
Bahamas, The	...	3	33.0	146.0	89.7§
Barbados	...	1	21.0	85.0	99.3§
Belize	...	2	7.0	41.0	90.0
Bermuda	...	1	13.0	217.0	90.0‡
Canada	1751	126	5,700.0	241.0	99.5
Costa Rica	1832	4	155.0	70.0	84.7
Cuba	1764	9	891.0	91.0	98.0
Dominican Republic	1804	7	220.0	42.0	67.2§
El Salvador	1820	12	334.0¶	...	49.0
Grenada	...	1	...	...	97.8§
Guadeloupe	...	1	18.0	58.0	...
Guatemala	1729	9	91.0¶	...	37.6
Haiti	1804	4	32.0	7.0	23.4♀
Honduras	1830	7	223.0	63.0	59.5
Jamaica	1718	3	128.0	59.0	96.1§
Martinique	...	1	26.0	83.0	86.9¶
Mexico	1722	...	...	...	86.7
Netherlands Antilles	...	5	54.0	206.0	93.6§
Nicaragua	...	8	170.0	69.0	87.0
Panama	1822	6	148.0	79.0	82.0
Puerto Rico	1807	4	475.0	139.0	90.5
St. Kitts-Nevis	...	1	1.5	22.0	97.6§
Trinidad and Tobago	...	4	193.0	171.0	94.0§
United States	1690	1,787	62,223.0	282.0	99.5
U.S. Virgin Islands	...	4	19.0	176.0	87.0§
America (South)					
Argentina	1801	133	2,556.0 ♀ †	97.0♀	92.6
Bolivia	1825	14	214.0	39.0	39.8
Brazil	1808	328	5,094.0 ♀ †	44.0♀	83.0
Chile	1810	37	945.0†	87.0	90.7
Colombia	1785	38	1,273.0	48.0	98.5
Ecuador	1785	38	400.0	49.0	79.0
French Guiana	...	1	1.5	22.0	74.3§
Guyana	...	3	67.0	77.0	...
Paraguay	1850s	5*	118.0*†	37.0*	79.7
Peru	1744	76*	1,265.0*†	...	76.0
America (South) (continued)					
Suriname	...	5*	32.0*†	91.0*	65.0
Uruguay	1807	21*	553.0*†	191.0*	93.9
Venezuela	1808	69	2,383.0	176.0	84.9
Asia					
Afghanistan	...	14	69.0†	5.0	16.2
Bangladesh	...	30	404.0	5.0	22.2
Burma	1889	7	329.0	10.0	68.3
China	618	392§	38,470.0*	38.0*	95.0
Cyprus	...	9	67.0	108.0	89.0
Hong Kong	...	47*	3,709.0*†	713.0*	80.9
India	1780	1,087	13,033.0	20.0	34.2
Indonesia	1744	95*	3,272.0*†	22.0*	64.0
Iran	1851	24	972.0§	26.0§	36.1
Iraq	1914	5	325.0†	27.0	50.1
Israel	1904	26*	824.0*†	212.0*	93.4
Japan	1860	178	65,880.0	569.0	100.0♁
Jordan	...	6*	126.0	59.0*	60.0
Kampuchea	...	17	...	...	36.1
Korea, North	...	11	1,000.0†	...	92.7¶
Korea, South	...	30*	7,248.0*†	190.0*	88.5
Kuwait	...	8	610.0	452.0	59.6
Laos		3	...	...	60.0
Lebanon		25*	273.0*†	...	88.0
Malaysia		44*	2,477.0*†	187.0*	60.8
Mongolia		1	112.0	69.0	100.0♁
Nepal	...	29*	104.0*†	7.0*	12.5
Pakistan	1823	119	1,094.0	14.0	26.7
Philippines	1811	19	972.0†	20.0*	83.4
Saudi Arabia	...	13*	217.0*†	31.0*	25.0
Singapore	...	11	587.0	249.0	77.9
Sri Lanka	...	23*	1,589.0*†	110.0*	82.0
Syria	...	6	104.0	12.0	46.6
Taiwan	...	32*	3,553.0*†	205.0*	85.9
Thailand	...	18	1,943.0†	42.0	81.8
Turkey	1831	255*	1,400.0*	31.0	54.7
United Arab Emirates	...	3	28.0†	27.0	44.2‡
Vietnam	...	3	500.0†	9.0	65.0§
Yemen (Ader	...	3	12.0†	6.0	0.5
Europe					
Albania	1930	2	145.0	54.0	71.0
Austria	1620	31	2,634.0	351.0	98.0
Belgium	1605	26	2,242.0	228.0	99.5
Bulgaria	1846	12	2,093.0	234.0	95.0
Czechoslovakia	1860	30	4,641.0	304.0	99.5
Denmark	1634	49	1,876.0	367.0	100.0♁
Finland	1771	62	2,289.0	480.0	100.0♁
France	1631	96	10,863.0	205.0	100.0♁
Germany, East	1609	39	8,658.0	517.0	100.0♁
Germany, West	1609	1,229	19,298.0	312.0	99.0
Gibraltar	...	1	2.0	73.0	77.8§
Greece	1821	118	1,200.0	130.0	86.0
Hungary	1721	27	2,585.0	242.0	98.2
Iceland	...	6	127.0	557.0	100.0♁
Ireland	1737	7	770.0	229.0	100.0♁
Italy	1636	73♀	5,484.0♀	97.0♀	94.0
Liechtenstein	...	2	12.0	477.0	100.0♁
Luxemboure	...	5	130.0	358.0	100.0♁
Malta	...	7*	71.0*†	223.0*	87.1§
Netherlands, The	1618	80	4,553.0	325.0	100.0♁
Norway	1763	83	1,859.0	456.0	100.0♁
Poland	1661	44	8,433.0	237.0	98.0
Portugal	1864	28	527.0	54.0	71.0
Romania	1828	35	3,998.0	181.0	100.0♁
Spain	1810	143	3,300.0	128.0	90.1
Sweden	1645	112	4,359.0	526.0	100.0♁
Switzerland	1610	88	2,501.0	395.0	99.5
United Kingdom	1622	120	25,221.0	513.0	100.0♁
Yugoslavia	1791	27	2,282.0	103.0	83.5
Oceania					
American Samoa	...	2	10.0	310.0	97.0
Australia	1803	63	4,581.0	336.0	98.5
Cook Islands	...	1	2.0	106.0	91.8§
Fiji	...	2	54.0	87.0	79.0
French Polynesia	...	3	11.0†	...	97.8¶
Guam	...	1	25.0	222.0	...
New Caledonia	...	1	15.0	109.0	89.4¶
New Zealand	1840	37	1,067.0	345.0	100.0♁
Papua New Guinea	...	1	19.0	6.0	30.4
Soviet Union	1703	694	103,796.0♀	400.0♀	100.0♁

*Figures for 1980. †Partial circulation total. ‡Figures for 1975. §Latest data available. ‖Figures for 1979. ¶Figures for 1977. ♀Figures for 1978. ♁Regarded by nation concerned as 100 percent literacy rate although actual rate is probably somewhat lower.

Sources: ***UNESCO Statistical Yearbook, 1981; Editor and Publisher Yearbook 1981;*** official country sources.

depended on the arrival of boats. Domestic news was confined by censorship to trivialities, and even the foreign news got the publishers into trouble. In reporting the progress of the Thirty Years' War, their partiality for the anti-imperial forces brought complaints from Spain and Austria and upset Charles I. As a result, all news books were suppressed by a decree of Star Chamber, from October 17, 1632, until the lifting of the ban on December 20, 1638. During this period, the demand for foreign news, especially of the campaigns of Gustavus Adolphus, was met partly by corantos printed in Amsterdam and partly by a half-yearly compilation brought out by Butter and Bourne, called *The Swedish Intelligencer.*

The English *Mercurius*

After the abolition of Star Chamber in 1641, the English press had a short spell of freedom, and domestic news at last began to appear. At the same time, the news book finally became a newspaper in form. Under pressure on space and the urgency of events, the old title page and its blank verso (underside) were dropped. The news began directly beneath the title, which was now constant; and aeries became increasingly regular, numbered, and dated. It has been estimated that between 1640 and 1660, nearly 300 distinct news publications were brought out. Many of these were occasional papers reporting the English Civil War, such as *News from Hull* or *Truths from York;* but there were also the propaganda papers on each side, mostly weekly but becoming more frequent. Mercuries abounded; the *Mercurius Aulicus,* followed by the *Mercurius Academicus* on the Royalist side and the *Mercurius Civicus* and *Mercurius Britannicus* (the second of this title; the first in 1625–27) on the side of Parliament. The *Mercurius Civicus* was the first paper to use illustration at all regularly. After the Mercuries came the *Intelligencers, Scouts, Spies,* and *Posts.* Then, under Cromwell (1649–58), strict control was reimposed, and only two official publications were permitted—the *Mercurius Politicus* (which Milton edited for a time) and the *Publick Intelligencer—both* run by Marchmont Needham. When the move to restore the monarchy began, these were superseded by two more official papers, *The Parliamentary Intelligencer* (later *The Kingdom's Intelligencer*) and the *Mercurius Publicus,* both started in 1659 by Henry Muddiman, who was to become the best known journalist of the century. With the Licensing Act of 1662, control became tighter still; and in 1663 Roger L'Estrange was appointed surveyor of the press, with the sole privilege of publishing newspapers; his two were the *Intelligencer* and the *Newes.* L'Estrange soon fell out of favour, however, and in 1665 Muddiman was licensed to start the official *Oxford Gazette,* so called because the court had removed to Oxford to escape the plague. When it returned to London, the paper became the *London Gazette* (1666), under which name it has continued to appear ever since, on Tuesdays and Fridays, though it now contains only public appointments and announcements.

As censorship slackened after the revolution of 1688 and the Licensing Act was allowed to lapse (1694), the English press began to expand and move into its second phase. In 1690, the first provincial paper was started, the *Worcester Post Man* (later *Berrow's Worcester Journal)* and, in 1699, the first in Scotland, the *Edinburgh Gazette;* others followed, mostly biweekly. In 1696, Edward Lloyd, whose coffeehouse had become a centre of marine insurance, began to issue *Lloyd's News,* a single sheet with both general and shipping news. After offending the House of Lords in 1697, he stopped publication rather than apologize but restarted in 1726 what became in 1734 *Lloyd's List and Shipping Gazette,* London's oldest newspaper still in existence. The improvement in packet and post systems after 1691 led to the start of daily papers, which eventually replaced the biweeklies and triweeklies. The first was *The Daily Courant,* a single sheet 13¼ by seven inches (34 by 18 centimetres), begun on March 11, 1702, and lasting until 1735. Like the old corantos, it consisted largely of extracts from foreign journals. The regular voicing of opinion on current political topics and thus the start of the leading article, or editorial, began with Daniel Defoe, in his triweekly *Review* (1704–13), which he wrote almost entirely himself. Finally, the social, artistic, and entertaining aspects of the newspaper got off to a magnificent start with the famous *Tatler* (triweekly, 1709–11) and *Spectator* (daily, 1711–12), both at one penny, the first begun by Sir Richard Steele, the second by Steele and Joseph Addison. These had a tremendous influence, especially on the development of the magazine (see the section of this article entitled *Magazine publishing)*; *The Spectator's* subscribers often numbered as many as 3,000.

The Tatler and *The Spectator*

The nascent power of the press was bound to disturb the government. In 1712 an attempt was made to curb it with the notorious Stamp Act. This imposed a duty of one-half penny on every periodical of half a sheet and one penny on every whole sheet (*i.e.,* four pages). In addition, a duty was levied on advertisements, which had already become a valuable source of revenue to newspapers; and there was an excise on paper. These "taxes on knowledge," as they came to be called, had the desired effect of at once killing off many newspapers, including *The Spectator;* but their effect was only temporary. Despite various increases in the rates the number of newspapers grew throughout the century. To defend itself against attacks—Whig and Tory could be equally scurrilous and venomous—the government relied partly on organs of its own, or on those it subsidized, and partly on bribery. Secret handouts to journalists or regular annual stipends were taken for granted right down to the Reform Act of 1832. For anything really "obnoxious," the government's powers of arrest and seizure of papers and type under the law of libel were still all too ample. Among countless examples, Daniel Defoe was imprisoned and pilloried in 1702 for his pamphlet "The Shortest-Way With The Dissenters"; and as late as 1810 William Cobbett was imprisoned and fined for denouncing flogging in the army. A great fighter for the freedom of the press was John Wilkes, who founded the *North Briton* in 1762 to attack the government of Lord Bute. Wilkes was also involved in the struggle for the full right of juries to decide libel cases (instead of the judge) and for the right to report the proceedings of the houses of Parliament. The latter were regarded as privileged, and only the bare votes were allowed to appear. Before long, however, discreet accounts of debates began to be published in *The Political State of Great Britain* (1711–40), which appeared monthly, and these were developed by several journals, often under a transparent disguise. Fulminations and prosecutions failed to stop them, and in 1771 Parliament was forced to concede this important right.

Improvements in typography

Of the many newspapers that came and went in 18th-century London, the steadiest was probably *The Daily Advertiser* (1730–1807), which offered a good news service, political, social, and commercial news, and plenty of advertising space. One of the longest lived foundations was the *Morning Post* (1772–1937), which finally merged with the *Daily Telegraph.* It was started by a remarkable publisher, John Bell, who in one of his other papers, *The World* (1787), revolutionized newspaper typography and layout. *The World* was printed in Bell's own type and was the first paper to allow sufficient space between lines of type for ease of reading; it also abolished the long "s," the symbol used up until then for the letter "s." Its style was copied at once and in full by the *Daily Universal Register,* the paper founded in 1785 by John Walter, which in 1788 assumed its rather more familiar name, *The Times.* Perhaps typically, however, *The Times* retained the long "s" until 1803. One other distinguished paper of the period was *The Observer* (1791), a "quality" Sunday paper that is still flourishing. In 1795, a circulation of 2,000 was regarded as good; the *Morning Post* was down to 350, but *The Times* had risen to 4,800.

The European continent. On the Continent, the authoritarian phase lasted longer and was more stringent than in England. In France, apart from the official *Gazette de France* (1631–1914), the only pre-Revolution papers of any consequence were the weekly *Mercure de France* (1672–1853, originally founded as the *Mercure galant*), which was more of a magazine (see below); and the first daily, the *Journal de Paris,* which was not started

Paris' Revolutionary papers

until 1777. Those responsible for clandestine papers were persecuted without mercy. During the Revolution, the number of papers in Paris rose to 350; but under the consulate they dropped back to 13 and under the empire to only four. The *Journal des Débats* (1789–1944), a daily founded by François Jean Badouin to report the sessions of the National Assembly, managed to maintain a moderate, liberal line under all regimes. Under Napoleon, *Le Moniteur Universel* (1789–1869), founded by the bookseller Panckoucke, became the official organ. In Germany most cities had incipient newspapers in the early 17th century, but the Thirty Years' War set them back to little more than fly sheets. When they developed again and increased in number, few were of great quality. Among the exceptions were the *Augsburger Zeitung* (1689), the *Hamburgische Correspondent* (1714), and, in Berlin, the *Vossische Zeitung* (1705) and the *Spener'sche Zeitung* (1749), the latter surviving to 1827 as the *Berlinische Nachrichten von Staats und Gelehrten Sachen.* Censorship, though it varied and fluctuated, was, of course, general all over Europe. The first country to have a law guaranteeing freedom of the press was Sweden, as early as 1766. The official Swedish gazette, the *Post och inrikes tidningar,* is also the oldest continuously published paper; it was started as a weekly, the *Ordinarii post tijdender,* in 1645. The world's oldest daily still in existence is usually considered to be the Austrian *Wiener Zeitung,* which goes back, under different names, to 1703.

United States. In North America, where the newspaper was to acquire so much of its modern character, its beginning was entirely typical. In Boston in 1690, Benjamin Harris, a radical from London, issued No. 1 of *Publick Occurrences Both Forreign and Domestick,* intending to continue monthly. It was at once suppressed by the Governor of Massachusetts. Under the colonial governments, there was no question of free printing or even free speech. News was spread via proclamations and pamphlets and by newsletters and papers from London. Though printing had begun in America as early as 1638, the first regular paper did not appear until 1704 and then by the authority of the government. This was the weekly *Boston News-letter,* published by John Campbell, the postmaster, and distributed without mailing charge as a public service. A change of postmaster led to a fresh paper, the *Boston Gazette* (1719), printed by James Franklin, elder brother of Benjamin. A further change led James Franklin to start his own paper, the *New-England Courant* (1721), and this marked the beginning of independent journalism in the colonies. Philadelphia had its first paper in 1719, the *American Weekly Mercury,* founded by Andrew Bradford; and New York City's first paper, in 1725, the *New-York Gazette,* was founded by Andrew's father, William Bradford. The other colonies gradually followed; by the start of the War of Independence there were 37 newspapers in America. The colonial printing office, with its multiple functions, became an important centre of community life, a clearinghouse of information, often set up and maintained under great difficulties. Editors cooperated in passing news to each other and were conscious of their role in creating a historical record.

Establishing the right to criticize government

The right to criticize the government was established in 1735, when John Peter Zenger, publisher of the *New-York Weekly Journal* (1733), was acquitted of criminal libel. From then on, the press became increasingly polemical. After the Stamp Act of 1765, newspapers were sharply divided into patriot and royalist, with passions rising high. The Boston Tea Party (1773) is supposed to have been planned in a backroom of the *Boston Gazette,* nicknamed "Monday's Dung Barge" by the royalists. With independence came the first dailies. The *Pennsylvania Evening Post,* founded by Benjamin Towne in 1775 as a triweekly, became a daily in 1783; and the *Pennsylvania Packet,* founded in 1771 by John Dunlap, became the *Pennsylvania Packet, and Daily Advertiser* in 1784, under Dunlap and David C. Claypoole. The first paper to be founded as a daily was the New York City *Daily Advertiser,* in 1785, by Francis Childs. By the end of the century, dailies were general in the larger cities, smaller towns having weeklies. Burning issues in the new republic soon gave rise to a highly partisan national press. On the side of Hamilton and the Federalists was the *Gazette of the United States* (1789–1818), published by John Fenno in New York City and then in Philadelphia; and on the side of Jefferson and the Republicans (later Democrats), the *National Gazette* (1791–93) of Philip Freneau, followed by the *Philadelphia Aurora,* started in 1790 by the grandson of Benjamin Franklin, Benjamin Franklin Bache. William Cobbett was in exile in the United States at this time and publishing *Porcupine's Gazette* (1797–99), in which he maintained the British standpoint. In the First Amendment to the U.S. Constitution, Congress was restrained from in any way "abridging the freedom of speech or of the press." In 1798, under the threat of war with France and the vitriolic tone of the newspapers, the Alien and Sedition Acts were introduced, which involved an element of censorship. But when these expired in 1801, the right to criticize the government was reaffirmed. Since then the press has been subject only to the law of libel.

Japan. Broadsheets called *yomiuri kawara-ban,* published in the Edo period (1603–1867), are regarded as the traditional forerunners of the Japanese newspaper. *Yomiuri* means "selling by reading out loud," which is what the vendors did, and *kawara-ban* describes printing with engraved tile (engraved woodblocks were also used). The artist-writers who produced these broadsheets usually chose anonymity rather than risk open confrontation with shogunate officials enforcing decrees against public discussion of political and social problems. Many *kawara-ban* featured only popular festivals; personal scandals, notably the double suicides fashionable during the Genroku era (1680–1709); and natural disasters. Some, however, dealt with dramatic events, the oldest known example (1615) describing the Battle of Abeno Osaka between the forces of Tokugawa Ieyasu and Toyotomi Hideyori. *Kawara-ban* conveyed the news of Commodore Perry's arrival (1853) to the Japanese public and were still appearing in the first years of the Meiji era.

19TH-CENTURY DEVELOPMENTS

During the 19th century, the newspaper grew to full independence in most literate countries, beginning with Britain and the U.S. At the same time, by broadening its appeal and reducing its price, it began to reach the mass of the population in a way that no other reading matter had ever done before. As a result, it became big business, largely through the value of its advertising revenue.

Technological advances. The impact of technical developments was enormous and many-sided. At the start of the century, when circulations were measured in thousands, newspapers were printed, like books, on the hand-operated, flatbed wooden press. By the end, rising circulations and the drive for ever greater speed had led to the development of highly specialized, reel-fed (web) rotary machines, driven by steam (after 1814) and later electricity (after 1884), which produced hundreds of thousands of cut and folded copies in a matter of hours. Behind this evolution lay numerous other, equally necessary inventions: methods of stereotyping, special inks and inking devices, cheap cellulose paper, and mechanical typecasting and typesetting.

No less startling was the transformation in news-gathering techniques, which occurred because of the invention of the telegraph and the telephone. In 1815, when the mounted courier was the chief means of getting quick news, it was four days before the Battle of Waterloo was reported in London, only 240 miles (390 kilometres) away; and that was a record, set up by the *Morning Chronicle.* After the laying of cables (*e.g.,* Dover–Calais, 1851; transatlantic, 1866), the time lag all but disappeared. Finally, the distribution of the daily newspaper on any great scale would have been impossible without a general speeding up of transport, especially through the railway.

Growth of an independent press. In Europe, the first signs of a move toward popularization appeared in France, with the founding of *La Presse* (1836) by Émile de Girardin, who might be called one of the earliest press magnates. He introduced new methods, serials, and "features" to raise circulation (to as much as 20,000) and to bring down the price of his papers. Also influential in the first half of the century was Louis Désiré Véron, who founded ***Revue de Paris*** (1829), a magazine to encourage new writers, and revived the liberal daily ***Le Constitutionnel*** (1835). French journalism, however, was already characterized by its literary slant; it was to develop little social responsibility until much later. Characteristic, too, was the signed article, made compulsory by the Tanguy Law of 1850, which gave the individual journalist greater prominence than he received elsewhere. In Germany, the press was still severely restricted; but a distinguished daily of some independence was the *Allgemeine Zeitung*, founded in Tiibingen in 1798 by Johann Friedrich Cotta, the publisher of Schiller and Goethe. It soon got into trouble with the authorities and had to move; but it eventually came to rest in Augsburg, where it achieved wide influence. Scandinavia continued to lead the Continent in press freedom. Sweden had long had favourable laws (1766); but its modern press is usually dated from 1830, when Lars Johan Hierta founded *Aftonbladet*, a liberal organ combining good news coverage with lively comment and human-interest material. Denmark enjoyed a similar freedom after 1849, when the nation received a new constitution, which included among its articles the abolition of censorship. Switzerland, too, abolished censorshio with its federal constitution of 1848. Elsewhere in Europe, revolution or threat of revolution caused the authorities to keep the press on a tight rein.

Success of *The Times*

Even in Britain, where the press had largely won its freedom, fear of the spread of revolutionary ideas prompted an increase in the Stamp Tax, in 1815, and the passing of the Six Acts of 1819, considered a high point in legislation restricting the press in England. But despite prosecutions for sedition and blasphemy, publications obnoxious to authority continued to appear; and William Cobbett's radical weekly, ***The Political Register*** (1802–35), had a wide circulation; Cobbett even managed to issue it from prison. The "respectable" press was naturally less affected; and one paper, ***The Times,*** rose to great heights, dwarfing all others. In the 19th century, under the management of its proprietors, John Walter II and John Walter III, and the editorship of two outstanding journalists, Thomas Barnes and John Thaddeus Delane, it kept free of all subsidy and political pressure and became a model of what a responsible, independent newspaper should be: quick and reliable with the news (during the Napoleonic Wars it was often ahead of the government), influential in opinion (it opposed slavery and backed reform), and technically well produced (it led the way in the use of new machinery). Its circulation rose from 5,000 in 1815, when it cost seven pence, to 50,000 in 1850 (price five pence), at a time when its nearest rivals, the ***Morning Advertiser*** and the ***Daily News,*** were selling only 5,000 copies and other dailies a mere 2,000 to 3,000. Its appeal was still limited, being confined to the mercantile middle class, which was just coming into its own; but it demonstrated for the first time the absolute value of journalistic independence.

On the other side of the Atlantic, an independent press developed even more strongly, and there it appealed from the beginning to the mass of the people. In the first part of the century, American newspapers were fiercely partisan and political, and presidents made early and ample use of them; Pres. Andrew Jackson had 60 full-time journalists on the government payroll. It was largely disgust with this party involvement that led James Gordon Bennett, in 1835, to found the *New York* ***Herald,*** with which modern American journalism may be said to begin. It was not the first of the cheap, popular papers, the so-called Penny Press, for Benjamin H. Day had successfully exploited the human-interest formula with his New York *Sun* (1833); but the ***Herald*** was the first to proclaim and maintain complete independence. As Bennett wrote in his prospectus, his paper would support no party; it would endeavour to record facts, "with comments suitable, just, independent, fearless and good-tempered." His public was the teeming immigrant population of New York City, whose significance he grasped and whom he presented with an exciting amalgam of news and views in an unvarnished language they could understand and at a price they could afford. In little over a year, the circulation of the ***New York Herald*** was running at 30,000 to 40,000; and as it prospered, Bennett led the way in rapid news-gathering and efficient production methods. His nearest rival was Horace Greeley, who founded the ***New York Tribune*** (1841). Both men came up the hard way, but where Bennett was primarily an entertainer, Greeley was an idealist and crusader, a fierce opponent of slavery and coiner of the phrase "Go West, young man!" The weekly edition of the ***Tribune*** had a great influence beyond New York City. Bennett and Greeley, with their personal brand of popular journalism, dominated the New York City press for 30 years; but a third paper, of equal independence though less startling character, was the *New* ***York Times*** (1851), co-founded and first edited by Henry Jarvis Raymond. There were, of course, many papers outside New York City, notably the ***National*** *Intelligencer* (1800) in Washington, D.C., the ***Philadelphia*** *Public* ***Ledger*** (1836), and the ***Baltimore Sun*** (1837); by 1850, with the westward expansion of the country, the number of dailies had risen to nearly 400, and the weeklies were correspondingly more numerous. Though the majority of papers continued to be party political organs, Bennett and Greeley and those who followed them had established a firm tradition of journalistic independence.

The newspaper reporter

News gathering. In the early days of journalism, news was gathered by a more or less anonymous system of newsletter writers, agents, and couriers; the key men were inclined to be those at home, who offered their comments on events. As newspapers grew, however, staffs were built up of full-time reporters who were sent out to get the news. This development was greatly stimulated by the several wars of the 19th century, which necessitated special correspondents and drew attention to them. One of the earliest and greatest was William Howard Russell, correspondent of the London ***Tintes*** in the Crimea during the war of 1853–56. His dispatches exposed the scandalous conduct of the war and moved Florence Nightingale to take up her mission. Similarly, the U.S. Civil War (1861–65), in which more than 150 correspondents followed events in the field, gave the reporter a new status. His emergence as something of a folk hero, the man on the spot fearlessly finding out facts in the name of democracy, may be said to date from this period. As newspapers became increasingly competitive, especially in their efforts to "scoop" each other—*i.e.*, be the first to publish an item of news—the reporter more and more replaced the editor in the popular imagination.

Development of Reuters and other agencies

At the same time, the cost of large-scale news gathering, prohibitive for the smaller paper, led to the rise of news agencies, especially after the invention of the telegraph. The earliest of note was founded in Paris, in 1835, by Charles Havas, who bought the Correspondance Garnier, mainly a translating office, and turned it into an agency culling extracts from the chief European papers for the French press. Before long, he extended his business to other subscribers, ran a regular carrier-pigeon service between London, Brussels, and Paris and had correspondents in most European capitals. In 1856, an advertising agency was added, news being paid for with space, which put the Havas Agency in a very strong position. In 1849, the opening to the public of the Prussian State telegraph line from Berlin to Aachen prompted a former Havas employee, Bernhard Wolff, to start the first telegraphic bureau in Berlin. Commercial news, stock-exchange prices, and market rates were the main elements in this service to begin with, as they had been in the early newsletters—*e.g.*, of the Fuggers. Another former Havas man, Paul Julius Reuter, saw his opportunity at the other end of the line, in Aachen, and ran a small office there before moving to England in 1851. In London, Reuter's first venture was, again, to supply the city with overseas commercial information; but in 1858 he was ready to inaugurate his famous service of foreign telegrams to the press. This was steadily expanded, along with British influence and the cable system, until it covered a large part of the world. Initially, these three agencies, Havas, Wolff, and Reuters, struggled fiercely for territorial news monopolies; but in 1870, in a remarkable series of Agency Treaties, they agreed to divide the world among them into exclusive preserves, with Reuters getting the lion's share. The smaller European agencies, such as Stefani (Italy, 1854), Korrespondenz-Bureau (Austrir, 1857), and Ritzaus (Denmark, 1866), all fell into line. In the U.S., meanwhile, a very different type of agency had arisen. The cost of covering the war with Mexico (1846--48) inspired six New York City newspapers to form a cooperative news-gathering organization, the New York Associated Press, which eventually, after major changes in 1892 and 1900, became the modern Associated Press. In 1893, this agency, too, lined up with the European triumvirate in a mutual agreement with Reuters for the exclusive supply of European news to America and American news to Europe. The Associated Press was an important development because it was the first agency run by newspapers themselves, which proved in the long run to be the most desirable arrangement. Naturally, newspapers continued to maintain correspondents and roving reporter's according to their resources; but the news agencies greatly helped the worldwide flow of "spot news"—*i.e.*, the bare facts about events as they occur.

Journalistic changes in the United States. In the second half of the 19th century, the press of the United States, in keeping with American society, expanded with enormous vigour. Between 1850 and 1880, the number of papers more than doubled (to about 850), and by 1900 they had done so again (to over 1,950); a high proportion, up to two-ihirds, were now afternoon papers. The press became the great means by which the American way of life was brought to the heterogeneous, often illiterate masses of immigrants. Its keynotes were sensationalism, to ensure that it was read, and idealism, to right wrongs and build civic and national consciousness. Typical of the sensationalism in the Middle West was the *Chicago Times* (1854) under Wilbur F. Storey (after 1861); and of civic consciousness, the *Chicago Tribune* (4847) under Joseph Medill (after 1855) or the *Kansas City* (Missouri) *Star* (1880) under Rockhill Nelson: who used the weapons of investigation and exposure to check corruption and promote good government. In the South, a new outlook was needed after the desolation of the Civil War; and here again newspapers helped to create it, notably the *Atlanta* (Georgia) *Constitution* (1868) under Henry W. Grady (after 1880) and the *Louisville* (Kentucky) *Courier-Journal* under Henry Watterson (after 1868). Among the many flourishing New York City papers, the Sun rose in style and prominence under Charles A. Dana (after 1868); the *Times* struck its own crusading note by exposing "Boss" Tweed of Tammany Hall, in spite of his attempt to bribe the editor with $5,000,000; and the *Evenirig Post* acquired great infiuence tinder a scholarly editor, Parke Godwin.

Joseph Pulitzer and the "new journalism"

Then, in 1883, the press was carried to fresh heights by Joseph Pulitzer, a Hungarian immigrant whose name in particular is linked with what is known as the new journalism. He began in St. Louis in 1878 by merging two papers into the *St. Louis Post-Dispatch* and turning it into a vigorous crusading journal. In 1883, he bought the failing *New York World* and in threo years raised its circulation from 15,000 to 250,000, to make it the most successful paper ever. He did so by taking the old blend of sensationalism and idealism and presenting it with new vitality in an endless series of stunts and campaigns. The press, he believed, should always be devoted to the public welfare, drastically independent and never afraid to attack wrong.

Exciting journalism, it would seem, could go no further; yet it was given a final, often excessive twist by William Randolph Hearst, who came from the far West, where journalism was at its brashest and toughest. Hearst ran the *San Francisco Examiner* (after 1880) and in his early days did his share of cleaning up corruption through exposure. By the rime he came to New Y'ork City, however, he was more interested in sensation at any price. He bought the *Morning Journal* in 1895 and within a year was challenging. though not exceeding, the *World* in circulation. Hearst's style of journalism, based on scare headlines, plenty of pictures, pseudoscientific articles, Sunday supplements with comic strips, and ostentatious campaigns, came to be known as yellow journalism, from a comic picture series called "The Yellow Kid," which played a part in the rivalry between the *World* and the *Journal.*

In the competition for circulation some of the other papers naturally suffered, and among them was the *Times.* In 1896, however, it was taken over by Adolph S. Ochs and re-established as the predominant serious journal with the slogans, "All the news that's fit to print" and "It will not soil the breakfast cloth." Many of the techniques of yellow journalism became permanent and spread far and wide, but the comeback of the *Times* could be said to mark the end of the era of highly personal journalism begun by James Gordon Bennett.

Expansion in Britain. In Britain, the press was somewhat slower to expand; but a major turning point came in 1855, when the stamp tax was abolished. It had been reduced to one penny in 1836; and this had released a flood of cheap crime sheets and a number of sensational Sunday papers, such as the *News of the World* (1843), whose great popularity dates from this period. The final abolition of the stamp tax and that of the other "taxes on knowledge" (the advertisement duty in 1853 and the paper excise in 1861) led to an immediate upsurge in the number of dailies. The tax was ended partly as a result of a campaign for cheaper papers but equally because the authorities could no longer endure what Lord John Russell called "the vile tyranny" of *The Times.* Among the new journals was London's first penny paper, the *Daily Telegraph* (originally the *Daily Telegraph aad Courier,* 1855). By 1861, it had a circulation of 130,000 (double that of the *Times*), and by 1870 this had risen to 240,000, then the largest in the world. Yet the *Daily Telegraph,* despite a rather more colourful style, was still not a popular paper in the American sense; its appeal was to the growing middle class, for whom it assembled a wealth of writing talent. The new editors followed Barnes in upholding journalistic independence; but in the age of Gladstone and Disraeli, when party lines grew sharper, they tended to support one side or the other.

American methods had as yet made little impact, but after about 1880 their influence became more noticeable. The evening *Star,* launched in 1888 at a halfpenny under T.P. O'Connor, aimed at a wider public, with bold display, short, lively items, and good racing tips; and another evening paper; the distinguished *Pall Mall Gazette,* adopted American tactics for some of its crusades. Popular journalism in Britain really begins in 1896. however. with the *Daily Mail,* founded by Alfred Harmsworth, later Lord Northcliffe. Harmsworth came into journalism through cheap magazines, which showed him the potential size of the new market and how to exploit it. In 1895, he bought the *Evening News* and by adopting a formula of short, snappy paragraphs quadrupled its circulation within months The *Daily Mail,* designed on the same pattern, was even mole successful; its circulation soon rose to 500,000 and to 1,000,000 at the time of the Boer War (1899–1902). It called itself "A Penny Paper for One Halfpenny" and "The Busy Man's Daily Journal." All its numerous features were broken down to conform to Harmsworth's cardinal precept: "Explain, simplify, clarify." Thus with Narmsworth, the new journalism could be said to have arrived in Britain but in muted form. The *Mail* was not so sensational as its American counterparts; it had no social conscience for crusades, and its public was not the real mass of the people but the aspiring lower middle class, whom it sought to inform quickly and painlessly Commercially, the *Daily Mail* was a portent, for it was the first British paper to be based deliberately on advertising. It cost mole than a halfpenny to produce, but by publishing net sales certificates (the first to do so), it gave advertisers concrete evidence of their possible market and enabled 'the management to charge accordingly—and to make handsome profits for

Popular journalism in Britain

its shareholders. The net sales certificate was a statement of sales based upon the number of copies printed less free and unsold copies as certified by a chartered accountant or professional auditor.

Developments in Europe. Between 1850 and 1900 the rest of Europe also moved toward journalistic independence and a popular press, though by no means uniformly. The continental press as a whole was inclined to remain more solidly middle class in its orientation and to favour the journal of opinion rather than that of information. Important papers founded were ***Le Figaro*** (1854, weekly; 1866, daily) and ***Le Temps*** (1861) in France; the ***Frankfurter Zeitung*** (1856) in Germany; and the ***Corriere della Sera*** (1876) in Italy, all of which marked progress toward modern concepts in journalism, made possible by improved political conditions. The popular press in France may be said to begin with La ***Petite Presse*** and ***Le Petit Journal,*** which had a circulation of 650,000 in 1878, based on a lively approach and an attacking style. Between 1880 and 1900 the number of newspapers in France almost doubled (to 2,400), with several strong provincial journals. Other European papers in a more popular vein were ***Aftenposten*** (1873), in Denmark, and ***Le Peuple,*** in Belgium.

In Spain and Portugal, censorship continued to prevent the development of valid journalistic independence; any periods of comparative freedom would be quickly followed by the reimposition of controls. In Russia, too, censorship remained in force. Alexander II allowed many new papers to be started between 1855 and 1865 but soon brought them under control when radical ideas began to be voiced; as a result, they became predominantly literary in tone. Outside the country, however, revolutionary Russian journals were published in a number of cities, notably ***Kolokol*** ("The Bell"), founded in London in 1857 by Aleksandr Herzen, first as a monthly, then as a fortnightly. These are regarded today as the forerunners of the modem Soviet press.

Newspapers in Russia

Japan. Japanese historians differ as to the first modern Japanese newspaper published by a Japanese. Some choose the ***Batabia shimbun*** (January 1862), 23 volumes of translated abstracts from the ***Javansche Courant,*** a Dutch government organ in Batavia (Djakarta), published by the shogunate "Office for Reviewing Barbarian Papers." This office subsequently published translations of newspapers from the China port cities, Hong Kong, and the United States. The status of these official publications as newspapers is questioned because of their lack of periodicity.

The claim of the ***Kaigai shimbun*** (1865–66; "Overseas Press"), published monthly for about a year, to be the first Japanese newspaper is also questioned because its publisher Joseph Heko, although native-born, had become a naturalized U.S. citizen (he was then an interpreter at the United States embassy).

During 1867–68, when the shogunate collapsed and the Meiji restoration was consolidated, more than a dozen newspapers were started by shogunate sympathizers. Among them was the ***Chugai shimbun,*** published by a group' of Japanese linguistic scholars who had worked on the ***Batabia shimbun.*** Another was the ***Koko shimbun,*** whose publisher, the dramatist and educator Fukuchi Genichiro, had studied Western newspapers while on government missions abroad.

The Meiji government suppressed these publications and promulgated a Newspaper Ordinance, which, in its 1871 version, established the principle that the contents of a newspaper should always be "in the interest of governing the nation." By 1876 the government had begun suspending or suppressing newspapers that displeased it and arresting their staffs. Many of the early Meiji newspapers consequently tended to be closely tied to the political world.

The first Japanese daily

The first Japanese daily newspaper was the ***Yokohama Mainichi shimbun*** (1870); it was among the first to use lead type. Some Japanese historians, however, regard the ***Tokyo Nichi-Nichi shimbun*** (1872), the first daily in the new capital, as being the first truly modem Japanese newspaper. The one-time publisher of the ***Koko shimbun,*** Fukuchi, became its president in 1874 and introduced an editorial column modelled on Western newspapers. The ***Nichi-Nichi,*** however, regarded itself as virtually an official gazette.

Two giants of contemporary Japanese journalism, the ***Asahi shimbun*** and ***Mainichi shimbun,*** were founded in the commercial centre of Osaka and placed more emphasis on general reporting than on political affairs. Murayama Ryuhei, scion of a warrior family who was invited to become president of ***Asahi,*** and the industrialist Motoyama Hikochi, who developed the ***Mainichi,*** are commonly regarded as the fathers of modern Japanese journalism. The ***Asahi*** began as the ***Osaka Asahi*** in 1879; the ***Tokyo Asahi*** was founded in 1888. The ***Osaka Mainichi*** developed from a foundering newspaper venture begun as the ***Osaka Nippo*** in 1876; the name ***Mainichi*** dates from 1888 when the paper became an organ of Osaka business interests. At first an adviser to the venture, Motoyama became president in 1903. In 1906, he added the ***Tokyo Nichi-Nichi*** to his publishing empire.

The third leading national daily of contemporary Japan, the ***Yomiuri shimbun,*** was founded in Tokyo in 1874 and gained a reputation as a "literary" newspaper. In 1924, it was acquired by a former police official, Shōriki Matsutaro, who added such innovations as entertainment sections, colour comics, and Hearst-style promotions.

Developments in other areas. In several other parts of the world, approaches to a modern press developed before the end of the century. Australia, whose first newspaper, ***The Sydney Gazette and New South Wales Advertiser,*** commenced publication as a weekly on March 5, 1803, went through the familiar pattern of censorship (ended 1824) and stamp tax (ended 1830) before being able to produce independent papers; but newspapers grew with the country, notable among them being the ***Sydney Morning Herald*** (1831), the Melbourne ***Argus*** (1846), ***The Age*** (1854), and the more popular ***Sydney Daily Telegraph.*** In South Africa, too, a fight had to be waged for press freedom by the editor of the country's first paper, the ***South African Commercial Advertiser,*** who secured *a* new Press Law in 1828. Later papers, such as the ***Cape Argus*** (1857), were tied to commercial and mining interests; but they gradually acquired some independence through the necessities of their existence.

The Australian press

The first issue of New Zealand's earliest newspaper, the ***New Zealand Gazette,*** was printed in London in 1839 before emigrants left for the island. The second issue did not come out until April 18, 1840, but was printed in Port Nicholson (now Wellington). Two months later, the ***New Zealand Advertiser*** appeared in the Bay of Islands. The oldest newspaper still publishing is the ***Taranaki Herald,*** which began in 1852; more than 20 New Zealand newspapers had celebrated centennials by the early 1970s.

Canada had its first newspapers, French and English, in the 18th century, and these developed regionally as had occurred in the U.S., though on a smaller scale, in Quebec, Montreal, and Toronto. A leader in journalistic standards was the ***Winnipeg Free Press*** (1872). In India, newsletters circulated as early as the 16th century. Under British rule, both English- and vernacular-language newspapers were begun, though the latter were subject to control and suppression. The ***Bombay Times*** (1838, later the ***Times of India)*** is usually considered the chief paper in English, while ***The Hindu*** (1878) was important for the growing national consciousness of Indians. In Argentina, finally, one of the world's greatest newspapers, La ***Prensa,*** was founded in 1869 by Dr. José C. Paz, whose descendants still own and edit it. With the watchwords, "Truth, Honour, Liberty, Progress, Civilization," it has long upheld the highest standards of journalistic independence.

The significance of the 19th-century independent newspaper was, of course, enormous. It provided a great sounding board through which political and social ideas could crystallize and become effective; its advertisements and steady supply of information oiled the wheels of commerce; and at the most popular level it became a medium of education and entertainment in a thousand and one ways. Reference has been made to the political influence of ***The Times*** in Britain and to the muckraking crusades of many American newspapers. In Britain muckraking was inclined to be more sober but nonetheless effective. As an example, one might cite the series of articles "The Maiden Tribute to Modern Babylon" by W.T. Stead in the ***Pall Mall Gazette.*** In these Stead exposed the London trade in young girls by himself procuring one and serving a term in jail. The result was a change in the law, the passing of the Criminal Law Amendment Act (1885). On occasion the press has been at the very centre of national events. Such a case was the Dreyfus Affair in France, in which Émile Zola's famous letter '*J'accuse,'* published in Georges Clemenceau's *L'Aurore* (January 1898), laid bare the political realities in such a way that they demanded debate by the whole nation. Zola was convicted of libel and fled the country. Finally, nothing better illustrates the power of the press, the misuse of that power, and its ultimate limitations than the notorious case of William Randolph Hearst and the Spanish-American War over Cuba in 1898. For months before America declared war, Hearst's ***New York Journal*** stirred the country to a high pitch of hysteria with exaggerations and even complete fabrications. Hearst is reported to have cabled his artist in Cuba, who found no atrocities to illustrate, "Please remain. You furnish the pictures and I'll furnish the war." Yet this he was, in fact, unable to do until events themselves brought it about, especially the blowing up of the American battleship "Maine" in the harbour of Havana with great loss of life. Hearst nevertheless claimed "credit" for the war in a banner headline: "How Do You Like the ***Journal's*** war." It was quite clear by the end of the century that for good and ill the Fourth Estate had arrived.

Hearst and the Spanish-American War

THE PRESS IN THE 20TH CENTURY

Since 1900 the number of newspapers in the world has risen rapidly, and every conceivable stage of press development and type of journal can be found in one country or another, from state-controlled propaganda organs on the one hand to a more or less free press on the other and from small sheets in areas of low literacy to the huge newspapers of a large industrial society. The methods and problems of producing a paper have become correspondingly diverse. Where the methods are most advanced, in the United States, Britain, western Europe, and Japan, the press has become big business and its problems largely financial, centring on advertising revenue and the struggle for circulation.

Technical developments. Technical development has continued to refine and speed the process of production, from the gathering of the news to the printing of the paper. The rotary presses turn even faster, thanks to mechanical improvements, quick-drying inks, and such devices as automatic ink-pumping (1915) and a system of reel joining to provide a continuous flow of paper. High speeds mean, of course, that larger editions can be rushed through the machine rooms in shorter time, and thus copies printed as late as 4:00 AM can still reach breakfast tables many miles away. The city (early) edition of the ***New York Times,*** published in Manhattan, could, until the establishment of a regionally printed national edition in 1980, be purchased in Chicago and on the U.S. West Coast, 3,000 miles (5,000 kilometres) away, on the day of publication.

Economic aspects. ***Labour.*** The increased amount of money in the newspaper industry led to a sharp rise in the pay and status of the journalist, and the growing strength of printing trade unions saw to it that their members were not neglected. Higher wages for those whose skills operated typesetting machines fast and accurately, cast the curved stereoplates without fault, and manned the giant presses at midnight were only to be expected; a much more serious matter was the establishment in succeeding years of systems of restrictive practices, under which precise limits were laid down within which trade union members should work. Any form of extra cooperation or helping out of a workmate was frowned upon and could be the subject of a fine or even a strike. The practice resulted in considerable overstaffing in newspaper printing departments—*i.e.,* the employment of a much larger staff than was necessary to produce the paper efficiently; it inflated costs and was to become a major cause of the financial difficulties that many daily papers found themselves facing in the mid-century. It has been argued that restrictions that were imposed by the unions with the object of creating more jobs for their members in fact had a reverse effect owing to the economic collapse of a number of once-powerful papers in Great Britain and the United States.

Trade unions

Consolidations and chains. As financial interests moved into the press, in place of the old editor-proprietor, the trend was toward consolidations, groups, and the growth of the newspaper chain. Its earliest, but least mercenary, exponent was E.W. Scripps, who began collecting American newspapers in 1878. His main technique was to buy up small, financially insecure papers and set them on their feet by putting in capable young editors, who were given a share of the profits. He always urged them to remember that their first duty was "to serve that class of people and only that class of people from whom you cannot even hope to derive any other income than the one cent a day that they pay for your newspaper." Scripps wanted his papers to be of genuine service to their communities, and though he succeeded in making money from them, his aim was not to extract maximum profits. He did, however, utilize to the full the benefits of large ownership to enable him to buy newsprint and syndicated articles on the most favourable terms. Many of Scripps' methods were adopted by his rivals and by newspaper proprietors in other countries, as the idea of the chain spread with the more commercial approach to newspapers. In the 1920s in the United States there were some 50 chains embracing about 300 papers, which meant that anyone moving about the country was liable to encounter the same news stories and features, all drawn from the same source. William Randolph Hearst, who at one time controlled 42 papers, was among the most acquisitive of the early chain owners; but the purely commercial aspects of proprietorship were epitomized, between 1916 and 1924, by Frank A. Munsey, who bought and merged newspapers, including the New York ***Sun*** and the ***Herald,*** with "the talent of a meat packer, the morals of a money changer and the manners of an undertaker," according to the eminent U.S. editor and author William Allen White.

The Scripps papers

Changes in newspaper content. ***Minority group newspapers.*** The U.S. press had long included a large number of papers for minority groups, especially in foreign languages. The first of these, the ***Philadelphische Zeitung,*** was published by Benjamin Franklin as early as 1732, though only two issues appeared. It was soon followed by others, among them the first foreign-language daily, the ***Courrier Français*** (Philadelphia), which supported the French cause in 1794–98. With successive waves of immigration, especially from Germany in 1830–60, the number of such papers rose phenomenally, until they amounted to 10 percent of all newspapers in 1860. Two-thirds of them were in German, others being in French, Italian, Spanish, Dutch, Swedish, Norwegian, and Welsh. This publishing activity reached its peak between 1900 and 1914, with the addition of papers in many more languages; thereafter, as immigration rates fell and assimilation became more complete, it steadily declined in importance, without by any means vanishing. Another minority press, on the other hand, that of blacks, has grown enormously, especially since World War II. Its origins go back to ***Freedom's Journal*** (New York City, 1827), one of several short-lived anti-slavery papers that sprang up in the period before the Civil War. In the late 1970s, 165 black newspapers were being published, with a total circulation of almost 3,000,000.

Tabloids and their effect. A major innovation that appeared at the beginning of the century was the tabloid—*i.e.,* a newspaper characterized by half the usual page size, abundant pictures, and text correspondingly reduced and condensed. The term was coined by Alfred Harmsworth when he designed and edited an experimental issue of the ***New York World*** for New Year's Day in 1900. The United States, however, ignored the new form for a time, and the first tabloid was Harmsworth's own ***Daily Mirror***

(1903). This was originally founded, in the usual format and style, as a newspaper for "gentiewomen"; when it failed as such, it was quickly revived with the help of a new technique for reproducing photographs as a halfpenny picture paper. Its success was immediate, with circulation running at over 1,000,000 by 1914. A second British tabloid, which proved equally successful, was the *Daily Sketch.* founded by Sir Edward Hulton, in Manchester, in 1909.

The first tabloid in America was the *New York Daily News* (1919), started by Joseph Medill Patterson to a ruthless recipe of sex and sensationalism, which put it in the lead and kept it there against all comers. Its chief rivals were the *Daily Mirror* (1924), siarted by Hearst, and the *Daily Graphic* (1924), begun by Bernarr Macfadden. These three fought a veritable "war of the tabs" in New York City, with the *Graphic* perishing in 1932. The *Mirror* survived until 1963, when it was swallowed by the *Daily News*. The tabloid has long since proved its convenience and popularity acd become a widespread alternative format.

World War I and after. During World War I, the press suffered severe restriction not only through a sharp decline in the quantity and quality of newsprint in Europe but above all through censorship. In Britain, under the Defence of the Realm Act, the British Official Press Bureau was set up within a week of the outbreak of -war, and it continued to function until April 1919. Its nominal purpose was to conceal naval and military intelligence from the enemy, which was naturally accepted by the press: but for want of clear principles, it operated unevenly and was used all too often to suppress opinion and save the face of the politicians. Over the vital matter of the shell (artillery ammunition) shortage, it was boldly defied by Lord Northcliffe, now owner of *The Times* as well as of the *Daily Mail*, who thereby rendered the country a great and necessary service, illustrating afresh the true responsibility of the press. After the United States entered the war, censorship became a little more rational, with some movement toward a self-imposed code. In the United States, more than 75 papers had their mailing privileges withdrawn under the Espionage Act, and the German-language press shrank to half its former size.

Struggle for circulation. After the war, the struggle for circulation became even more intense, especially in the U.S. and Britain. It had now become a matter of life and death; for circulation determined the advertising revenue that made a newspaper financially viable.

Political columnists

In America, readers were constantly wooed with better coverage, more pictures, new features, ever larger papers. New York dailies grew to 40 or even 60 pages, while the Sunday edition of the *New York Herald-Tribune* (merged in 1924) often had 100-140 pages, divided into various sections covering different affairs. The personal impact of the old editors had long since gone, but in its place came that of the new opinion makers, the columnists—*i.e.*, men and women who wrote regular columns of views, comment. and reflection or mere social gossip. The first political columnist was David Lawrence of the *United States News* in the 1920s. Other pioneers were Frank Kent of the *Baltimore Sun* and Mark Sullivan of the *New York Herald-Tribune*. One of the best known was Walter Lippmann, who originally wrote in the New *York World*. Such columns were often syndicated in newspapers all over the country; when this occnrred, the columnist acquired a considerable national following. Syndication of material became a growing practice, either through the news agencies, the picture agencies, and the newspaper chains or through such special feature syndicates as the Newspaper Enterprise Association (NEA), which supplied ready-to-use material of all kinds, from comic strips and crossword puzzles to editorials, book reviews, and medical columns. This service enabled a small-town editor to assemble a respectable paper without a large staff: he needed reporters only for local affairs. Despite the "jazz journalism" of the 1920s, in the tabloids above all, standards of serious news coverage continued to rise, led by the *New York Times*. Such a small-circulation daily as the *Christian Science Monitor* (1908), for instance. gained a high reputation, especially for foreign affairs. A characteristic of the American press, which went back lo the previous century, was the predominance of the afternoon paper, which tended to have the largest circulation. Typical of the time was the. purchase of the New York *Morning*, *Evening*, and *Sunday World* in 1931 by the Scripps-Howard chain: the morning and Sunday editions were. killed, and the *Evening World* was merged with the *New* York *Evening Telegram*. Many such mergers were imposed by economic pressures: the trend toward fewer papers with higher circulations, both individually and collectively, was well under way.

Circulation-boosting stunts

In Britain, a new force appeared after World War I in the person of an iconoclastic Canadian, Max Aitken, created Lord Beaverbrook in 1917 for political services. At about the same time, he acquired the *Daily Express*—which had been limping along since 1900. when it was founded by Arthur Pearson—-and made it Britain's second popular daily through sheer journalistic panache. With his two other papers, the *Sunday Express* (started 1918) and the *Evening Standard* (acquired 1923), he campaigned tirelessly for Empire Free Trade but without great result, showing again the limited influence of the press when it is not in accord with the public will. Among the many circulation-boosting stunts of the 1920s was free insurance for readers, which proved so costly, however, that it had to be abandoned. Another was free gifts, started by the *Daily Herald* in 1930. Despite the condemnation of the Newspaper Proprietors' Association, gift schemes proliferated until the *Herald*, by spending most, achieved a circulation of 2,000,000, the first paper in the world to do so. In this fierce competition, there were casualties, of course; and by 1934, even the popular London tabloid the *Daily Mirror* was failing, with circulation down to 700,000. But it was rescued in striking fashion by Harry Guy Bartholomew, who gained editorial control from a divided board of managers and transformed it into a true working-class paper, the first in Britain, with a radical voice, all the most outré tricks of yellow journaiism. and a compelling strip cartoon about a young lady called Jane. He eventually raised its circulation to.over 4,000,000, the largest in the world. At the same time. Britain's first "popular" paper, the old *Daily Telegraph,* now settled down as a conservative middle class journal (owned by Lord Camrose), absorbed the *Morning Post* in 1937, and rose through solid merit from a circulation of 100,000 in 1927 to 750,000 in 1939. For all the shoddiness of much of the journalism of the 1930s, readership of the popular newspaper.; rose by 1,500,000, and the British press could be said to reflect at last the full social pyramid, including the underprivileged at its base.

Newspaper organization. As the newspaper grew in size and complexity, so too did its organization, which had now assumed its modern form and often required a vast city block or skyscraper to house it. Apart from the business side, which takes care of advertising and circulation, and the mechanical side, which includes composition (*i.e.*, typesetting and correction), engraving, stereotyping (*i.e.*, making duplicate nlates of the printing surface), and the actual printing, the editorial function itself had expanded greatly from its original simplicity. Under the editor in chief and his assistants, responsible for the editorials and overall policy, there are now usually a number of editors in charge of particular departments, such as the news editor, who assigns stories to reporters and rewrite men; the picture editor, with a staff of photographers; the features editor, responsible for the supply of pieces and columns of an entertaining or instructional nature; the sports editor, who usually has his own reporters and photographers; the society editor: the financial editor; the women's editor: the correspondence editor; and many other specialists, covering the arts and such subjects as science and education, with a thoroughness appropriate to the resources of the paper and the interests of its readers. An ever-growing volume of information flows into a newspaper office, from agencies and bureaus, from government departments and foreign embassies, from organizations and individuals of every kind. All has to be assessed, selected, checked, and amplified, if necessary, from the paper's reference library and "morgue" of past material, with a constant watch for possible libels. Finally, before the stereotypes are cast and the paper is "put to bed," each page has to be made up in characteristic style, with the blend of headlines, news, pictures, features, and advertisements that the reader has grown to expect.

Breaking the European news monopoly

News agencies. Since the Agency Treaties of 1870, to which the U.S. Associated Press acceded in 1893 (see above), the international flow of news had remained largely in European hands, dominated by Reuters. During World War I, however, the possibility of national bias became very apparent, though Reuters itself, a private profit-making company, kept a remarkable degree of independence. As a result and as part of the general expansion of American influence after the war, the so-called European news monopoly began to be strongly challenged by the Associated Press, under the inspired leadership of its general manager, Kent Cooper. After a bitter struggle, which came to a head over Japan, a "Reuters territory," a new agreement was signed in 1934 that ended the old exclusiveness. This represented a great advance in international news gathering, for world news could be received from more than one source. The free interchange of news was seen, largely as a result of Kent Cooper's crusade, to be an essential safeguard against governmental control. Reuters eventually became a newspaper-owned trust (1941), with articles guaranteeing its integrity and independence; but despite a renewed campaign for freedom of news after World War II, many national news agencies continued to be state owned or controlled.

Press freedom. In the years leading to World War II, the European press reflected all too faithfully the unhappy state of the Continent. In Germany, Italy, Spain, and Portugal it was subject to censorship and made to serve the ends of the government, as it was also, in a different cause, in the Soviet Union. In France, where the press was more or less open to the highest bidder, articles might be inspired from any quarter prepared to bribe or subsidize; and in Britain it fell far short of its old traditions. ***The Times*** under Geoffrey Dawson became an instrument of government policy, and the ***Express*** maintained a fatuous optimism; apart from the muted tones of the ***Manchester Guardian*** and the ***News Chronicle,*** only the raucous voice of the ***Daily Mirror,*** to its credit, pointed out unpalatable truths. In these circumstances, where the main press was failing in its duty, there was an interesting return to the ancient newsletter. Two in particular, the sober ***News Letter,*** issued by Stephen King-Hall, and the more highly flavoured, somewhat Marxist ***The Week,*** put out by Claud Cockburn, were influential.

Wartime censorship

World War II and after. In World War II the press was handled far more sensibly than in World War I, both in Britain and in the United States. In Britain, apart from censorship in the field and any temporary embargo on items of news, there was no compulsion, nor was comment restricted. Newspapers voluntarily submitted any doubtful material to the censorship department, thus clearing themselves in the event of any proceedings under the Official Secrets Act. A system of Defense Notices issued by the government indicated sensitive areas. In the United States, the Office of Censorship set out a "Code of Wartime Practices for the American Press," which again provided a workable basis for cooperative self-censorship. From the government side, the British Ministry of Information and the American Office of War Information (OWI) both issued a vast amount of official news and propaganda. The level of war reporting and photography, in particular, reached great heights; and many journalists risked, and lost, their lives. The rationing of newsprint, which also affected the United States, though less stringently than Britain, pegged circulations and imposed difficult decisions—*e.g.*, whether to use the limited space more for advertising, and hence income, or for news coverage, and hence the loyalty of readers.

Financial trends. Since World War II the pressure of rising costs of production has accentuated the trend in most of the highly industrialized countries toward monopolistic control—there came to be fewer newspapers and these, in the main, were owned by large, monopolistic corporations. In the United States by the late 1970s almost 60 percent of the daily papers were chain owned, and eight out of 10 of all newspapers were monopolies—*i.e.*, they had no competition in the city in which they were published. New York City, which supported eight major dailies in 1960, had only four in 1980: the ***New York Post,*** an afternoon tabloid; the ***Daily News,*** a morning tabloid; the full-sized ***New York Times;*** and the full-sized ***Wall Street Journal.*** There are about 170 chains in the United States, the largest being Gannett Company, Inc. Lord Thomson, the Canadian and British proprietor, owned 67 U.S. newspapers in 1980. The effect of monopoly is heightened by the all-pervading influence of the two main news agencies, Associated Press (1848, a non-profit-making cooperative) and United Press International (1958, a commercial, profit-making agency, the outcome of a merger between United Press Association, founded by Scripps-Howard in 1907, and International News Service, formed by Hearst in 1909). It has been claimed that monopoly relieves a newspaper from the twin evils of competing for advertising and competing in sensationalism; but lack of competition can also lead to lack of entemrise in exploring subjects and airing opinions.

A serious problem, which has arisen in connection with automation, has been prolonged newspaper strikes. Coupled with the ready availability of news on radio and, especially, on television, it may have led some people to drop the newspaper habit altogether. Though total daily circulation has risen, it has not kept pace with the rise in population. In Canada, an alternative type of publishing group, Free Press Publications, was formed in the early 1960s to provide the financial benefits of a commercial alliance without the drawbacks of uniformity—each of its papers retained its identity and full editorial independence. Although Free Press Publications grew to become one of the largest chains in Canada, in 1980 the Thomson organization purchased the eight papers of the group, selling one, merging two, and terminating the publication of a fourth. Thomson and Southam then owned 48 percent of the total daily circulation, and eight smaller chains held an additional 28 percent.

The trend toward consolidation has also been visible in Britain, but the fact that the principal newspapers are national, available all over the country, has preserved a somewhat greater element of competition and choice for the reader. By 1980 almost two-thirds of Britain's daily newspapers were controlled by six chains: Associated Newspapers Group, Express Newspapers, Mirror Group Newspapers, News Group Newspapers, Thomson Regional Newspapers, and United Newspapers Publications.

Springer group

The same trend toward the large complex has taken place elsewhere in Europe. The Springer group in Germany, built up by Axel Springer after World War II, controlled almost 40 percent of the total daily circulation in West Germany and as much as 70 percent in West Berlin until 1968, when the main press commission *(Günther)* set limits on the proportions of circulation one group would be allowed to control; Axel Springer then reduced his share of the periodical market to 11 percent. The group owns two of the country's most influential national dailies, the quality ***Die Welt*** and the sensational ***Bild Zeitung,*** and two of the most influential national weeklies, ***Welt am Sonntag*** and ***Bild am Sonntag.*** Its influence, essentially conservative, is offset by the national ***Frankfurter Allgemeine*** and the local press (every large city has its daily), as well as by the news magazine ***Der Spiegel.*** In France, where the number of dailies has declined steadily since 1946, the press is controlled mainly by four groups, Hachette, Amaury, Hersant, and Expansion, the biggest being the Hachette Group, which came under the control of Matra, an electronics-defense conglomerate, in 1980. The postwar French press is less open to political pressure and is more informational than in prewar days, with a professional element that asserts itself strongly against the managerial. The editorial staffs of several papers, including ***Le Figaro*** and ***Le Monde,*** a daily of immense integrity, have won complete independence. In Italy, where only 12 percent of adults read a daily paper, the number of papers has fallen sharply, and those that remain have become financially dependent on large industrial groups, political parties, or other organizations with sufficient capital to withstand financial losses. The country's leading daily is ***Corriere della Sera;*** the voice of the Vatican is ***L'Osservatore Romano,*** a weekly; and ***L'Unita,*** the official organ of the Italian Communist Party, has been one of the country's leading newspapers since World War II.

By 1937 there were 1,200 dailies and 600 weeklies in Japan. National mobilization decrees beginning in 1938 led to totalitarian government "guidance" of the press and virtually enforced mergers that reduced the number of dailies to 53 by 1942. The postwar Allied occupation brought another form of "guidance" censorship until 1948. Top management personnel were removed, and for several years radical unions seized virtual control of such papers as the ***Yomiuri.*** After the Korean conflict erupted in 1950, the occupation authorities directed newspaper managements to purge the many Communists who had infiltrated the editorial staffs.

After sovereignty was regained in 1952, the Japanese government passed legislation intended to prevent any new attempts to employ newspapers for the purpose of subversion. By the early 1970s, however, Japanese newspapers had established a vigorous and independent stance, exposing secret government documents before their existence had been publicized in the Diet or elsewhere. Furthermore, because of the popular clamour for news, leading newspapers have increased their standard number of pages and broadened their coverage of global, national, and local affairs. Japanese newspaper circulations have remained among the largest in the world, but the trend has continued away from traditional reliance upon circulation revenues toward an increasing ratio of advertising revenue. The leading newspapers have large magazine and book divisions and are increasingly involved in affiliated television-broadcasting operations.

Though Australian newspapers were originally in the hands of such families as the Fairfax family ***(Sydney Morning Herald)*** or the Davies ***(Hobart Mercury),*** the trend in the 20th century has been toward ownership by joint stock companies, so that, by the early 1970s, no major Australian paper was privately owned. By the early 1980s three principal groups owned the bulk of the print media; the John Fairfax Group, The Herald and Weekly Times Ltd. Group, and News Ltd., and although there has been criticism of the interlocking ownership of the Australian press, no national inquiry into the extent of concentration or cross-ownership has taken place.

Of the 44 daily newspapers in New Zealand in the early 1980s, the largest were the ***New Zealand Herald*** and ***Auckland Star,*** both published in Auckland, the ***Evening Post*** and ***The Dominion*** of Wellington, and ***The Press*** and ***Christchurch Star,*** both of which appear in Christchurch. Although there has been a trend toward amalgamation, papers do prosper on their own.

Press freedom. Press freedom remains as problematic as ever, even in countries where it might be taken for granted. In France, under the Gaullist Constitution of 1959, seizures of newspapers were made as recently as 1965. A successful case against the state brought them to an end, but government sensitivity remained and press coverage of both Pres. Georges Pompidou's illness in the early 1970s and the war in Chad was minimal. In 1980 the government brought criminal charges against the editor of ***Le Monde,*** alleging that "discredit" had been done to the courts of France in articles criticizing controversial judicial decisions. In West Germany the ***Der Spiegel*** affair of 1962, in which the arrests of several members of the staff in retaliation for personal attacks on Germany's Defense Minister were upheld by the Constitutional Court, illustrated the lack of a strong tradition of press freedom in the country. A later controversy, however, surrounding published reports about the former empress of Iran, was decided in favor of the West German press when an amendment that would have extended protection for foreign heads of state against journalistic intrusion was dropped because of the opposition of German publishers and the journalists' association. Strict control of the press was imposed in Rhodesia (now Zimbabwe) in 1965 and in Greece in 1967. In South Africa, where the press is nominally free, controls and intimidation are applied under a dozen different statutes. As Dr. Hermann Knorr, editor of the ***Rhein-Neckar Zeitung,*** observed: "Press freedom guarantees are only as good as the will of the public to make them good."

Freedom of the press in Communist countries

In Communist countries, of course, press freedom is viewed in a wholly different light. The press is an arm of government, an instrument of social control, and a means to educate and mobilize the masses, with great emphasis on "self-criticism." The state censorship agency in the Soviet Union is known as Glavit, but in any case only leading party members become editors of important papers. To Communists, Western press freedom is illusory because the wealthy few control what is to be printed, whereas in Communist countries all have free access to the press. Certainly, much information is supplied by worker and village correspondents on achievements and shortcomings. One government report elicited 126,000 letters and articles, of which 102,941 were published. To this extent the newspapers are a platform for the people. In the Soviet Union, besides ***Pravda,*** the All-Union daily with a circulation of 11,000,000, and ***Zzvestiya*** (8,600,000), the official voice of the Soviets, both fine productions if perhaps somewhat staid to Western eyes, there are hundreds-of publications representing different interests and many wall newspapers in factories, collectives. and institutions. The press of all Communist countries' is modelled on that of the Soviet Union. Even in Yugoslavia, which is independent of the Soviet bloc and where the 1974 constitution provides certain freedoms, the same strict controls are applied in practice. In China, where the country's many papers take their lead from ***Jen-min Jih-pao,*** the ***People's Daily*** of Peking, the wall newspaper, often mimeographed or hand written, is a strong feature, especially in remote communities. It is maintained by a "keeper," who reads it aloud to illiterates.

Underground presses

Wherever there are strong currents of dissent from prevailing notions, an underground press is likely to develop; clandestine printing is almost as old as the art itself. In Nazi-occupied Europe, there were many such publications, which maintained morale by spreading news of the Resistance and of Allied successes. Since the war the term underground press has been applied to ***samizdat*** ("self-publishing") material in the Soviet Union, where some free literary life continues in spite of official repression. News of this, together with reports of such matters as trials and persecutions for ***samizdat*** activity, religious belief, and demonstrations, is circulated secretly in a publication called "A Chronicle of Current Events," copies of which have reached the West. Nor does the West lack its own vigorous "underground," which is also known more appropriately, since publication does not carry the same risks, as "the alternative press." It consists of a host of small, evanescent productions, chiefly in the United States and Britain but also in western Europe, published by young people as an expression of social dissent. At its most evangelical, it puts forward a radically different, alternative set of values; at the level of entertainment, it leans heavily on pornography and youthful trends. In some cases—*e.g.*, ***The Village Voice*** of New York City and ***Time Out*** of London—the underground becomes virtually overground, with a high, stable circulation.

The press in developing countries. A flourishing press presupposes some degree of literacy, yet approximately three out of every 10 adults in the world are illiterate. During a general conference in 1976, the United Nations Educational, Scientific and Cultural Organization (UNESCO) reported that only about 12 percent of the newsprint consumed worldwide went to the developing countries of South America, Africa, and Asia, representing an average of about 2 pounds (one kilogram) of newsprint per head of population, compared with an average of almost 40 pounds per head of population in the industrially developed countries of Europe and North America. This is one more measure of the disparity between advanced countries and those who are making what has been called the "terrible ascent to modernity." The correlation between economic growth and growth in communications is well established; newspapers are as essential to developing countries as they are elsewhere. Apart from providing a flow of information and acting as a two-way link between leaders and people far more effectively than radio or television can do, newspapers are themselves a vital step in the acquisition of literacy, with all of its benefits. But the problems are immense. Besides the obvious economic ones of equipment and trained staff and the fact that there have to be people who can afford to buy a paper and businesses that can afford to advertise in it, there is the multiplicity of languages in Asia and Africa and the problem of typesetting in local scripts; in some cases a paper is written by calligraphers, reproduced photographically, and printed by lithography. A newspaper in a developing country is also finely balanced in its policy: though it may aspire to be a public watchdog, it is often compelled to mute its bark. In Ghana and Nigeria, for instance, which have a press tradition of more than a century—rooted in colonialism, mission activity, and the national awakening of the 1930s—the press is less free than it was before the countries became independent. There are, of course, fine newspapers in all of the large urban centres in South America, Asia, and Africa; but the rural masses are less easy to reach. For this journalistic techniques have to be adapted to local needs. Basic English is sometimes used, and one newspaper is published in pidgin English, the Nu Gini Toktok (*i.e.*, "New Guinea Talk Talk"). Since 1950 great progress has been made, partly through UNESCO's program of fundamental education. Throughout the underdeveloped world the number of newspapers has risen sharply, and many more national news agencies have been founded, despite a hampering lack of telecommunication facilities.

CONTEMPORARY AND FUTURE TRENDS

New printing techniques

During the past 100 years, technical developments have transformed newspaper production, and they are continuing to do so. A recent innovation has been the application of colour printing to the daily paper. While this may be done partly for its own sake, as a natural evolution, the chief hope is that it will enable the newspaper to compete for advertising on equal terms with colour television and the glossy magazines. More far-reaching is the development of photocomposition and web-offset printing, which eliminate hot-metal typesetting and reduce the amount of skilled labour required. Though capital costs are still high, such new techniques are particularly suited to small-circulation journals and may provide the long-term answer to the needs of emerging nations. Computer typesetting and electronic control systems, despite strong resistance from unions, clearly point the way to future development. So, too, does the increasing use of telecommunication facilities. In 1959 the highly advanced Japanese press led the world in facsimile transmission and teletypesetting by publishing simultaneous editions of newspapers in cities hundreds of miles apart. Such transmission processes, via land line, cable, radio, and space satellite, are bridging ever greater distances at ever greater speeds of operation. The trend points toward multicentre printing of identical material and, hence, stronger national and even supernational newspapers, though regional loyalties—*e.g.*, in the United States—may prove resistant. In 1962 the world news agencies also entered the space age by using the first communications satellite to relay dispatches by telephone and teleprinter.

The trend toward monopoly in advanced countries, owing to the dependence of the modem newspaper on advertising and the high cost of keeping pace with technical development, seems irreversible, for only the big and strong can survive. But although it must be deplored, there are limits to its dangers. Whatever the inhibiting influence of advertisers or the caution of large corporations, the public-service function of a newspaper cannot be denied beyond a certain point without damage to its nature and so, ultimately, to its commercial existence. In addition, there is much evidence of the worldwide strengthening of the professional element in journalism, promoted by the great amount of study it is receiving. Where managements, however monopolistic, uphold the integrity and independence of their editors, the public interest can still be served. Finally, there is always the leaven of the small publication. Though the cost of starting a daily paper has become prohibitive, it has never been easier to issue an alternative paper in some form or other. So easy is it to produce such a newspaper that in Czechoslovakia, where the government has been unable to suppress the strong underground press, anyone who buys a typewriter has to leave his name and address.

Government security

Far more serious for a free press is the trend toward greater government secrecy. In the name of internal stability or external threat or even the projection of a good national image, criticism is all too readily labelled "unpatriotic." As government grows more complex and bureaucratic, important decisions are taken by faceless officials, who are then protected by their ministers. To avoid unwanted disclosures, the press handout has come increasingly to take the place of journalistic investigation. In the United States in 1955, the Moss Sub-committee of Congress reported the development of "An attitude novel to democratic government—an attitude which says that we, the officials, not you, the people, will determine how much you are to be told about your Government." Since then this attitude has become all too widespread and not only in the United States, of course, which at least preserves an unrivalled capacity for self-criticism. There have been many instances of courageous individuals who have set what they judged to be the public interest above their official loyalties and passed on information to newspapers, which, in turn, have risked publication—and both parties have been subsequently justified. This involves the difficult question of obligation to reveal sources; Scandinavian countries are among the few that fully recognize the right of a news source to anonymity. In the last analysis, the answer lies with the public: the health of a free press depends on the expectations and demands of its readers.

As to the future of the newspaper itself, this seems assured; there is nothing on the horizon likely to replace it, though it will doubtless undergo further changes of form. Radio and for similar reasons television, vivid and stimulating though these media are, have proved to be no substitute for newspapers. It is hard to see how the electronic media can ever match the newspaper in its variety of material handled in convenient form for assimilation and reference at will. The most serious effect of television, in countries with a commercial system, has been to draw off advertising revenue. This has added considerably to the financial problems of the press and may eventually lead to higher prices for newspapers. But it seems likely that these will always be paid—for a good product. Even in developing countries, where radio may become the primary means of mass communication, newspapers follow hard on its heels.

III. Magazine publishing

BEGINNINGS IN THE 17TH CENTURY

Though there may have been approaches to a magazine in antiquity, especially perhaps in China, the magazine as it is now known began only after the invention of printing in the West. It had its taproot in the spate of pamphlets, broadsides, ballads, chapbooks, and almanacs that printing made possible. Much of the energy that went into these gradually became channelled into publications that appeared regularly and collected a variety of material designed to appeal to particular groups of interests. The magazine thus came to occupy the large middle ground, incapable of sharp definition, between the book and the newspaper.

The earliest magazine or periodical appears to have been the German Erbauliche Monaths-Unterredungen (1663–68; "Edifying Monthly Discussions"), started by Johann Rist, a theologian and poet of Hamburg. Soon after, there appeared a group of learned periodicals; the Journal des *Sçavans* (later Journal des Savants; 1665), started in France by the author Denis de Sallo; the Philosophical Transactions (1665), of the Royal Society in England; and the Giornale de' *letterati* (1668) published in Italy, issued by the scholar and ecclesiastic Francesco Nazzari. A similar journal was started in Germany a little later, the *Acta eruditorum Lipsiensium* (Leipzig; 1682); and mention may also be made of the exile-French *Nouvelles de la République des Lettres* (1684), published by the philosopher Pierre Bayle mainly in Holland, to escape censorship. All these sprang from the revival of learning, from the need to review its fruits, and the wish to diffuse its spirit as widely as possible. It is notable that England's Royal Society, for instance, embraced the whole field of knowledge and had as one of its aims the improvement of English prose.

While the learned journals might summarize important new books, there were as yet no literary reviews. Book advertisements, by about 1650 a regular feature of the news sheets, sometimes had brief comments added, and regular catalogues began to appear, such as the English quarterly *Mercurius librarius, or A Catalogue of Books* (1668–70). But in the 17th century the only periodicals devoted to books were very short-lived: the *Weekly Memorials for the Ingenious* (1682–83), which offered some critical notes on books, and the *Universal Historical Bibliothèque* (January–March 1686). The latter invited scholarly contributions and could thus be regarded as the true forerunner of the literary review.

The lighter type of magazine, or "periodical of amusement," may be dated from 1672, which saw the first appearance of the *Mercure Galant* (renamed *Mercure de France* in 1714). It was founded by the writer Jean Donneau de Visé and contained court news, anecdotes, and short pieces of verse—a recipe that was to prove endlessly popular and become widely imitated. This was followed in 1688 by a German periodical with an unwieldy title but one that well expressed the intention behind many a subsequent magazine: "Entertaining and Serious, Rational and Unsophisticated Ideas on All Kinds of Agreeable and Useful Books and Subjects." It was issued in Leipzig by the jurist and publicist Christian Thomasius, who made a point of encouraging women readers. England was next in the field, with a penny weekly, the *Athenian Gazette* (better known later as the *Athenian Mercury;* 1690–97), run by a London publisher, John Dunton, to resolve "all the most Nice and Curious Questions." Soon after came the *Gentleman's Journal* (1692–94), started by the French-born Peter Anthony Motteux, with a monthly blend of news, prose, and poetry; and in 1693, after devoting some experimental numbers of the *Athenian Mercury* to "the Fair Sex," Dunton brought out the first magazine specifically for women, the *Ladies' Mercury,* which offered forthright advice on some perennial problems. Finally, another note, taken up time and again later, was struck by *The London Spy* (1698–1700), issued by a tavern keeper, Ned Ward, and containing a sort of running narrative of the sights and sounds of London.

DEVELOPMENTS IN THE 18TH CENTURY

With the gathering momentum of trends at the turn of the century—increasing literacy, especially among women, and a quickening interest in new ideas—the magazine filled out and became better established. In England, three early "essay periodicals" had enormous influence: Daniel Defoe's *Review* (1704–13; thrice weekly), Sir Richard Steele's *Tatler* (1709–11; thrice weekly), to which Joseph Addison soon contributed, and Addison and Steele's *Spectator* (1711–12, briefly revived in 1714; daily). Though they resembled newspapers in the frequency of their appearance, they were more like magazines in content. *The Review* introduced the opinion-forming political article on domestic and foreign affairs; while the cultivated essays of the *Tatler* and *Spectator,* designed "to enliven morality with wit, and to temper wit with morality," did much to shape the manners and taste of an age. The latter had countless imitators, not only in England, where there were in addition a *Female Tatler* (1709–10) and a *Female Spectator* (1744–46), but also on the Continent and later in America. The Stamp Tax of 1712 had a damping effect, as intended (see above), but the magazine proved endlessly resilient, easy to start and easy to fail, then as now.

Addison and Steele's *Spectator*

So far various themes had been tried out; they were first brought together convincingly by the English printer Edward Cave, who began to publish *The Gentleman's Magazine* in 1731, with the motto "*e pluribus unum*." It was originally a monthly collection of essays and articles culled from elsewhere, hence the term magazine—the first use of the word for a periodical. Cave was joined in 1738 by Doctor Johnson, who was later to publish his own *Rambler* (1750–52); thereafter *The Gentleman's Magazine* contained mostly original matter, including parliamentary reports. Rivals and imitators quickly followed, notably the *London Magazine* (1732–85) and the *Scots Magazine* (1739–1817; to 1826 published as the *Edinburgh Magazine);* and, among the increasing number of women's periodicals, there were a *Ladies' Magazine* (1749–53) and a *Lady's Magazine* (1770–1832). Their progenitor, however, outlived them all and perished only in 1907.

The literary and political rivalries of the day produced numerous short-lived periodicals, from which the critical review emerged as an established form. Robert Dodsley, a London publisher, started the *Museum* (1746–47), devoted mainly to books, and a Nonconformist bookseller named Ralph Griffiths founded *The Monthly Review* (1749–1845), which had Oliver Goldsmith as a contributor. To oppose the latter on behalf of the Tories and the Church of England, *The Critical Review* (1756–1817) was started by an Edinburgh printer, Archibald Hamilton, with Tobias Smollett as its first editor. Book reviews tended to be long and fulsome, with copious quotations; a more astringent note came in only with the founding of the *Edinburgh Review* in 1802 (see below).

Europe. In Europe, development was similar but was hampered by censorship. French periodicals containing new ideas had to appear in exile, such as Bayle's (see above); some 30 were published in Holland up to the time of the Revolution. Within France, there were the short-lived *Spectateur français* (1722–23) and *Spectateur suisse* (1723); and *Le Pour et le Contre* (1733–40; "For and Against"), issued by Abbé Prévost. Of more literary interest were the *Gazette littéraire de l'Europe* (1764–84) and *La Décade philosophique, littéraire et politique* (1794–1804). In Germany, the poet and philosopher J.C. Gottsched issued the country's first women's periodical, *Die vernünftigen Tadlerinnen* (1725–26; "The Rational Woman-critics"), and the first literary review, *Beiträge zur kritischen Historie der deutschen Sprache* (1732–44; "Contributions to the History of the German Language"), published in Leipzig. Literary movements were connected more profoundly than in England with the production of new periodicals, the influence of which was often greater than their duration, among them, Schiller's *Horen* (1795–97) and Goethe's *Propylaen* (1798–1800). Of more general and lasting influence was the *Allgemeine Literatur-zeitung* (1785–1849), founded by F.J. Bertuch, "the father of the German periodical."

The first Russian periodical, published by the Academy of Sciences, was a learned journal called "Monthly Works" (1755–64). The first privately published Russian magazine, a critical periodical with essays and translations from England's *Spectator,* was called "Industrious Bee" and began in 1759. Catherine II used her *Olla Podrida* (1769–70), also modelled on the *Spectator,* to attack opponents, among them Nikolay Novikov, whose "Drone" (1769–70) and "Windbag" (1770) were suspended and whose "Painter" (1770–72) escaped only by being dedicated to the Empress.

The United States. In America, magazines got off to a tentative start in 1741. The first was Andrew Bradford's *American Magazine,* beating by a mere three days Benjamin Franklin's *General Magazine;* both appeared in Philadelphia. Neither lasted long, however; Bradford's survived only three months and Franklin's six. Franklin was more widely known for another of his publications, *Poor Richard's Almanack* (1733–58), containing maxims and proverbs. Before the end of the century, some 100 magazines appeared, offering miscellaneous entertainment, uplift, or information, mostly on a very shaky, local, and brief basis. Among the more important were, in Philadelphia, the *Perznsylvania Magazine* (1775–76), edited by Thomas Paine, and the *American Museum* (1787–92) of the bookseller Mathew Carey; the *Massachusetts Magazine* (1789–96); and the *New-York Magazine* (1790–97).

THE 19TH CENTURY AND THE START OF MASS CIRCULATION

General periodicals. Most of the early periodicals, designed for the few who could afford them, reflected the tastes of the leisured classes; they were largely "quality" magazines. In the 1830s, however, cheap magazines began to appear amed at a wider public. At first they were strong on improvement, enlightenment, and family entertainment, but, toward the end of the century, they evolved into a popular version of the old "periodical of amusement."

The pioneers in England were Charles Knight, publisher for the Society for the Diffusion of Useful Knowledge, with his weekly *Penny Magazine* (1832–46) and *Penny Cyclopaedia* (1833); the Chambers brothers, William and Robert, with *Chambers's (Edinburgh) Journal* (1832–1956), which reached a circulation of 90,000 in 1845; and teetotaler John Cassell, with his *Working Man's Friend and Family Instructor* (1850) and the *Quiver* (1861). Besides popular magazines, many standard works appeared in parts, often with illustrations. Typical of the family type of entertainment were Charles Dickens' *Household Words* (1850), followed in 1859 by *All the Year Round;* several similar periodicals such as *Good Words* (1860); and, for young people, the *Boy's Own Paper* (1879) and the *Girl's Own Paper* (1880). Germany, too, had its *Pfennigmagazin* (1833), edited by J.J. Weber, and a *Familienblatt* ("Family Magazine") modelled on that of Dickens, such as the *Gartenlaube* (1853–1937; "Arbour"), which enjoyed great popular influence, with a circulation of 400,000, in the 1870s. The United States had no national magazines before about 1850, but two of its most well-known early periodicals were the *Saturday Evening Post* (1821–1969) and *Youth's Companion* (1827–1929). The latter, published in Boston, was typically wholesome in content, intended to "warn against the ways of transgression" and to encourage "virtue and piety."

German popular magazines

By the last quarter of the century, largely as a result of compulsory education, the potential public for magazines had greatly increased and was avid for miscellaneous information and light entertainment. The first man to discover this in Britain was George Newnes, who liked snipping out any paragraph that appealed to him. In 1881, he turned his hobby to advantage by publishing a penny magazine, *Tit-Bits from all the Most Interesting Books, Periodicals and Contributors in the World,* soon shortened to *Tit-Bits* (still in publication). It was a great success and the beginning of an empire that was to include *Country Life* (1897–), *Wide World Magazine* (1898), and, above all, the *Strand* (1891–1950), one of the first monthly magazines of light literature with plenty of illustration; it became enormously popular for its Sherlock Holmes stories by Arthur Conan Doyle. Among the early contributors to *Tit-Bits* was Alfred Harmsworth (later Lord Northcliffe), who had a similar appetite for odd bits of information. In 1888, after editing *Youtlz* and *Bicycling News,* he launched a rival to *Tit-Bits* called *Answers to Correspondents,* or *Answers,* which he successfully promoted by contests. Within five years he had a string of cheap magazines for the same popular market, including *Comic Cuts* and *Home Chat.* A similar empire was built up by Arthur Pearson, another former *Tit-Bits* man, with *Pearson's Weekly* and *Home Notes,* among others.

Early popular monthly magazines in the U.S.

In the United States, magazine publishing boomed after the Civil War, as part of the general expansion, and it was also helped by favourable postal rates for periodicals (1879). But a gulf remained between expensive magazines geared to the genteel, such as *Harper's* and *Scribner's* (see below), and the cheap weeklies and miscellanies. The first man to aim at a popular monthly in between and thus spark off a revolution in the industry was S.S. McClure, who began publishing *McClure's Magazine* in 1893, at 15 cents instead of the usual 25 or 35 cents. John Brisben Walker, who was building up *Cosmopolitan* (1886–) after acquiring it in 1889, cut his price to 12% cents, and, in October 1893, Frank A. Munsey reduced the price of *Munsey's Magazine* (1889–1929) to 10 cents. All three saw that, by keeping down the price and gearing contents to the interests and problems of the average reader, high circulations would become possible. Frank Munsey estimated that, between 1893 and 1899, "the ten-cent magazine increased the magazine-buying public from 250,000 to 750,000 persons." This in turn led to high advertising revenue, making it possible to sell a magazine, like a newspaper, for less than its cost of production, a trend that was to become increasingly pronounced in the next century. The role of technical development was also important; mass-production methods and the use of photoengraving processes for illustration enabled attractive magazines to be produced at ever lower unit costs.

The first periodical published in Australia was the *Austıalian Magazine,* which began in 1821 and lasted for 13 monthly issues. The *South Asian Register* began quarterly in 1827 but published only four issues. The *Hobari Town Magazine* (1833–34) survived longer and contained stories, poems, and essays by Australian writers; the *Sydney Literary News* (1837) was first to contain serial fiction and advertisements. Illustrations were introduced in the 1840s; the *Australian Gold Digger's Monthly Magazine and Colonial Family Visitor* (1852–53) was followed by the *Melbourne Punch* (1855–1925; incorporated in *Table Talk,* 1885–1937). In India, the first published magazine was the *Oriental Magazine; or, Calcutta Amusement* (1785–86); it was followed by a number of short-lived missionary publications. The first periodical founded and edited by an Indian was the *Hindustan Review,* which commenced in 1900.

Magazines in China

Missionaries founded the first periodical in China; printed in Malacca, the *Chinese Monthly Magazine* lasted from 1815 to 1822. It was followed by the *East-West Monthly Magazine,* printed in Canton from 1833 to 1837 and in Singapore from 1837 until its end in 1847. Although learned journals were founded in the 19th century, only in the 1900s, with the literary revolution, did they circulate widely.

Illustrated magazines. Woodcuts had been used to illustrate newspapers from the beginning, but the early ones were inclined to be decorative and imaginative rather than accurate and topical. The first reliable pictures to portray events shortly after their occurrence date from 1806, when *The Times* (London) published a picture of Nelson's funeral car. Thereafter, news illustration became more common, especially in the case of great public occasions such as coronations. The first man to notice their effect on sales and grasp their possibilities was a British newsagent in Nottingham, Herbert Ingram, who moved to London in 1842 and began publishing the *Illustrated London News,* a weekly consisting of 16 pages of letterpress and 32 woodcuts. It was successful from the start, winning approval from the Archbishop of Canterbury and hence the support of the "rectory" public. Though it suffered at first from the old defect—its pictures were by well-known artists but were not taken from life—it later sent artists all over the world. Drawings made on the spot during the South African War, sometimes at considerable risk, were a great popular feature. Among its competitors was the monthly *English Illustrated Magazine* (1883–1913).

News in pictures

The idea of presenting the news largely in pictures was quickly taken up on the Continent, in France by *L'lllustration* (1843–1944) and in Germany by the *Leipziger illustrierte Zeitung* (1843) and *Die Woche* (1899–1940).

In the United States, the main early forerunners of the genre were *Leslie's Weekly* (1855–1922) and *Harper's Weekly* (1857). Soon after its founding, *Leslie's* had a circulation of 100,000, which doubled or trebled whenever there was something sensational to portray. During the Civil War, of which it gave a good pictorial record, it had as many as 12 correspondents at the front.

The invention of photography and the development of the halftone block began to transform this type of magazine in the last decade of the century; from then on the artist was gradually replaced by the camera.

Women's magazines. Women's magazines are barometers of fashion in more ways than one; they reflect the changing view of women's role in society. In the 18th century, when women were expected to participate fully in social and political life, they were robust and stimulating in content; in the 19th, when domesticity became the ideal, they were inclined to be insipid and humourless. After about 1880, they began to widen their horizons again.

Typical of the Georgian and Regency magazines in Britain were The Lady's Magazine (1770), a sixpenny monthly that, along with its literary contributions and fashion notes, gave away embroidery patterns and sheet music; The Lady's Monthly Museum (1798), which had a half-yearly "Cabinet of Fashion" illustrated by coloured engravings, the first to appear in a women's periodical; and La Belle *Assemblée* (1806), which encouraged its readers to unburden themselves in its correspondence columns. These three merged in 1832, the first instance of what was to become a common occurrence, but ceased publication in 1847. The Victorian type of magazine included The Ladies' Pocket Magazine (1824–40), The Ladies' Cabinet (1832–52), The New Monthly Belle *Assemblée* (1847–70), and The Ladies' Treasury (1857–95). All contained verse, fiction, and articles of high moral tone but low intellectual content. There were attempts to swim against the tide, such as The *Female's* Friend (1846), which was one of the first periodicals to espouse women's rights, but they seldom lasted long.

In 1852, a wider market began to be tapped by The Englishwoman's Domestic Magazine, a monthly issued by Samuel Beeton at twopence instead of the usual one shilling; it was also the first women's periodical to concentrate on home management, to offer service to women rather than upper class entertainment. Beeton's wife (author of the classic Book of Household *Management*, 1861) visited Paris regularly and acquired the use of the fashion plates from Adolphe Goubaud's Moniteur de la Mode. A feature of the magazine was the "Practical Dress Instructor," a forerunner of the paper dressmaking pattern. In 1861, Samuel Beeton followed up his success with The Queen, a weekly newspaper of more topical character.

The great expansion of women's magazines into a major industry may be dated in Britain from Myra's *Journal* of Dress and Fashion (1875–1912) and *Weldon's* Ladies' Journal (1875–1954), both of which supplied dressmaking patterns and met the needs of a mass public. There were several new quality magazines, such as *The* Lady (1885–) and The Gentlewoman (1890–1926), one of the first to acknowledge the financial necessity of advertisements, but many more cheap weeklies, such as Home Notes (1894–1957), Home Chat (1895–1958), and Home Companion (1897–1956), which were of great help to the average woman in meeting new standards of hygiene, nutrition, and child care.

Among the earliest women's magazines in the United States was a monthly published in Philadelphia called Godey's Lady's Book (1830–98), which employed up to 150 women to hand-tint its famous fashion plates. Of the early national magazines, one of the best and hardiest was Harper's Bazar (1867; Harper's Bazaar after 1929), modelled on the Berlin women's periodical Der Bazar, Srom which it obtained its fashion material. The practical trend was begun in 1863 by Ebenezer Butterick, who devised the tissue-paper clothing pattern and, to popularize it, brought out the Ladies' Quarterly Review of Broadway Fashions and, later, Metropolitan. These merged in 1873 into the Delineator, which had a highly successful career down to 1937. The field of women's magazines was finally transformed, however, by Cyrus Curtis with his Ladies' Home Journal (1883–), edited by his wife. This soon reached a circulation of 400,000, and, under the editorship of Edward W. Bok, after 1889, it broke with sentimentality and piety and became a stimulating journal of real service to women. Other popular magazines were Ladies' Home Companion (1886; called Woman's Home Companion 1897–1957), *McCall's* Magazine (1897–), and Pictorial Review (1899–1939). Two requiring special mention were Good Housekeeping (1885–), which established a testing station for consumer goods early in the 1900s, and *Vogue* (1892–), a fashion weekly (originally) dedicated to "the ceremonial side of life," which was designed for the elite of New York and had Cornelius Vanderbilt among its backers.

Literary and scientific magazines. The critical review, often as an adjunct to a book-publishing business, developed strongly in the 19th century. It became a forum for the questions of the day, political, literary, and artistic, to which all the great figures contributed. There were also many magazines with a literary flavour, which serialized some of the best fiction of the period, and a few that marked the beginning of specialization—*e.g.*, in science.

The Edinburgh Review

Britain was particularly rich in reviews, beginning with the Edinburgh Review (1802–1929), founded by a trio of brilliant young men, none over 30, but all gifted critics, Francis Jeffrey, Henry Brougham, and Sydney Smith. The higher and more independent tone they adopted was said by Coleridge to mark an "epoch in periodical criticism." Though Tories, including at first Sir Walter Scott, wrote for it, it gradually became Whig in attitude. When it did, Scott transferred his allegiance to the Quarterly Review (1809–), its Tory rival, founded by the London publisher John Murray and first edited by William Gifford, who had previously edited Canning's review The *Anti-Jacobin* (1797–98). In opposition to these and more political than either of them was the *Westminster* Review (1824–1914), started by Jeremy Bentham and James Mill as an organ of the philosophical radicals. Two other early reviews were the *Athenaeum* (1828–1921), an independent literary weekly, and the Spectator (1828–), a nonpolitical weekly that nevertheless supported Reform and the cause of the North in the American Civil War. Later reviews included the Saturday Review (1855–1938), which had George Bernard Shaw and Max Beerbohm as drama critics (1895–1910); the Fortnightly Review (1865–1954), which had John Morley as editor (1867–83); the *Contemporary* Review (1866–); the Nineteenth Century (1877; now quarterly as Twentieth Century); and W.T. Stead's Review of Reviews (1890–1936), a limited kind of *Reader's* Digest.

Of the closely related literary magazines, one of the earliest and best was Blackwood's Edinburgh Magazine (1817–), founded by a book publisher, William Blackwood, as a rival to the Edinburgh Review, but a less ponderous one than the Quarterly. It provoked in turn the founding of the London Magazine (1820–29), in which Lamb's Essays first appeared. The rivalry between these two led to a duel in which John Scott, the first editor of the London Magazine, was mortally wounded. Other literary periodicals included the Examiner (1808–80), edited by Leigh Hunt, who introduced Shelley and Keats through its columns; the New Monthly Magazine (1814–84); Bentley's Miscellany (1837), which had Dickens as its first editor and "Oliver Twist" as one of its serials; and the *Cornhill* (1860–), first edited by William Thackeray and the first magazine to reach a circulation of 100,000. Finally, two rather different periodicals must be mentioned: Nature (1869–), which began to make scientific ideas more widely known and to which Darwin and Huxley contributed; and *Punch* (1841–), which provided a weekly humorous comment on British life illustrated by many distinguished draftsmen.

European reviews tended to be more literary than political, perhaps because of the persistence of censorship. The most notable in France were the Revue des Deux Mondes (1829–), with such contributors as Sainte-Beuve and Victor Hugo, and its rival the (Nouvelle) Revue de Paris (1829–), which published authors disapproved by the other, notably Flaubert. In Germany, F.A. Brockhaus, the book publisher, tried to emulate the Edinburgh Review with Hermes (1819–31) but had more success with Literarisches Wochenblatt (1820–98). Later reviews were the conservative *Deutsche Rundschau* (1874–) and the liberal Freie *Bühne* (1890). Two influential Italian reviews were the ***Nuova Antologia*** (1866–) and ***La Cultura*** (1881–1935).

Early U.S. literary magazines

The early literary magazines in the United States included, among many others often of more local interest, the ***Philadelphia Literary Magazine*** (1803–08); the ***Monthly Anthology*** (Boston, 1803–11), which became the quarterly ***North American Review*** (1815–1940), with a host of famous contributors; the ***New York Monthly Magazine*** (1824); ***Dial*** (1840–44), the organ of Emerson's Transcendental Club (there was a second, literary ***Dial,*** 1880–1929); and ***De Bow's Review*** (New Orleans, 1846–80). The cultured weekly ***Home Journal*** (1846–1901; then as ***Town and Country)*** introduced Swinburne and Balzac to Americans, while ***Harper's New Monthly Magazine*** (New York, 1850–) founded by the book-publishing Harper brothers, serialized many of the great English novels and became one of America's finest quality magazines. It was rivalled only by the ***Atlantic*** (Boston, 1857–), which had a long line of distinguished editors, beginning with James Russell Lowell, and published most of the great American writers, from Emerson, Longfellow, and Holmes onward: it seemed to enjoy “a perpetual state of literary grace. Not far below in excellence was ***Scribner's Monthly*** (1870), which became the ***Century*** (1881–1930) but was restarted as ***Scribner's Magazine*** (1887–1939); and a fine magazine in the Far West was ***Overland Monthly*** (San Francisco, 1868–1935), first edited by Bret Harte. Of more specialized magazines, three require mention: ***Scientific American*** (1845–), founded by Alfred Ely Beach, a talented inventor whose magazine encouraged inventors; ***Popular Science Monthly*** (1872–), for the spreading of scientific knowledge, which had William James and John Dewey among its contributors; and the ever-popular ***National Geographic Magazine*** (1888–), published by the National Geographic Society, which used some of the proceeds to sponsor scientific expenditions.

Scholarly journals. The publishing of scholarly journals, begun in the 17th century, expanded greatly in the 19th as fresh fields of inquiry opened up or old ones narrowed into specialties. Numerous learned societies were formed, dealing with such fields as classical studies, biblical studies, archaeology, philology, Egyptology, the Orient, and all the branches into which science was dividing, each with its regular bulletin, proceedings, or "transactions," which enabled scholars to keep in touch with what others were doing. In the sober pages of these journals, seldom read by the general public, some of the most far-reaching discoveries were first made known and then filtered through into the world at large, to change man's view of himself and the universe. Notable at random were ***Annali del Institutio di Corrispondenza Archaeologica*** (1829–), the ***Revue Archéologique*** (1844–), ***Philologus*** (1846–), ***Mind*** (1876–), the ***Journal of Hellenic Studies*** (1880–), the ***American Journal of Philology*** (1880–), the ***Asiatic Quarterly*** (1886–), the ***Geographical Journal*** (1893–), or an interesting informal aid to scholars, ***Notes and Queries*** (1849), with the motto: "When found, make a note of." In every advanced country the professions too began to have their journals, such as medicine's ***Lancet*** (1823–), in Britain, originally started to attack abuses in hospital administration, the ***Mining Journal*** (1835–), the ***British Medical Journal*** (1840–), the ***Engineer*** (1856–), and the ***Solicitors' Journal*** (1857–), to cite only a few examples. In the course of time, these developed endless technical ramifications. The economics of all such journals are based on necessity. Though their circulation is small, anyone working in a particular field must subscribe to them or at least have access to them in appropriate libraries. They can be described as reference books in installments; for further description see INFORMATION PROCESSING.

THE 20TH CENTURY

The advertising revolution. There is a tradition of resistance to advertising in magazines, in keeping with their literary affinities. When the advertisement tax in Britain was repealed in 1853 and more advertising began to appear, the ***Athenaeum*** thought fit to say: "It is the duty of an independent journal to protect as far as possible the credulous, confiding and unwary from the wily arts of the insidious advertiser." In the United States many magazines, such as ***Harper's,*** took a high line with would-be advertisers until the 1880s; and ***Reader's Digest,*** with its mammoth circulation, only admitted advertisements to its American edition in 1955. Yet today some sectors of the magazine industry are dominated by advertising, and few are wholly free from its influence.

This revolution grew out of the social and economic conditions of the late 1880s, when industrialization was producing a swelling flood of goods to advertise; when education, income, and leisure were rising among an increasing number of potential consumers; and when technical innovations, such as cheap paper from wood pulp, the rotary press, and the halftone block, were opening up fresh possibilities in magazine production. Its pioneers, as has been noted, were McClure, Walker, and Munsey in the United States and Harmsworth in Britain, who discovered the possibility of mass circulation by reducing prices below costs of production and recouping themselves through advertising revenue. This technique was developed with growing vigour during the next decades.

Magazine advertising economics. Among its most vivid exponents in the United States was Cyrus Curtis with the ***Saturday Evening Post.*** He bought the magazine for $1,000 in 1897, when it was on its last legs, and sank into it $1,250,000 of his profits from the ***Ladies' Home Journal*** before it finally caught on. But when it did, through an appeal based on well-founded stories and articles about the business world, a prime interest at the time, its success was enormous; by 1922 it had a circulation of more than 2,000,000 and an advertising revenue in excess of $28,000,000. It was a classic demonstration of modern magazine economics: as circulation rose in the initial phase of low advertising rates, money had to be poured in to meet the cost of producing more copies; but, as soon as high advertising rates could be justified by a high circulation, profitability was assured. Conversely, when high rates are maintained on a falling circulation, it is the advertisers who lose, until they withdraw their support.

Verifying circulation

Once circulation figures became all-important, advertisers naturally asserted their right to verify them. The first attempt, made in 1899 by the Association of American Advertisers, only lasted until 1913, but fresh initiatives in 1914 created the Audit Bureau of Circulation. Though resented at first by publishers, it was eventually seen as an asset to them, as a guarantee of their claims. From the publishers' side, interest in circulation led them into market research. The first organization for this purpose was set up by the Curtis Publishing Company in 1911; but such research did not become general until the 1930s. When it was purely promotional—*i.e.*, designed to convince advertisers of the value of a particular medium—it came to be viewed with a certain skepticism. Reader research, to ascertain what readers wanted, was also developed in the 1930s and proved to be a useful tool, though no substitute for editorial flair. As was once observed by the features editor of ***Vogue:*** "If we find out what people want, it's already too late."

Thus the popular magazine in the United States, expanding with the economy, became a part of the marketing system. By 1900 advertisements might form up to 50 percent of its contents; by 1947, the figure was more often 65 percent. A proprietor was no longer just selling a bundle of attractive editorial matter to a segment of the public; he was also selling a well-charted segment of the public to the advertiser. Though the process was most pronounced in the United States, a vast country where, in the absence of national newspapers, national magazines had a special function, the same principles came to apply, in varying degrees, in Britain and the rest of Europe.

The effects of advertising on the appearance of the magazine have been, on the whole, stimulating. At the turn of the century, advertisements began to move forward from the back pages into greater prominence among the editorial matter, and this was often regretted by readers. At the same time, advertising agencies were developing from mere space sellers into copywriters and designers; their efforts to produce work of high visual appeal forced editors to make their own editorial typography and layout more attractive. The use of colour, in particular, was greatly fostered by advertisers, as they discovered its effectiveness. In the 1880s, it was rare, but, after the development of the multicolour rotary press in the 1890s, it steadily became more common; by 1948, nearly half the advertising pages of the leading American magazines were in two or more colours.

Advertising and editorial content

The effect of advertising on editorial content is more open to question. There is no doubt that advertisers have not been slow to exercise their financial pressure and have often succeeded in suppressing material or modifying policy. In 1940, for instance, Esquire lost its piano advertisements after publishing an article recommending the guitar for musical accompaniment; six months later it tried to win them back with a rueful editorial apology. Yet many magazines, notably the Saturday Evening Post, Time, and the New Yorker, insisted on editorial independence. Something like a balance of power has come into being, which can tip either way. What can safely be said is that advertising pressure as a whole has been a socially conservative force, playing on conformity, inclining magazines to work on the principle of "minimum offense," and holding them back from radical editorial departures until they are clearly indicated by changes in public taste. This has tended to make the large-circulation magazine an exploiter rather than a discoverer of fresh talent or new ideas. Yet in the last analysis, advertisers have been forced to recognize that magazines, like newspapers, cannot forgo too much of their independence without forfeiting the loyalty of their readers and hence their value as an advertising medium.

Women's magazines. The bond with advertising is probably most evident in magazines for women, since they are the greatest buyers of consumer goods. In the United States, up to the mid-1930s, such magazines were largely "trade-papers for home-makers." There were exceptions, such as True Story (1919–), which concentrated on entertainment, and Vogue, which had windows onto a wider world, but more typical was Better Homes and Gardens (1922–), which gave fresh impetus to the trend toward "service" by helping women and their husbands in the running of their homes. In this area, of course, advertising pressure can be considerable—*e.g.*, for editorial support of a new product—but editors have usually contained it within some limits. An innovation in the 1930s was the store-distributed magazine. One of the first and most successful was Family Circle (1932–), given away in Piggly Wiggly supermarkets until 1946, when it was sold as a family monthly for five cents, increasing to 69 cents in the early 1980s. Equally successful were Woman's Day (1937–), published by a subsidiary of the Great Atlantic and Pacific Tea Company, and Better Living (1951–), sponsored by the Super Market Institute. During the 1930s, to combat falling circulations and to meet changes in taste, women's magazines broadened their base, as they did again in the 1950s, in a similar crisis. Among the casualties in 1957, however, was the long-running Woman's Home Companion; and in 1956 even *McCall's* came under "outside" commercial control. By 1980 more than 65 feminist-oriented political and literary magazines had been established, the best known being Ms. (1972–), a nonprofit magazine with a circulation of 500,000. Another general trend, in keeping with the growing youth of the United States, has been to direct appeal toward younger women, not only in the old magazines but also in such new ones as Seventeen (1944–), Ingenue (1959–), and 'Teen (1957–).

The muckraking era

Before they entered their most commercial phase, general magazines in the United States went through a muckraking period, in which they were responsible for some major social reforms. This began when *McClure's* Magazine for January 1903 contained three articles on social problems, including one by Ida M. Tarbell on the Standard Oil trust. They were in accord with the public's interest at the time and prompted a spate of others, in Arena, Collier's, Cosmopolitan, Everybody's, *Hampton's*, and, for a brief time, American Magazine (1876–1956). The movement ended in 1912, helped by the pressure of financial interests. Thereafter, magazines grew and survived largely on their formula—*i.e.*, their particular blend of material and style of presentation. Collier's, for instance, became known for its brevity, and American for its success stories, until 1929, when it adopted a family formula. Competition for market leadership led to promotion schemes, such as cut-price subscriptions canvassed through the post, and also to casualties, including *McClure's* in 1933. When Collier's died in 1957 with a circulation of 3,750,000, it was due to falling income through cut-price promotion, rising costs through bigger printing bills, and insufficient advertising revenue to bridge the widening gap. Before the veteran Saturday Evening Post ceased publication in 1969, it tried to save itself, ironically, by cutting its circulation to a third. Since the 1930s the ranks of the general magazine have been thinned steadily, partly by the effects of paperbacks and television and partly by the advent of specialized publications—news, picture, and digest magazines.

Advertising in Britain and Europe. Though the advertising revolution began in Britain at much the same time as in the United States, its course has been less explosive. By 1898, The Gentlewoman was pointing out in its first issue that every copy cost "nearly double the price for which it is sold." Yet Britain's Audit Bureau of Circulations was not set up until 1931, and membership remained small until the 1960s; for it was only then that consumer spending in Britain and hence advertising really began to soar, to be reflected characteristically in a boom in women's magazines. In the early part of the century, the old general magazines continued to flourish, with such additions as the *Windsor* Magazine (1895–1939), Pearson's Magazine (1896–1909), Argosy (1926–), which published only fiction, and the popular weekly John Bull (1906–64), which thrived on "revelations." Several American magazines, especially women's, began to come out in British editions, such as Vogue (1916–), Good Housekeeping (1922–), and Harper's Bazaar (1929–). Society periodicals lost ground after World War I to those catering to the so-called new poor and new rich created by the social upheaval. The fortnightly Queen (1861–1970), Woman's Weekly (1911–), and the monthly Woman and Home (1926–) and Woman's Journal (1927–) were joined by such popular weeklies as Woman's Own (1932–), Woman's Illustrated (1936–61), and, above all, Woman (1937–), the first to be printed by colourgravure. After reaching a circulation of 750,000 before the war, it climbed to 3,500,000 in the late 1950s and has remained Britain's leading advertising medium in the women's market. During World War II some of these magazines achieved a high degree of reader identification with their advice on how to make fruit flans from carrots and brassieres from old lace curtains. Following the arrival of mass affluence in the late 1950s, Britain had its first store-distributed magazines with Family Circle (1964–), a service-only Anglo-American production, and its sister publication, Living (1967–); while the trend toward youthful markets was indicated by She (1955–), broad and robust in outlook, Honey (1960–), *Annabel* (1966–), for newlyweds, and Petticoat (1966–), for 14–19 year olds, and 19 (1968–), a market leader with a circulation of more than 184,000 in the early 1980s. The death of many of the old general magazines, under the pressure of paperbacks and television, and the dearth of weekly illustrated~(see below) left room for a new advertising vehicle. The first to perceive this was Lord Thomson, who in 1962 brought out a colour magazine as supplement to the Sunday Times. Though it lost £900,000 in its first 18 months and was derided as a failure, it went on to make money and obliged the Observer and the Daily Telegraph to compete with their own colour supplements.

Youth magazines

In the rest of Europe the impact of advertising on magazines has been more delayed and less pronounced, partly because market prices of Continental magazines tend to be closer to the production cost. General magazines were fairly limited before World War II, but since then, as part of the economic expansion, there has been a rich crop, most of the Time and Life type (see below) but also of magazines for women. France has several with large circulations, including *Nous Deux*, *Elle*, and *Intimité*, while those in Germany include entries for all age groups, such as *Jasmin* for newlyweds and *Eltern* for parents. Though Scandinavia has fewer periodicals, it is worth noting that the top-circulation magazine in Finland, *Pirkka,* is a giveaway distributed through grocery stores.

Publications outside Europe and the United States. *Japan.* The outstanding early 20th-century personality in Japanese magazine publication was Noma Seiji, who published nine magazines, nearly all with six-figure circulations. World War II did not seriously affect periodicals; and, at the end of occupation in 1952, there were more than 2,000 of all kinds, including *Shufu No Tomo* (1917–56; "Woman's Friend"), *Yoiko No Tomo* (1924–1957; "Child's Friend"), and Le-no-Hikari (1925– ; "Light of Home").

Africa. Important publications in Africa have included the quarterly East Africa *Africana* (Nairobi, 1962–); the Zimbabwean *Africa Calls* (1960–), published every two months; the quarterly *Nigeria Magazine* (1933–); the quarterly *Pan African Journal* (Nairobi, 1967–); and, in South Africa, journals in Afrikaans. Elsewhere, magazines in African languages have increased, as have those in English and French—*e.g.*, *Black Orpheus* (Ibadan, Nigeria, 1957–), containing creative writing by Africans and West Indians.

India, Bangladesh, and Pakistan. Important 20th-century magazines in India, Bangladesh, and Pakistan include the *Illustrated Weekly of India* (Bombay, 1880–), a topical review for educated readers; the *Statesman Weekly* (Calcutta, 1924–), an illustrated digest of Indian news and views; the monthly general review *Current Events* (Dehra Dun, 1955–); *Thought* (New Delhi, 1949–), a political and economic weekly; the monthly *Akhand Anand* (1947–); and the weekly *Akashvani* (New Delhi, 1936–), *Dharmayug* (Bombay, 1950–), and *Mukhabir-I-Alam* (Morādābād, India, 1903–). *Sport and Pastime* (1947–), with offices in nine cities, is well illustrated. *Eve's Weekly* (Bombay, 1947–), in English, Urdu, and Hindustani, is a popular women's magazine. Bangladesh weeklies include *Bangladesh Sangbad* (1972–). Pakistani periodicals include the monthly *Subrang Digest* (1970–) and the weekly *Muslim World* (1961–).

South America. Argentina had a greater magazine circulation than any other nation in South America until the mid-1970s, when total circulation decreased by almost one-half (it has been slowly rising since then). The weekly rotogravure *Maribel* (1932–56) long had the highest periodical circulation in that area, closely followed by that of the women's weekly *Para ti* (1922–). Mexico's leading magazine in the early 1980s was the weekly *Selecciones del Reader's Digest;* others included the weeklies *El Libro Semanal* (1954–) and *Alarma* (1963–). Venezuelan periodicals include the weekly *Resumen* (1973–) and *Elite* (1925–).

News and photo magazines. The accelerated tempo of life in the 20th century, coupled with the bewildering amount of information appearing in print, suggested the need for more concise ways of presenting it. The first to show how it could be done and so give rise to a whole new class of periodical was the U.S. news magazine *Time* (1923–), founded by Briton Hadden and Henry Luce.

The Luce magazines

Time magazine. There had, of course, been news magazines before, both in Europe and the United States. *Time* magazine's immediate forerunner was the folksy *Pathfinder* (1894–1954), a weekly rewriting of the news for rural audiences. There had also been attempts at compression of the digest type (see below). But *Time* was the first to aim at a brief and systematic presentation of the whole of the world's news. It was based on the proposition that "People are uninformed because no publication has adapted itself to the time which busy men are able to spend simply keeping informed." Its beginning was amateurish and precarious; neither Hadden nor Luce had much experience when they started summarizing the news from bundles of daily papers (being perishable, news was not protected by copyright). But after 1928 it grew steadily, finding its market chiefly among the rising number of college graduates. What came to be known as the *Time* style was characterized, in the words of a later critic, by two great democratic ideals, disrespect for authority and reverence for success. It presented the news in narrative form, well researched and checked, in tightly packed sentences, with a general air of omniscience. In the 1930s, to ensure adequate sources of information, Time Inc. built up a large news-gathering organization of its own. It also branched out into other publications, including *Fortune* (1930–), summarizing business news, and *Life* (see below). In the early 1980s *Time* had a circulation of about 4,300,000.

Among the direct followers of *Time* in the United States were *Business Week* (1929–), *United States News* (1933–), and *Newsweek* (1933–), its nearest rival. Simiiar magazines appeared in Shanghai, *East* (1933), and in Britain, the *News Review* (1936), though the latter did not have a comparable success, partly because Britiain was so well supplied with national dailies. After World War II the United States had several news magazines of a regional nature, such as *Fortnight* (1946) in California and *Texas Week* (1946). *Time* has had its greatest influence, however, in postwar Europe, where such magazines as *L'Express* (1953–) in France, *Der Spiegel* (1947–) in West Germany, and *Panorama* (1962–) in Italy derive directly from it. J.-J. Servan-Schreiber, the owner of *L'Express*, remodelled it following a visit to Time-Life in America; he also started two business weeklies, *L'Expansion* and *Le Management. Der Spiegel* ("The Mirror") became famous for its aggressive, anti-authoritarian exposures of scandal and malpractice, while *Panorama* achieved a high standing in Italy and is often cited as a reliable source. The influence of *Time* can probably be traced in most of the news magazines, as in *Tiempo* (1942–) in Mexico or *Primera Plana* (1962–) in Argentina.

Picture magazines. Conciseness can also be achieved through pictures, which obviate the need for description. Illustrated news magazines began in the last century (see above), but they took an altogether new form as photography developed. The most influential, though by no means the first of the modern type, was undoubtedly the United States weekly *Life* (1936–72), started by Henry Luce and the staff of Time Inc.

Pictorial journalism grew up alongside advertising techniques, the tabloid, and the documentary film and was spurred on by the new miniature camera of the Leica type, with wide-aperture lens and high-speed shutter, which enabled top-grade photographs to be taken quickly under almost any conditions. Its pioneers were particularly active in Germany, until many had to flee the Nazis. One of them was the Hungarian Stefan Lorant, who developed the story in pictures with *Bilder Courier* in Berlin in 1926 and with the *Munchener illustrierte Presse* in the period 1927–33. He then went to Britain, where he started a pocket picture magazine, *Lilliput* (1937–60), and was the first editor of *Picture Post* (1938–57). Another pioneer and one of the originators of candid camera shots was a German, Erich Salomon, whose pictures in the London *Tatler* in 1928 prompted *Fortune* to invite him to America, where he inspired *Life* photographer Thomas McAvoy.

In November 1936, therefore, when *Life* first appeared, picture magazines were very much in the air. Only a month before, *Mid-Week Pictorial* (1914–37), an American weekly of news pictures, was restyled along the lines *Life* was to take but was quickly overwhelmed by it. Though expected to have a circulation of well under 500,000 copies, *Life* was running at 1,000,000 within weeks. Its first issue, 96 large pages of pictures on glossy paper for 10 cents, was a sellout, the opening picture brilliant: an obstetrician slapping a baby into consciousness, with the caption "Life begins." Over the years, it kept the promise of its prospectus: "To see life; to see the world; to witness great events; to watch the faces of the poor and the gestures of the proud; to see strange things. . . ." During World War II, which it covered with great accomplishment, it enlarged its operations with a fortnightly international edition, and in 1952 a Spanish-language edition was added for Latin America, *Life en Español.* In 1971 *Life* magazine's circulation was about

7,000,000, but its high costs were no longer being met by advertising income, and it ceased publication in December 1972; it was revived as a monthly in October 1978.

Imitators of *Life* magazine

Of the countless imitators of *Life*, many were American, such as *Focus, Peek, Foto,* and two of longer duration, *Pic* (1937–48) and *Click* (1938–44). Best known was *Look* (1937–71; revived 1979), a popular biweekly. It was founded by Gardner Cowles, Jr., who also started *Quick* (1949–53), a miniature magazine. Britain had two news picture magazines, *Picture Post* (1938–57), which acquired much prestige through its social conscience, and *Illustrated* (1939–58); their place was taken to some extent by the Sunday colour supplements (see above). Preeminent in the rest of Europe is the French *Paris-Match* (1949–), exceptionally well produced and well supplied with photographers; while West Germany has *Stern* (1948–), a glossy blend of light and serious material, and Italy, where magazines are read more than newspapers, has *Oggi* (1945–), which thrives on not-too-sensational disclosures, and the elegant *Epoca* (1950–). There are magazines similar to *Life* in a number of other countries, such as *Cruzeiro (c.* 1908–) in Brazil and *Perspectywy* (1969–) in Poland, and still more that follow the cheaper style of *Look,* such as *Manchete* (1952–) in Brazil, *Caretas* in Peru, or the Australian *Pix–People.*

Digests and pocket magazines. The need for concise reading matter that gave rise to *Time* and *Life* was met even more successfully, in terms of circulation, by a U.S. magazine that reprinted in condensed form articles from other periodicals. This was the pocket-size *Reader's Digest* (1922–), founded by DeWitt Wallace.

Reader's Digest

Reader's Digest magazine. Its forerunners in the United States (apart from those in Europe) were the *Literary Digest* (1890–1938), "a repository of contemporaneous thought and research as presented in the periodical literature of the world," started by two former Lutheran ministers, Isaac K. Funk and Adam W. Wagnalls; the *Review of Reviews* (1890–1937), founded by Albert Shaw to condense material about world affairs; and Frank Munsey's *Scrap Book* (1906–12), "a granary for the gleanings of literature." The *Literary Digest,* in particular, with a circulation of more than 1,000,000 in the early 1920s, was something of a U.S. institution. Its famous straw votes successfully predicted the result of presidential elections after 1920, and its highly publicized wrong prediction of the outcome in 1936 played a decisive part in its collapse. *Reader's Digest,* however, was more specific in content and more universal in appeal. It aimed to supply "An article a day from leading magazines in condensed, permanent, booklet form." Each article, moreover, had to satisfy three criteria: "applicability" (it had to be of concern to the average reader); "lasting interest" (it had to be readable a year later); and "constructiveness" (it had to be on the side of optimism and good works).

The story of *Reader's Digest* is a classic of ingenuity and enterprise, The idea for it occurred to DeWitt Wallace, the son of a Presbyterian minister, when he was recuperating from war wounds in 1919. He prepared a sample, but no publisher was interested. So in 1921 he raised $5,000, and, with the help of his wife-to-be, Lila Bell Acheson, whose father was also a Presbyterian minister, he began to produce the magazine himself (first issue February 1922) from a basement office in Greenwich Village in New York City. After a year, with subscriptions running at about 7,000, the Wallaces moved to a garage and pony shed in Pleasantville, 40 miles north of New York City. Three years later, they were able to build their own house there, and in 1939, when circulation had reached 3,000,000, *Reader's Digest* moved into large premises at nearby Chappaqua. Until 1930 it was produced entirely by amateurs. Condensed books began to be added at the end of the magazine in 1934, and from this grew the Reader's Digest Condensed Book Club, with 2,500,000 members four years later. Overseas editions were started in 1939 (British), and foreign-language editions in 1940 (Spanish), others being steadily added over the following 10 years. In the early 1980s, *Reader's Digest* had one of the largest circulations of any magazine in the world: about 18,000,000 in 13 languages.

Problems of condensing articles

This success was not achieved entirely without setbacks and criticism. At first, permission to reprint was easy to obtain and without charge; but after a while, and especially after competitors entered the field and sometimes reprinted without permission, magazines began to regard the digests as parasitic. Charges were made, which rose steadily, and the major proprietors withheld their permission at various times. To guard against this and because articles of the sort he wanted were in short supply, Wallace began to print original material in the *Digest* in 1933. To keep up the appearance of a digest, articles were commissioned and then offered to other magazines in exchange for the right to "condense" and reprint them. Such articles, "cooperatively planned" accoiding to the *Digest,* "planted" according to critics, were naturally welcome to many magazines with slender budgets, but they did lead to controversy. In 1944, the *New Yorker,* fearing that *Reader's Digest* was generating too big a fraction of magazine articles in the United States, attacked the system as "a threat to the free flow of ideas and to the independent spirit"; but, in the more general view, the matter was regarded as a private one for the parties concerned. Internationally, too, the *Digest* has been attacked since World War II for its part in "American cultural imperialism"; even a U.S. senator has called it a "world cartel." But it has continued to find favour with the magazine public. Its formula, consistently applied from the beginning, has revealed the common denominator in men and women all over the world.

The digest idea was soon taken up by others, often in direct competition but also in more limited areas, such as *Science Digest, Catholic Digest, Negro Digest,* and *Children's Digest.* There was also a *Cartoon Digest* (1939), an *Editorial Digest* (1947), and a *Column Digest* (1949). Most of the general digests were forced to use original articles, since competition for the limited amount of highly popular material became too keen, and *Reader's Digest,* as first in the field, was always able to outbid everyone else. One of the more successful was *Magazine Digest* (founded 1930), based in Canada, with a good deal of scientific and technical matter. One that tried a new formula, based on timeliness and a liberal slant, was *Reader's Scope* (1943–48). The most successful book digest was probably *Omnibook* (1938–57), each issue of which contained abridgments of several popular books, fiction and nonfiction. The digests originally carried no advertising, but, since World War II, they have gradually been driven to it by rising costs. One of the last to capitulate was *Reader's Digest* in 1955 (though some of its foreign editions had always carried advertising), to avoid raising its price (25 cents until 1961); the proportion of advertising was restricted, however, to 20 percent. In general, competitors of *Reader's Digest* have tended to be short-lived and utterly dwarfed by it.

Types of pocket magazines. The success of *Reader's Digest* also had an influence through its format; it popularized the pocket magazine as a type. Several of the self-improving variety, such as *Your Life* (founded 1937) and *Success Today* (1946–50), were started by.Wilfred J. Funk on the proceeds from his father's *Literary Digest* (sold to *Time* in 1938). Of those more directly inspired by *Reader's Digest, Coronet* (1936–61), an offshoot of Esquire Inc., built up a large circulation during World War II, and when it closed, a victim of the promotion race, it was still running at more than 3,000,000. Somewhat livelier and glossier was·*Pageant*, first published in 1944. Britain had several pocket magazines, such as *London Opinion, Men Only,* and *Lilliput,* but these owed rather less to *Reader's Digest.* Finally, there have been a few "superdigests," miniature news magazines with pictures and a minimum of text, such as *Tempo* (1950), *People Today* (founded 1950), and *Jet* (1951–).

Specialized magazines. Though general magazines take the limelight, by far the largest number of periodicals are of the type that cater to specialist interests or pursuits. They vary greatly in circulation, but, even where it is small, it is usually stable over the short term and has the advantage, for an advertiser, of offering a well-defined market. Such magazines may be broadly classified into professional (including trade and technical journals) and nonprofessional.

Professional types. The professional magazine, often the organ of an association, keeps members informed of the latest developments, helps them to maintain standards, and defends their interests. Many were started in the last century, but since then, with the growth of specialization or with the need for a different viewpoint, they have greatly increased in number. Instead of two or three medical journals, for instance, there are now likely to be as many as a dozen, besides those in specialized areas such as dentistry, ophthalmology, and psychiatry. The same can be said of accounting, law, engineering, or any other profession. Though most of these magazines are of little interest to the general public, a few have authoritative articles of broader scope.

Controlled circulation magazines

Trade and technical journals serve those working in industry and commerce. During this century they too have grown enormously in numbers. Apart from all the traditional occupations, every major discovery in science, in manufacturing methods, or in business practice has led to a fresh subdivision of technology, with its own practitioners and, more often than not, its own magazine. The *Standard Periodical Directory* lists more than 65,000 magazines in 230 different categories, in the United States with subjects ranging from accounting and advertising through funeral service and hypnosis to sewage and zoology. Benn's Press Directory lists almost 5,600 periodicais in the United Kingdom, covering more than 340 topics. Such papers tend to be highly factual and accurately written, by people deeply immersed in their subjects. Most are well produced, often on art paper for the sake of the illustrations, and heavily dependent on advertising. Indeed, many are issued for a controlled circulation; *i.e.*, a publisher undertakes to distribute a magazine free of charge to a given number of specialist concerns, who can be relied upon to want a certain range of products. The manufacturers of these products, for their part, are naturally glad to have an advertising medium guaranteed to reach their particular market. The business papers may lack glamour, but they play a vital and highly influential part in economic life.

Nonprofessional types. Of the nonprofessional magazines, quite a number serve broad interest groups, religious, political, or social. Most churches have their journals and often more than one to meet the needs of ministers or sustain faith among the laity. As examples in Britain, one might mention the Catholic *Tablet,* the Anglican *Church Times,* and the Jewish *Chronicle* and, in the United States, the Catholic *Commonweal,* the Protestant *Christian Century,* and the Jewish *Commentary.* Though some of these magazines are subsidized as part of a drive to spread their message, most of them merely aim to foster corporate feeling among co-religionists. Much the same applies to political magazines in the narrow sense—*i.e.*, where they are issued by political organizations: they discuss doctrine, give news of activities, and forge links among members. More general political discussion tends to take place in the literary magazines (see below). Certain periodicals spring from the needs of particular social groups, examples being student magazines, magazines for blacks, or "underground" magazines, all of which see life from some shared point of view, which they encourage and clarify.

By far the largest number of specialized magazines for the layman fall into the hobby category. Very often a professional magazine has its amateur counterpart, as, for instance, in electronics, where the amateur has a wide range of technical magazines on radio, television, hi-fi, and tape recording. Other popular subjects are photography (the British *Amateur Photographer* dates from 1884) and motoring (Hearst's *Motor* was founded, as *Motor Cycling and Motoring,* in 1902); specialization even extends to types of camera and makes of car. No hobby or sport, indoor or outdoor, is without its magazine: boating, fishing, skin diving, philately, model railways, golf and tennis, ballroom dancing, chess, and many others. As soon as any activity becomes sufficiently popular, a magazine appears to cater to its adherents and to provide an advertising medium, not only for manufacturers and suppliers of special goods and services but also for readers, to help them buy and sell secondhand equipment.

"Pulp" and "comic" magazines

Some special tastes in entertainment are met by the "pulp" and "comic" magazines. In 1896, Frank Munsey turned his *Argosy* into an all-fiction magazine using rough wood-pulp paper. The "dime novel" did not qualify for cheap postal rates in the United States, but the pulp magazine did, and so an industry was born. Pulps began as adventure magazines but soon split up into further categories: love, detective, and western. Such magazines sold in millions up to the mid-1930s, when they gradually lost ground to the comics. These began as collections reprinted from the comic strips in newspapers; the first to appear regularly was *Famous Funnies* (1934). After 1937, however, with *Detective Comics,* they came into their own as original publications, and, like the pulps, they grew into a major industry, dividing up into much the same types. They may be seen, in effect, as pictorial condensations of the pulps. Though mainly for children, they were widely read by adults. "Comic" rapidly became a misnomer, as they played increasingly on horror and violence. While some defended them as harmless and cathartic, others condemned them as incitements to imitation. Attempts at control were made through legislation in the United States and elsewhere, and the industry itself tried to regulate standards. Television has since drawn off much of the criticism onto itself and aiso reduced the demand for comics; but they remain big business. One type of magazine, originally classed as pulp but growing to a significant stature, is the science-fiction magazine, which was pioneered by Hugo Gernsback's *Amazing Stories* in 1926.

The "fan" magazines offer glimpses of life behind the scenes in the world of entertainment and sport. In the heyday of motion pictures, numerous magazines appeared about films and their stars, beginning with *Photoplay* (1911) and *Picture Play* (1915) and a succession of others, such as *Movie Mirror* (1930) and *Movieland* (1942). When radio and TV became popular, similar magazines sprang up centring on broadcast programs and their personalities. One of their main functions was to provide a weekly timetable of programs. Today most countries have them, such as the *Radio Times* and *T.V. Times* in Britain, the *TV Guide* in America, *Télé 7 Jours* in France, and *Hor Zu!* in Germany, all with seven-figure circulations.

Finally, there are a number of what might be called "special service" magazines—*e.g.*, financial magazines to help the private investor, magazines of advice issued by consumer associations, magazines specifically for house hunters, racegoers, or for trading in secondhand goods, and so on. It is safe to say that no recognizable group of any size lacks its periodical.

Scholarly and literary magazines. Under the changed social circumstances of the 20th century and with many alternative forums for public debate, the old critical review lost some of its former glory, but it continued to have an influence quite out of proportion to its circulation. One may distinguish broadly between the scholarly type of review, the more widely read politico-cultural periodical, and the purely literary magazine.

Britain. Many of the British reviews founded in the 19th century (see above) have continued to flourish. Among additions of the scholarly type were the *Hibbert Journal* (1902–70), a nonsectarian quarterly for the discussion of religion, philosophy, sociology, and the arts; the *Times Literary Supplement* (1902–), important for the completeness of its coverage of all aspects of books and bibliographical matters; *International Affairs* (1922–), the journal of Chatham House, the Royal Institute of International Affairs, containing serious, long-term articles; and the *Political Quarterly* (1930–), for the discussion of international social and political questions from a progressive but nonparty point of view. Of the weekly political reviews, the long-established *Spectator* (1828–) represents the right and the *New Statesman* (1913–), founded by Sidney and Beatrice Webb, the left, though both in a broad context; while *Time and Tide* (1920–79), originally founded by Lady Rhondda as an independent

Political reviews

journal, was an influential news magazine. Several other periodicals meet the need for serious articles on current questions; among them are the *Economist* (1843–), which carries more general material than its name implies; *The Listener* (1929–), published by the British Broadcasting Corporation and consisting mainly of radio talks in printed form; the *New Scientist* (1956–), drawing attention to current scientific work; and *New Society* (1962–), which does the same for sociology. Among the literary magazines that came and went, but not without leaving their mark, were the *Egoist* (1914–19), associated with Ezra Pound and the Imagists; the *London Mercury* (1919–39), started by J.C. Squire, one of the Georgian poets; the *Criterion* (1922–39), founded and edited by T.S. Eliot; the *Adelphi* (1923–55), of John Middleton Murry, friend of D.H. Lawrence; *New Writing* (1936–46), edited by John Lehmann, who also later revived the old *London Magazine* (1954–); and *Horizon* (1940–50; 1958–), which Cyril Connolly started as a medium for literature during the war years. Since then, *Encounter* (1953–), an international review originally sponsored by the Congress for Cultural Freedom, has proved an intellectual magazine of value and distinction. In addition, many "little magazines" have struggled along, as always, providing essential seedbeds for new writers.

The United States. U.S. counterparts to British scholarly journals would include the *Political Science Quarterly,* edited by the political science faculty of Columbia University; the *American Scholar* (1932–), "a quarterly for the independent thinker" edited by the united chapters of Phi Beta Kappa; *Foreign Affairs* (1922–), a quarterly dealing with the international aspects of America's political and economic problems; and *Arts in Society* (1958–), a forum for the discussion of the role of art, which also publishes poetry and reviews. Of general political journals, the oldest still in publication in the early 1980s was *The Nation* (1865–), founded by E.L. Godkin and edited in the period 1918–34 by Oswald Garrison Villard. By tradition it adopted a critical stand on most matters, disdaining approval by the majority; it was notable for the "casual brilliance" of its literary reviews. When the muckraking phase in the popular magazines died down, zeal for reform was left to a succession of little magazines that led precarious lives, often needing extra support from loyal readers or rich individuals. Such were the *Progressive* (1909), of the La Follette family; *The Masses* (1911–17), run by the Greenwich Village Socialists; and the *New Republic* (1914–), which was started by Herbert Croly with the backing of the Straight family as "frankly an experiment" and "a journal of opinion to meet the challenge of the new time" and which survived as a liberal organ, after many triumphs and vicissitudes. Between the wars came the Marxist *Liberator* (1918–24); the *Freeman* (1920–24 and 1950–54), founded to recommend the single-tax principle of Henry George and later revived as a Republican journal; the *New Leader* (1927), for 10 years the organ of the American Socialist Party; and the extreme left *New Masses* (1926–48). Postwar foundations included the anti-Communist *Plain Talk* (1946–50); the fortnightly *Reporter* (1949–68), strong on "facts and ideas"; and the conservative *National Review* (1955–). Of the literary magazines, the *Atlantic* and *Harper's* were joined by the *American Mercury* (1924–), which had a brilliant initial period under Henry L. Mencken and George Jean Nathan, when it published work by most of the distinguished writers of the time; and the *Saturday Review* (1924–), which began as a purely literary magazine but broadened its scope in the 1940s. In 1972 a new ownership brought additional changes. A powerful influence on American writing has been exerted by the *New Yorker* (1925–), mainly through its founder Harold Ross, a perfectionist among editors. It became famous for its cartoons and probing biographical studies. Finally, there has been no lack of "little magazines" to foster talent; in the early 1980s there were more than 2,000 such magazines.

Europe. Among the numerous literary magazines in Europe, several in France and Germany in particular may be mentioned. The *Mercure de France* was revived in 1890 as an organ of the Symbolists; the influential *Nouvelle Revue française* (1909–) aimed at a fresh examination of literary and intellectual values; and the *Nouvelles Littéraires* (1922–) was founded by Andre Gillon as a weekly of information, criticism, and bibliography. Since World War II there have been Jean-Paul Sartre's left-wing monthly *Les Temps modernes* (1945–); *La Table ronde* (1948–); and *Les Lettres nouvelles* (1953–). In Germany, political magazines at the beginning of the century included the radical *Die Fackel* (1899; "The Torch") and *Die neue Gesellschaft* (1903–07; "The New Society") of the Social Democrats. An important literary influence was *Blatter fur die Kunst,* associated with the Neoromantic movement of Stefan George. The Nazi period imposed an artificial break in development, but since the war there have been the liberal weekly *Die Zeit* and a number of literary journals such as *Westermanns Monatshefte, Neue deutsche Hefte,* and *Akzente.*

Reviews in Communist countries

The political involvement of the literary review has been especially marked in the Soviet Union and iron curtain countries. The *Literaturnaya Gazeta* (1929–) and the influential *Novy Mir* (1925– ; "New World") have often been the centre of controversy in the Soviet Union when writers have been condemned for their views or denied the opportunity to publish. This has led to a strong underground press. In Czecholovakia the *Literárne Listy* played a prominent part in the freedom movement of 1968 and was later suppressed at Soviet insistence, along with the *Reportér* and *Student,* leading again to the starting of several underground magazines. Even *Sinn und Form* (1949–), a Marxist critical journal in East Germany, has been subject to temporary suspensions by the government for publishing such banned authors as Sartre, Kafka, and Hemingway.

CURRENT PROBLEMS AND TRENDS

Magazines in the second half of the 20th century are faced with rising production costs, falling advertising revenue, and competition from paperbacks and television. Their most expansive phase is almost certainly over, and there is a marked trend away from the general-interest magazine toward specialization. Yet the industry as a whole is far from declining; it is merely changing, as it always has done, in response to social change and technical development.

Production costs. Rising costs of production are a fairly universal phenomenon that can only be offset by increased efficiency. Some help in stabilizing costs in the magazine industry may come eventually through electronic techniques, but these are still in their infancy, and the cost of installing equipment is itself an inhibiting factor. It has been suggested that greater gains in economy might be obtained by rationalizing the process of manufacture. Traditionally, a magazine is set up, printed, and bound under one roof, but successful experiments have been made in separating these operations. Such specialization has obvious organizational advantages, and overheads are generally confined to each specific operation without penalizing any other. Closer collaboration between publisher and printer is also currently recommended, to avoid false lines of development, leading to either bottlenecks or idle capacity.

Competition with television

Increased costs have naturally been reflected in higher charges for advertising, which have weakened the competitive position of magazines with respect in particular to television advertising. It seems that in the early days of the television boom any advertising through that medium was extra and did not cut into magazine advertising, but this is no longer true. Television's share of advertising has been rising, and that of magazines has been falling. Yet there are signs that this trend is past its maximum. The cost of advertising, per message per 1,000 of audience, is higher in magazines than in the daily press or on television, but it has a much greater "recall value," which gives it a strong fundamental position.

Circulation. High circulations are, of course, the *sine qua non* of high advertising rates, but there are levels at which returns begin to diminish, and this has led to the anomaly of deliberate reductions in circulation. In 1970, *Life* magazine, with a circulation of 8,500,000 and a charge of $67,000 for a full-page colour advertisement, found it was losing 14 percent of its advertising revenue. It therefore restricted circulation to 7,000,000 and dropped its charge to the more viable rate of $54,000, but even this reduction could not prevent the magazine's demise. Similarly in Britain, when its circulation of 3,500,000 was proving an embarrassment to *Woman* in 1957, *Woman's Realm* (1958–) was launched to absorb the excess. A solution adopted in America is "regionalization." For the sake of the advertising, *Time,* for instance, has split up into 200 different area editions. It seems likely that much can be done through flexibility of makeup to enable a magazine to catch national, regional, and local advertising.

Playboy magazine

While some magazines, as always, are on the way down, numbers of periodicals continue to rise in almost every country, and there is no indication of any decline in the total amount of money that is spent on them. It is true that magazines relying heavily on entertainment have been affected by television, just as those relying on fiction have been hit by paperbacks. But categories of magazines have become somewhat blurred; *Life,* for instance, was "illustrated," "information," and "entertainment." In general, the "all or none" principle applies—*i.e.*, the media tend to reinforce each other; keen television viewers, for instance, are also keen magazine readers. Moon shots are exciting on television but are more permanent in a magazine; news is fresher in a daily paper, but interpretation is less transient in a magazine. The mortality rate among periodicals is traditionally high, as are the odds against success for a new one. Yet a clear editorial intent, backed by adequate capital, can still succeed on the scale of *Playboy* (1953–) in the United States, for instance, with a circulation of about 5,200,000, or, more modestly, of the woman's magazine *Nova* (1965–75) in Britain. *Nova* got off to a magnificent start as "The New Magazine for a New Kind of Woman," but it lost ground, significantly, when it began to waver editorially between features and femininity. As society grows more complex, the need is not so much for a comparatively small number of mass-circulation magazines, as in the past, as it is for a large number of periodicals to serve the many subgroups that are continually forming. As leisure time has increased, so too has the number of magazines of the hobby or special-interest type; and as educational standards have risen, so too has the broadly educational content of many magazines. Every new area of interest, such as ecology and conservation, quickly produces a spate of new magazines. These trends seem likely to continue and become more pronounced.

BIBLIOGRAPHY. O.H. CHENEY, *Economic Survey of the Book Industry,* 1930–31 (1931, reprinted 1960), a massive work covering all aspects of U.S. publishing and bookselling that still has relevance to present-day problems; COLIN CLAIR, *A History of Printing in Britain* (1965), a detailed account of individuals and their contributions, particularly useful for the early period in England; EUGENE EXMAN, *The House of Harper* (1967), a full-scale account of a famous American publishing firm, from its founding down to the present—especially useful for the 19th century and early U.S. copyright complications; GINN AND COMPANY, *Seventy Years of Text Book Publishing: A History of Ginn and Company,* 1867–1937 (1937), outlines the state of U.S. illiteracy after the Civil War and the opportunities for educational publishing that were ably exploited; HENRY HOLT, *Sixty Years As a Publisher* (1924), personal memories of a famous American publisher, valuable for details of trade practices and early copyright arrangements (or lack of them): ALLEN KENT and HAROLD LANCOUR (eds.), *Copyright: Current Viewpoints on History, Laws, Legislation* (1972), a useful collection of essays on copyright from varied professional viewpoints; R.J.L. KINGSFORD, *The Publishers Association* 1896–1946 (1970), a detailed and fully documented history of the first 50 years of the association that covers all the main problems of book trade organization and attempts at their solution in Britain; W.J. LEAPER, *Copyright and Performing Rights* (1957), a legal handbook covering the history of copyright in England and the implications of the Berne Convention and the Universal Copyright Convention; H. LEHMANN-HAUPT, *Das Amerikanische Buchwesen* (1937; Eng. trans., *The Book in America,* 2nd ed., 1951), one of the few general books on the subject, especially good in relating American publishing to its historical background; LENOX LOHR, *Magazine Publishing* (1932), a detailed account, department by department, of the work involved in the editing and management of a weekly or monthly magazine; RUARI MacLEAN, *Magazine Design* (1969), a handsomely produced royal octavo volume containing reproductions of the covers of famous magazines of the United States and of Britain and other European countries; JOHN C. MERRILL, CARTER R. BRYAN, and MARVIN ALISKY, *The Foreign Press* (1970), a very complete and concise survey of the current state of newspaper publishing throughout the world, taking the American press as its point of departure and also containing some data on magazines; F.A. MUMBY, and IAN NORRIE, *Publishing and Bookselling,* 5th ed. (1974), traces the history of the trade from classical and medieval times to the beginnings of the modern approach in the 17th century and then deals most fully with the 19th and 20th centuries in Britain; THEODORE PETERSON, *Magazines in the Twentieth Century,* 2nd ed. (1964), a scholarly survey describing not only the course of the development of some principal U.S. magazines but also the social and economic factors that affected their fortunes; G.H. PUTNAM, *Memoirs of a Publisher,* 1865–1915 (1915), useful background material on post-Civil War social economics and the effects on book buying, also on Japanese literary piracy following the Perry expedition; E. LLOYD SOMMERLAD, *The Press in Developing Countries* (1966), a discussion from every angle of the many problems involved in bringing out a newspaper in countries just beginning to have a regular press; S.H. STEINBERG, *Five Hundred Years of Printing,* 2nd ed. rev. (1962). a most complete and readable account for the general reader, covering all its aspects of printing and its impact on civilization; GRAHAM STOREY, *Reuters' Century,* 1851–1951 (1951), a thorough account of the history of news gathering, told primarily from Reuters' point of view, which does full justice to much else, including the great U.S. agencies; JOHN TEBBEL, *A History of Book Publishing in the United States,* 4 vol. (1972–81), a detailed and interesting account of book publishing in the United States from 1930 to 1980; UNESCO, *World Communications,* 5th ed. (1975), a worldwide survey, including statistics on the press, news agencies, radio, television, and films; CYNTHIA WHITE, *Women's Magazines,* 1693–1968 (1971), a history of women's magazines in Britain, with a chapter on American equivalents, by a social historian; FRANCIS WILLIAMS, *The Right to Know* (1969), a detailed account of the world press, with emphasis on Britain, the United States, and leading European countries, by a former editor of the London *Daily Herald* and a governor of the BBC, who regards newspapers as unique barometers of their times that indicate the climate of the societies from which they spring; RICHARD WINCOR and IRVING MANDELL, *Copyright, Patents and Trademarks* (1980), a legal handbook covering the history of copyright in the United States, including the copyright act of 1976; JAMES P. WOOD, *Of Lasting Interest: The Story of the Reader's Digest* (1958), outlines the early life of DeWitt Wallace and describes in vivid detail the conception, launching, and brilliant development of the most famous and successfui of the digests.

(P.U./G.U.)

Puccini, Giacomo

Almost exclusively a composer of opera, Giacomo Puccini virtually brought the history of Italian opera to an end. He wrote what are perhaps the only Impressionistic Italian operas, at least, to the extent that they avail themselves of Impressionistic musical technique.

Puccini's approach to dramatic composition is expressed in his own words: "The basis of an opera is its subject and its treatment." The fashioning of a story into a moving drama for the stage claimed his attention in the first place, and he devoted to this part of his work as much labour as to the musical composition itself. The action of his operas is uncomplicated and self-evident, so that the spectator, even if he does not understand the words, readily comprehends what is taking place on the stage.

Giacomo Antonio Domenico Michele Secondo Maria Puccini was born on December 22, 1858, in Lucca, Italy. He was the last descendant of a family that for two centuries had provided the musical directors of the Cathedral of S. Martino in Lucca. Puccini initially dedicated himself to music, therefore, not as a personal vocation but as a family profession. He was orphaned at the age of five by the death of his father, and the municipality of Lucca supported the family with a small pension and kept the position of cathedral organist open for Giacomo until he became of age. He first studied music with two of his father's former pupils and he played the organ in small local churches. A performance of Verdi's *Aida*, which he saw in Pisa in 1876, convinced him that his true vocation was opera. In the autumn of 1880 he went to study at the Milan Conservatory, where his principal teachers were Antonio Bazzini, a famous violinist and composer of chamber music, and Amilcare Ponchielli, the composer of the opera *La gioconda*. On July 16, 1883, he received his diploma and presented as his graduation composition *Capriccio sinfonico*, an instrumental work that attracted the attention of influential musical circles in Milan. In the same year, he entered *Le villi* in a competition for one-act operas. The judges did not think *Le villi* worthy of consideration, but a group of friends, led by composer-librettist Arrigo Boito, subsidized its production, and its premiere took place with immense success at Milan's Teatro dal Verme on May 31, 1884. *Le villi* was remarkable for its dramatic power, its operatic melody, and, revealing the influence of Wagner's works, the important role played by the orchestra. The music publisher Giulio Ricordi immediately acquired the copyright, with the stipulation that the opera should be expanded to two acts. He also commissioned Puccini to write a new opera for La Scala and gave him a monthly stipend: thus began Puccini's lifelong association with Giulio Ricordi, who was to become a staunch friend and counsellor.

Alinari—Mansell

Puccini.

Elopement with Elvira

After the death of his mother, Puccini fled from Lucca with a married woman, Elvira Gemignani. Finding in their passion the courage to defy the truly enormous scandal generated by their illegal union, they lived at first in Monza, near Milan, where a son, Antonio, was born. In 1890 they moved to Milan, and in 1891 to Torre del Lago, a fishing village on Lake Massaciuccoli in Tuscany. This home was to become Puccini's refuge from life, and he remained here until three years before his death, when he moved to Viareggio. But living with Elvira proved difficult. Tempestuous rather than compliant, she was justifiably jealous and was not an ideal companion. The two were finally able to marry in 1904, after the death of Elvira's husband. Puccini's second opera, *Edgar*, based on a verse drama by the French writer Alfred de Musset, was performed at La Scala in 1889 and it was a failure. Nevertheless, Ricordi continued to have faith in his protégé and sent him to Bayreuth in Germany to hear Wagner's *Die Meistersinger*.

Puccini returned from Bayreuth with the plan for *Manon Lescaut* based, like the *Manon* of the French composer Massenet, on the celebrated 18th-century novel by the Abbé Prévost. Beginning with this opera, Puccini carefully selected the subjects for his operas and spent considerable time on the preparation of the libretti. The psychology of the heroine in *Manon Lescaut*, as in succeeding works, dominates the dramatic nature of Puccini's operas. Puccini, in sympathy with his public, was writing to move them so as to assure his success. The score of *Manon Lescaut*, dramatically alive, prefigures the operatic refinements achieved in his mature operas: *La Bohème* (1896), *Tosca* (1900), *Madarna Butterfly* (1904), and *La fanciulla del west* (1910, *The Girl of the Golden West*). These four mature works also tell a moving love story, one that centres entirely on the feminine protagonist and ends in a tragic resolution. All four speak the same refined and limpid musical language of the orchestra that creates the subtle play of thematic reminiscences. The music always emerges from the words, indissolubly bound to their meaning and to the images they evoke. In *Bohème*, *Tosca*, and *Butterfly*, he collaborated enthusiastically with the writers Giuseppe Giacosa and Luigi Illica. The first performance (February 17, 1904) of *Madama Butterfly* was a fiasco, probably because the audience found the work too much like Puccini's preceding operas.

Characteristics of mature work

In 1908, having spent the summer in Cairo, the Puccinis returned to Torre del Lago and Giacomo devoted himself to *Fanciulla*. Elvira unexpectedly became jealous of Doria Manfredi, a young servant from the village who had been employed for several years by the Puccinis. She drove Doria from the house threatening to kill her. Subsequently, the servant girl poisoned herself, and her parents had the body examined by a physician, who declared her a virgin. The Manfredis brought charges against Mrs. Puccini for persecution and calumny, creating one of the most famous scandals of the time. Elvira was found guilty, but through the negotiations of the lawyers was not sentenced, and Puccini paid damages to the Manfredis, who withdrew their accusations. Eventually the Puccinis adjusted themselves to a coexistence, but the composer from then on demanded absolute freedom of action.

The premiere of *La fanciulla del west* took place at the Metropolitan in New York on December 10, 1910, with Arturo Toscanini conducting. It was a great triumph, and with it Puccini reached the end of his mature period. He admitted "writing an opera is difficult." For one who had been the typical operatic representative of the turn of the century, he felt the new century advancing ruthlessly with problems no longer his own. He did not understand contemporary events, such as World War I.

In 1917 at Monte-Carlo in Monaco, Puccini's opera *La rondine* was first performed, and then was very quickly forgotten.

Always interested in contemporary operatic compositions, Puccini studied the works of Claude Debussy, Richard Strauss, Arnold Schoenberg, and Igor Stravinsky. From this study emerged the *Il trittico* (*The Triptych;* New York, 1918), three stylistically individual one-act operas—the melodramatic *Il tabarro (The Cloak)*, the sentimental *Suor Angelica*, and the comic *Gianni Schicchi*. His last opera, based on the fable of *Turandot* as told in the play *Turandot* by the 18th-century Italian dramatist Carlo Gozzi, is the only Italian opera in the Impressionistic style. Puccini did not complete *Turandot*, unable to write a final grand duet on the triumphant love between Turandot and Calaf. Suffering from cancer of the throat, he was ordered to Brussels for surgery, and a few days afterward, on November 29, 1924, he died with the incomplete score of *Turandot* in his hands.

Turandot was performed posthumously at La Scala on April 25, 1926, and Arturo Toscanini, who conducted the performance, concluded the opera at the point Puccini had reached before dying. Two final scenes were completed by Franco Alfano from Puccini's sketches.

Solemn funeral services were held for Puccini at La Scala in Milan, and his body was taken to Torre Del Lago, which became the Puccini Pantheon. Shortly afterward, Elvira and Antonio were also buried there. The Puccini house became a museum and an archive.

The themes of the operas

The majority of Puccini's operas illustrate a theme defined in *Il tabarro:* "Chi ha vissuto per amore, per amore si mori" ("He who has lived for love, has died for love"). This theme is played out in the fate of his heroines—women who are devoted body and soul to their lovers, are tormented by feelings of guilt, and are punished by the infliction of pain until in the end they are destroyed. In his treatment of this theme, Puccini combines compassion and pity for his heroines with a strong streak of sadism: hence the strong emotional appeal but also the restricted scope of the Puccinian type of opera.

The main feature of Puccini's musicodramatic style is his ability to identify himself with his subject; each opera has its distinctive ambience. With an unfailing instinct for balanced dramatic structure, Puccini knew that an opera is not all action, movement, and conflict; it must also contain moments of repose, contemplation, and lyricism. For such moments he invented an original type of melody, passionate and radiant, yet marked by an underlying morbidity; examples are the "farewell" and "death" arias that also reflect the persistent melancholy from which he suffered in his personal life.

Puccini's conception of diatonic melody is rooted in the tradition of 19th-century Italian opera, but his harmonic and orchestral style indicate that he was also aware of contemporary developments, notably the work of the Impressionists and of Stravinsky. Though he allowed the orchestra a more active role, he upheld the traditional vocal style of Italian opera, in which the singers carry the burden of the music. In many ways a typical *fin-de-siècle* artist, Puccini nevertheless can be ranked as the greatest exponent of operatic realism.

MAJOR WORKS

OPERAS: *Le villi* (first performed 1884); *Edgar* (1889); *Manon Lescaut* (1893); *La Bohème* (1896); *Tosca* (1900); *Madama Butterfly* (1904); *La fanciulla del west* (1910, *The Girl of the Golden West); La rondine* (1917); *Il trittico: Il tabarro, Suor Angelica, Gianni Schicchi* (1918); *Turandot* (1926).

BIBLIOGRAPHY. Complete catalogs of Puccini's compositions may be found in MOSCO CARNER, *Puccini: A Critical Biography,* 2nd ed. (1977); and CLAUDIO SARTORI, *Puccini,* 2nd ed. (1963), in Italian; a complete catalog of recordings in EUGENIO GARA, *Carteggi Pucciniani* (1958). The principal collections of autographed compositions and letters are at the Puccini Villa in Torre del Lago; the archives of the Ricordi House in Milan; and the library of the Boccherini Musical Institute in Lucca. Puccini's letters have been published in *Carteggi Pucciniani* (cited above); and in GIUSEPPE ADAMI (ed.), *Letters of Giacomo Puccini,* new ed. (1974). English translations of the letters to Sybil Seligman have been published in VINCENT J. SELIGMAN, *Puccini Among Friends* (1938, reprinted 1971). Other works on the life and music of Puccini include: GEORGE R. MAREK, *Puccini* (1951), an exhaustive biography; DANTE DEL FIORENTINO, *Immortal Bohemian* (1952); RICHARD SPECHT, *Giacomo Puccini; das Leben, der Mensch, das Werk* (1931; Eng. trans., *Giacomo Puccini, the Man, His Life, His Work,* 1933, reprinted 1970); WILLIAM ASHBROOK, *The Operas of Puccini* (1968); and HOWARD GREENFELD, *Puccini* (1980).

(C.Sa.)

Puerto Rico

An autonomous political entity in voluntary association with the United States according to a 1953 UN resolution, the Commonwealth of Puerto Rico occupies a central position among the West Indian islands. Its land rises sharply from the tropical sea, and its coastal plain quickly ascends to meet the steep mountain ranges that give the island the appearance of a nearly rectangular pyramid. By air it is about 1,600 miles (2,600 kilometres) from New York City, 1,050 miles from Miami, and 550 miles from Caracas. To the west lie the other, larger islands of the Greater Antilles—Hispaniola (containing the Dominican Republic and Haiti), Jamaica, and Cuba—while to the east is St. Thomas, of the Virgin Islands group, the closest of the myriad islands of the Lesser Antilles.

Impact of Spanish tradition

San Juan was the name given the island when it was discovered by Columbus in 1493; its capital city was known as Puerto Rico (Spanish: "rich harbour"). In the course of centuries, during which it played an integral role in Spain's empire in the Americas, the names for island and city were interchanged. At the same time its people developed traditions and a way of life that remain deeply rooted in Spanish culture. This factor, even more than the island location, helps to isolate the Puerto Rican from the northern European culture that shaped most of the institutions of other U.S. citizens. Pride in this unique heritage has been a major emotional component in the continuing internal debate among proponents of U.S. statehood, independence, and the maintenance of a commonwealth status.

Despite their heritage, Puerto Ricans began a massive exodus to mainland cities in the 1950s, largely because of an exploding population, which by 1980 was more than 3,100,000. Many persons lacking the skills to participate in the island's industrial development found themselves crushed into a hilly land whose area, counting several small adjacent islands, totals only 3,435 square miles (8,897 square kilometrcs). By the early 1970s the exodus to the mainland had been slowed but not completely stopped, and escape from crushing poverty in some areas remained a definite factor of life in Puerto Rico, along with a critical shortage of adequate housing and an unemployment rate more than twice the national average of the United States. In spite of these continuing problems, Puerto Ricans in general enjoy the highest per capita personal income and the highest standard of living in Latin America. They have come to represent a unique social and political community within both the United States and the Spanish-speaking nations of the Americas. (For information on related topics, see the articles CARIBBEAN SEA; UNITED STATES, HISTORY OF THE; and SPAIN, HISTORY OF.)

THE HISTORY OF PUERTO RICO

Pre-Columbian habitation. The early inhabitants of Puerto Rico were Indians from either the Florida peninsula or the Orinoco River Basin in South America. The Indians whom Christopher Columbus met, when he discovered Puerto Rico on November 19, 1493, belonged to the South American Arawak stock, a peaceful, sedentary people, given to fishing and agriculture. They numbered from 30,000 to 50,000, but their hold on the island was already being challenged by the bellicose Caribs from the Lesser Antilles.

Exploration and early development. Puerto Rico became one of the earliest Spanish settlements in the New World. The Spanish explorer Juan Ponce de Leon landed from Hispaniola in 1508, and he founded the first town, Caparra, and initiated the first mining and agricultural efforts. Gold mining was briefly profitable, but production declined with dwindling Indian labour. Sugarcane became the leading agricultural product with the introduction of slaves from Africa, but by the end of the 16th century it still had failed to spur the development of the island.

The exposed position of the island, however, enhanced its strategic value to Spain and its enemies. It was attacked successively by the French, by the English under Sir Francis Drake in 1595 and the 3rd earl of Cumberland in 1598, and by the Dutch in 1625—attacks that forced Spain to turn the city of Puerto Rico into the most fortified place in the West Indies.

Puerto Rico remained largely undeveloped until the late 18th century, when efforts were made to improve conditions. The population increased from about 45,000 in 1765 to 155,000 in 1800, about which time trade, no longer restricted to the Spanish empire, began to develop with the United States and other nations.

The 19th century. Puerto Rico remained loyal to Spain throughout the Napoleonic Wars and the wars of Latin-American independence. In 1815 a royal order issued by Spain opened the island to trade and to colonization, especially by loyalists fleeing from rebellious colonies in South America. After 1830 Puerto Rico gradually developed into a plantation economy, based on three main crops: sugar, coffee, and tobacco. Sugar and molasses, sold mainly in the U.S. market, provided an important source of income for the Spanish government. Foreign settlers contributed to economic development, though the Spanish element attempted to maintain a tight monopoly and gave strong support to the military and administrative bureaucracies.

Reform movements

After the U.S. Civil War and Spain's attempted reconquest of Santo Domingo, a new generation pressed for the abolition of slavery, free trade, and political rights. The more radical members of the generation opted for independence and were responsible for a revolution in 1868 that was quelled quickly by Spanish forces. Though separatist elements, led by Ramon Emeterio Betances, joined the Cubans who were struggling for independence, they were unable to challenge Spanish power effectively. Moderate reformist groups eventually joined forces with Roman Baldorioty de Castro, a leader advocating self-government under Spain. In 1897 they succeeded in obtaining a charter from Spain granting the island broad powers of self-government.

Period of U.S. influence. The Spanish–American War (1898) soon ended this promising period of reform. Largely unopposed invading U.S. troops were received cordially by large segments of the population, and there was great hope for increased political rights and economic benefits. The Treaty of Paris ceded Puerto Rico to the United States and left to the U.S. Congress the determination of its political future.

The U.S. military ruled with little regard for political sensitivities. The Foraker Act of 1900, which provided for civil government and helped pave the way for free trade with the United States, was a step backward in the struggle for political rights, since most power was held by federal appointees.

Early U.S. governors were mainly preoccupied with Americanizing Puerto Rican institutions, language, and political habits, but had no clear policy on the island's eventual political status. This approach created strong resistance from many native leaders led by Luis Muñoz Rivera, who had fought for autonomy under Spain. In the meantime, free access of sugar to the mainland market created a profound change in the economy, with 75 percent of the population coming to depend on the sugar industry by 1920. The population increased from about 950,000 in 1899 to more than 1,540,000 in 1930. Glaring inequalities of wealth contributed to sharpened social and political tensions.

Politically, the U.S. Congress broadened the very limited basis of Puerto Rican self-government by granting an elective senate in 1917 and extending U.S. citizenship to Puerto Ricans, but the governor and other key executive officials continued to be appointed by the president. According to Supreme Court decisions, Puerto Rico was an organized but unincorporated territory of the United States.

Beginnings of status controversies

As economic conditions worsened, the 1930s became an era of hidden, potentially explosive crises. The Nationalist Party insisted on complete and immediate independence, while statehood was defended by the Republican Party. The Socialist Party, more concerned with social and economic reforms, formed a coalition with the Republicans that dominated the island's legislature in the 1930s. The Popular Democratic Party, founded by Luis Muiioz Marín, won the election of 1940 and pushed through a program with strong support from the rural electorate. In 1946 Pres. Harry S. Truman appointed Jesus T. Piñero as the first native Puerto Rican governor; a popularly elected governor followed in 1948. A congressional law of 1950, approved in a referendum by the people of Puerto Rico, called for a constitution to be drafted and approved by the insular electorate. On July 25, 1952, the Commonwealth of Puerto Rico was established, with Luis Muiioz Marín elected as its first governor.

Though enjoying strong popular support, the commonwealth failed to satisfy the independence and statehood groups. A moderate Independence Party was formed, and from 1952 to 1956 it was the second party on the island. Statehood-oriented leaders staged a comeback by 1960 and narrowly captured the governorship in 1968. An island-wide referendum in 1967 showed that 60 percent favoured the commonwealth and 39 percent preferred statehood, but most of the voters favouring independence abstained and continued to agitate for separation. In 1972 the Popular Democratic Party was returned to power over the New Progressive Party, while the Independence Party made small gains. The New Progressive Party won the governorship in 1976 and 1980 over the Popular Democratic Party, and the independence Party polled third.

THE NATURAL LANDSCAPE

Mountainous landscape

Surface features. Puerto Rico is a rugged and hilly island. Nearly half of the island lies 500 feet (150 metres) or more above sea level—about 20 percent between 500 and 1,000 feet and about 25 percent more than 1,000 feet. The island is divided into three main geographical regions: the mountainous interior, the northern plateau, and the coastal plains. The central mountain range, known as the Cordillera Central, rises to more than 3,000 feet, with the highest points at Cerro de Punta, 4,389 feet, and Monte Guilarte, 3,949 feet. The range slopes very steeply to the southeast. Although the northern slope is less steep, the rivers have eroded the area more thoroughly. The south-running rivers are short and torrential, but they are dry most of the time.

In the northwest the elevation of the northern plateau averages 100 feet near the coast and 700 feet toward the interior. The plateau is crossed by small hills, and toward the interior the land is covered by hundreds of hillocks and gullies. In the northeast the Sierra de Luquillo includes the rain forest of El Yunque, a jungle of tropical and subtropical trees and plants—giant ferns, orchids, and trailing vines. The whole area of 29,000 acres (12,000 hectares) is included in the Caribbean National Forest and is a major tourist attraction.

In the north the coastal plains run from Punta Borinquen in the west to Cabezas de San Juan in the extreme northeast. Adjacent to the coast the land is quite level, interrupted only by a few rock promontories and by lines of sand dunes along the shore. The only flatlands of importance are found in the alluvial plains of such rivers as the Grande de Arecibo, Cibuco, Grande de Loiza, and La Plata. The mountains reach close to the shore in the east, the rivers forming deep valleys; in the west the valleys are even deeper. In the south the coastal plains are narrower and more regular than on the northern coast and are studded with hillocks and sand dunes.

Relation of topography and precipitation

The island's varied precipitation is the direct result of its topography. The east–west mountains form a barrier to the dominant east-to-northeasterly winds, giving the north an abundance of rain. On rising over the Cordillera Central, the warm, humid air masses cool and lose much of their moisture, so that rain on the southern coast is scarce, and a dry climate predominates. The rugged and irregular topography accounts for the nearly 1,300 streams, but only 50 are true rivers. The northern slopes carry the main currents, many of them flowing into plains so low that marshes, moors, and some lakes are formed. Drainage is deficient and floods are common. The shorter rivers on the southern slopes are dry in winter but torrential in the wet season.

Climate. Lying within the tropical zone, Puerto Rico has a pleasant climate greatly influenced by the sea and the warm North Equatorial Current. Moisture-laden winds from the east and northeast bring on the frequent rainy periods of winter as they encounter occasional cold fronts that extend southward into the West Indies from the U.S. mainland. Temperatures in Puerto Rico very seldom fall below 60° F (16" C). Extreme temperatures are rare, with the highest recorded daily average at 89° F (32° C) and the lowest at 66° F (19° C).

Vegetation and animal life. Puerto Rico has relatively little animal life. Wildlife includes nonpoisonous snakes, several lizards, mongooses, and birds. The most common birds are the thrushes, tanagers, bullfinches, flycatchers, warblers, plovers, terns, and sandpipers. Fish abound in great variety, but with little economic exploitation. Plant life includes palm trees and mangrove, which flourish along the coast, and bamboo in great clumps along the roads and streams. African tulip trees, bougainvillea, hibiscus, poinsettias, a golden trumpet known as *canario,* and other plants splash vivid colour against a green and brown landscape throughout the year.

THE PEOPLE OF PUERTO RICO

Ethnic and racial blending. The people of Puerto Rico are historical products of a mixture of diverse ethnic strains. The main ethnic stocks are Spanish and African, and though the aboriginal Indians were either absorbed

Composite character of the people

or eliminated, some of their physical characteristics remain, particularly in the peoples of the mountains. Settlers from several European nations arrived at the island in the 19th century—Danish, French, Corsican, and some English and German. During the 20th century there was a limited flow of people from the United States and, during the 1950s and 1960s, a major influx of Cubans. Slavery, introduced in the early days of the Spanish conquest, never flourished as it did in Saint-Dorningue (Haiti under French control), Jamaica, and Barbados. In the early 19th century, when the emphasis was on sugar cultivation, thousands of African slaves were brought into Puerto Rico, reaching a total of more than 50,000 around 1850. Under Spanish rule many slaves earned their freedom. In 1846, for instance, free blacks numbered more than 175,000. Abolition of slavery was achieved in 1873 without violence.

Since the end of the 18th century the population gradually has become a blend of its various ethnic strains. The independent, isolated small farmers of the interior, known as jibaros, have become a symbol of the social and cultural heritage of the island. Despite the mixture of peoples and traditions, the Puerto Ricans are a very homogeneous people; no group is looked upon as a minority in either racial, ethnic, or linguistic terms.

Patterns of migration

Demography. Puerto Rico's annual growth rate increased from little more than 1 percent in the late 1960s to almost 3 percent in the 1970s, representing one of the world's highest growth rates. The establishment of family planning programs and other measures aimed at birth control have contributed to a sharp decline in birth rates since 1966. The birth rate still exceeds the death rate, however, yielding a natural rate of increase in 1977 of 1.7 percent, compared with 0.6 percent for the United States. Internal population growth has been affected by migrations to and from the United States. Prior to 1960 emigration was largely responsible for an extremely low rate of growth. From 1960 to 1965 migration declined, birth rates dropped slightly, and the annual growth rate rose to 2.5 percent. Migration increased again after 1965, though the great number of returning Puerto Ricans held the net emigration to about 7,000 persons in 1969. During the 1970s, however, the number of returning migrants exceeded the number of emigrants, resulting in an average in-migration of about 25,000 residents per year. In 1977 this trend reversed again, with a net emigration of approximately 20,000 people. Of the migratory workers and families that have stayed in the United States, most have concentrated in the New York City area, although there are growing Puerto Rican settlements in the Eastern and Midwestern states.

The majority of migrants are from rural areas. With low educational levels, little work experience, and often a poor command of English, their lot has not been easy in urban areas on the mainland, where they have suffered from discrimination, unemployment, cultural deprivation, and loss of identity.

Puerto Rico's population is predominantly young. The median age by 1978 was 22.2 years for women and 24.4 years for men. There has been a definite increase in the numbers of persons of middle or advanced years, along with steadily rising life-expectancy rates. About 60 percent of the Puerto Ricans lived in urban areas in the early 1970s, compared with 44 percent in 1960. Urban growth, one of Puerto Rico's most pressing problems, has increased the demand for housing, transportation, and services, problems that the government must resolve if it is to sustain and improve living standards.

Puerto Rico, Area and Population

	area		population	
	sq mi	sq km	1970 census	1980 census
Municipalities				
Adjuntas	67	174	19,000	19,000
Aguada	30	78	26,000	32,000
Aguadilla	36	93	51,000	55,000
Aguas Buenas	30	78	19,000	22,000
Aibonito	31	80	20,000	22,000
Añasco	40	104	19,000	23,000
Arecibo	127	329	73,000	87,000
Arroyo	15	39	13,000	17,000
Barceloneta	24	62	. . .*	19,000
Barranquitas	34	88	20,000	22,000
Bayamón	44	114	156,000	196,000
Cabo Rojo	71	184	26,000	34,000
Caguas	59	153	96,000	118,000
Camuy	47	122	20,000	25,000
Canovanas	28	72	. . .*	32,000
Carolina	45	116	108,000	166,000
Cataño	5	13	26,000	26,000
Cayey	50	130	38,000	41,000
Ceiba	27	70	10,000	15,000
Ciales	66	171	16,000	16,000
Cidra	36	93	24,000	28,000
Coamo	78	202	26,000	31,000
Comerio	28	73	19,000	18,000
Coroza!	42	109	25,000	28,000
Culebra	10	26	1.000	1,000
Dorado	23	60	17,000	26,000
Fajardo	31	80	23,000	32,000
Florida	10	26	. . .*	7,000
Guánica	35	91	15,000	19,000
Guayama	66	171	36,000	40,000
Guayanilla	42	109	18,000	21,000
Guaynabo	27	70	67,000	81,000
Gurabo	28	73	18.000	24.000
Hatillo	42	109	22,000	29.000
Hormigueros	11	28	11.000	14,000
Humacao	45	117	36.000	46.000
Isabela	56	145	30,000	37.000
Jayuya	39	101	14,000	15,000
Juana Diaz	61	158	36,000	44,000
Juncos	26	67	22,000	25,000
Lajas	61	158	17.000	21.000
Lares	62	161	25,000	27,000
Las Marias	47	122	8,000	9,000
Las Piedras	33	85	18,000	22,000
Loíza	25	65	. . .*	21,000
Luquillo	26	67	10,000	15,000
Manati	45	116	31,000	37,000
Maricao	37	96	6,000	7,000
Maunabo	21	54	11.000	12.000
Mayaguez	76	197	86,000	96,000
Moca	50	129	22,000	29,000
Morovis	39	101	19.000	21.000
Naguabo	53	137	18,000	21,000
Naranjito	28	73	20,000	24,000
Orocovis	63	163	20.000	19.000
Patillas	47	122	18,000	18,000
Peñuelas	45	116	16,000	19,000
Ponce	116	300	159.000	189,000
Quebradillas	23	60	16,000	20,000
Rincón	14	36	9,000	12,000
Río Grande	61	158	22,000	34,000
Sabana Grande	35	91	16,000	20,000
Salinas	69	179	22.000	26.000
San Germán	54	140	28,000	33,000
San Juan	47	122	463,000	435,000
San Lorenzo	54	140	28,000	32,000
San Sebastián	71	184	30,000	36,000
Santa Isabel	34	88	16,000	20,000
Toa 41ta	27	70	19,000	32,000
Toa Baja	24	62	46,000	78,000
Trujillo Alto	21	54	31,000	51,000
Utuado	115	298	35.000	34.000
Vega Alta	28	73	23,000	29,000
Vega Baja	46	119	35.000	47,000
Vieques	52	135	8,000	8,000
Villalba	37	96	19,000	21,000
Yabucoa	69	179	30,000	31,000
Yauco	55	142	35,000	38,000
Total Puerto Rico	3,422	8,866	2,712,000†	3,196,000†

*Adjusted estimates for post-1970 boundary changes not available.
†Figures do not add to total given because of rounding.
Source: Official government figures.

THE COMMONWEALTH'S ECONOMY

Industry and agriculture. Since the end of World War II the island has undergone an industrial transformation that has had profound social and economic effects. The shift from agriculture to industrial production has been largely the result of commonwealth policies that, since 1948, have encouraged private investors. Efforts at industrialization were begun in the 1940s by the Puerto Rico Industrial Development Company, a governmental corporation that was revamped in 1950 as the Economic Development Administration, popularly known as Fomento. Its program has been known as "Operation Bootstrap." Its early achievements were modest: 82 factories were created in the first three years and 6,200 people were employed. This well-published pioneer work succeeded in attracting new industries to the island. Income from manufacturing

Economic expansion

surpassed that from agriculture in 1955 and by 1978 was nine times greater. Net manufacturing income in 1978 was more than $3,000,000,000, whereas the agricultural income was less than $350,000,000. Per capita personal income rose steadily from about $280 to about $2,500 in 1977.

Natural resources. Puerto Rico's industrial achievement is the more remarkable because the island is poor in minerals. Iron, manganese, lead, and zinc occur in such small quantities that there is no important commercial exploitation. Explorations, however, have revealed rich copper deposits; negotiations between the government and the mining companies on their development have been undertaken. Several nickel deposits also have been discovered.

Trade relations. Rising living standards and purchasing power have made Puerto Rico one of the best customers of the United States; Puerto Rico's imports increased more than threefold from 1960 to 1976. The United States, in turn, has been Puerto Rico's best customer, though the balance of trade remains in favour of the United States. Puerto Rico's position is worsened by unfavourable trade balances with foreign countries. Tourism has become an important source of income, with tourist expenditures increasing more than sixfold from 1960 to 1977. Tourism expenditures represent about 5 percent of the gross national product, and the tourist industry is the third largest sector.

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of *government.* The commonwealth government resembles the government of a **U.S.** state. Separate

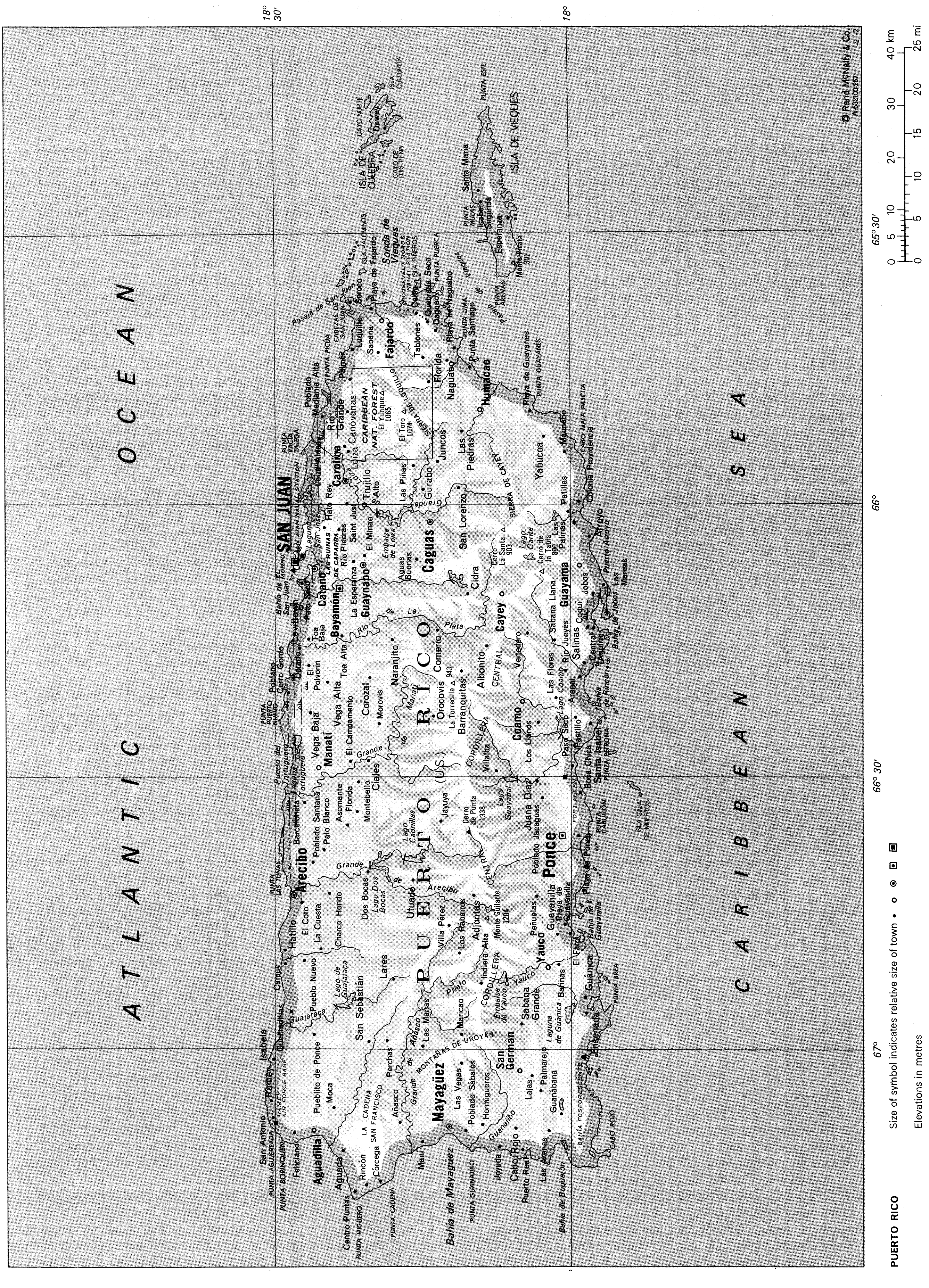
MAP INDEX

Cities and towns

Adjuntas	18.10n	66·43w
Aguada	18 23n	67·11w
Aguadilla	18·26n	67·09w
Aguas Buenas	18.15n	66·06w
Aibonito	18·08n	66·16w
Aiiasco	18·17n	67·08w
Arecibo	18·28n	66·43w
Arenal	17 59n	66·19w
Arroyo	17 58n	66·04w
Asomante	18·23n	66·36w
Barceloneta	18 27n	66·32w
Barinas	18.01n	66·51w
Barranquitas	18 11n	66·18w
Bayamón	18 24n	66·09w
Boca Chica	17 59n	66·32w
Cabo Rojo	18·05n	67·09w
Caguas	18 14n	66·02w
Camuy	18 29n	66·51w
Canóvanas	18·23n	65·54w
Carolina	18·23n	65·57w
Cataiio	18·27n	66·07w
Cayey	18·07n	66·10w
Ceiba	18·16n	65·39w
Central Aguirre	17·57n	66·13w
Centro Puntas	18·22n	67·16w
Charco Hondo	18·25n	66·43w
Ciales	18 20n	66·28w
Cidra	18·11n	66·10w
Coamo	18·05n	66·22w
Colonia Providencia	17·59n	66·00w
Comerío	18·13n	66·14w
Coquí	17 59n	66·14w
Corcega	18 19n	67·15w
Corozal	18·21■	66·17w
Daguao	18 14n	65·41w
Dewey	18·18n	65·18w
Dorado	18 28n	66·15w
Dos Bocas	18·20n	66·40w
El Campamento	18 22n	66·28w
El Coto	18 28n	66·44w
El Faro	38 00n	66·47w
El Minao	18 22n	66·05w
El Polvorín	18·26n	66·17w
Ensenada	17 58n	66·56w
Esperanza	18·06n	65·28w
Fajardo	18 20n	65·39w
Feliciano	18 28n	67·08w
Florida	18 14n	65·47w
Florida	18·22n	66·34w
Guanábana	18 01n	67·07w
Guánica	17 58n	66·55w
Guayama	17 59n	66·07w
Guayanilla. . . .	18·01n	66·47w
Guaynabo . .	18·22n	66·07w
Gurabo	18·16n	65·58w
Hatillo	18 29n	66·49w
Hato Rey	18 25n	66·03w
Hormigueros	18 09n	67·08w
Hurnacao	18·09n	65·50w
Indiera Alta	18·09n	66·53w
Isabela	18 30n	67·01w
Isabel Segunda	18·09n	65·27w
Jayuya	18·13n	66·36w
Jobos	17 58n	66·10w
Joyuda	18·07n	67·11w
Juana Díaz	18 03n	66·31w
Juncos	18 14n	65·55w
La Cuesta	18·25n	66·49w
La Esperanza	18 22n	66·07w
Lajas	18·03n	67·04w
Lares	18·18n	66·53w
Las Arenas	18·02n	67·09w
Las Flores	18·03n	66·22w
Las Mareas	17·56n	66·09w
Las Marías	18·15n	67·00w
Las Palmas	17 59n	66·02w
Las Piedras	18 11n	65·52w
Las Piñas	18·15n	65·55w
Las Vegas	18·11n	67·02w
Levittown	18 27n	66·09w
Loíza	18·23n	65·54w
Loiza Aldea	18 26n	65·53w
Los Llanos	18·03n	66·24w
Los Rabanos	18.11n	66·50w
Luquillo.............	18·22n	65·43w
Manatí.............	18·26n	66·29w
Maní................	18·15n	67·10w
Maricao	18·11n	66·59w
Maunabo	18·01n	65·54w
Mayaguez	18·12n	67·09w
Moca	18·24n	67·07w
Montebelo	18·22n	66·31w
Morovis	18·20n	66·24w
Naguabo	18·13n	65·44w
Naranjito	18·18n	66·15w
Orocovis	18·14n	66·23w
Palmareio	18·03n	67·05w
Palmer	18·22n	65·46w
Palo Blanco	18 26n	66·39w
Palo Seco	18·28n	66·09w
Paso Seco	17 59n	66·23w
Pastillo	17 59n	66·29w
Patillas	18·00n	66·01w
Peñuelas	18·03n	66·43w
Perchas	18·19n	66·59w
Playa de Fajardo	18·20n	65·38w
Playa de Guayanes	18·04n	65·49w
Playa de Guayanilla	18 01r	66·46w
Playa de Naguabo	18·12n	65·43w
Playa de Ponce	17 59n	66·37w
Poblado Cerro Gordo	1829'1	66·20w
Poblado Jacaguas	18·03n	66·32w
Poblado Medinía Alta	18·26n	65·50w
Poblado Sábalos	18·11n	67·09w
Poblado Santana	18·27n	66·40w
Ponce	18·01n	66·37w
Pueblito de Ponce	18 26n	66·58w
Pueblo Nuevo	18·28n	66·51w
Puerto Real	18·05n	67·11w
Punta Santiago	18·10n	65·45w
Quebrada Seca	18 14n	65·40w
Quebradillas	18 29n	66·56w
Rarney	18·28n	67·10w
Rincón	18·20n	67·15w
Río Grande	18·23n	65·50w
Rio Jueyes	18·01n	66·20w
Río Piedras	18·24n	66·03w
Sabana	18 20n	65·44w
Sabana Grande	18 05n	66·58w
Sabana Llana	18 02n	66·15w
Saint Just	18·23n	66·00w
Salinas	17 59n	66·18w
San Antonio	18 30n	67·07w
San German	18 05n	67·03w
San Juan	18·28n	66·07w
San Lorenzo	18 11n	65·58w
San Sebastián	18·20n	66·59w
Santa Isabel .	17·58n	66·24w
Santa Maria . .	18·09n	65·26w
Soroco	18·22n	65·38w
Tablones	18·15n	65·45w
Toa Alta	18 23n	66·15w
Toa Baja	18·27n	66·15w
Trujillo Alto	18·22n	66·01w
Utuado	18·16n	66·42w
Vega Alta	18·25n	66·20w
Vega Baja	18·27n	66·23w
Vertedero	18·05n	66·15w
Villalba	18 08n	66·30w
Villa Perez	18.121	66·47w
Yabucoa	18·03n	65·53w
Yauco	18 02n	66·51w

Physical features and points of **interest**

Agujereada Punta *point*	'8 31n	67·08w
Arenas Punta *point*	18·07n	65·35w
Arroyo, Puerto bay	17·58n	66·05w
At antic Ocean	18 55n	66 15w
Boquerón Bahía de bay	18 02n	67 12w
Bcr nquen Punta *point*	18·29n	67·10w
Brea, Punta, *point*	17·56n	66·56w
Cabullón, Punta *ooint*	17·57n	66·35w
Cadena Puita *ooint*	18 18n	67 15w
Caja de Muertos Isla *isiana*	17·54n	66·31w
Caon l as Lago laze	▮8 16n	66 39w
Caparra Las Ruinas de *ru ns*	18·24n	66·07w
Caribbean National Forest	18 18n	65·47w
Caribbean Sea	17 45n	66·20w
Carite Lago *lake*	18·06n	66·06w
Cayey Sierra de *mountains*	18·05n	65·52w
Central, Cordillera *mountains*	18·08n	66·35w
Cidra Lago de *reservoir*	18·12'1	66·08w
Coamo Lago *reservo r*	18011	66·23w
Culebra s a de *island*	18'91	65·18w
Culebrita Isla *island*	18·19n	65·14w
Dos Bocas Lago *reservo* r	18·19n	66·40w
El Morro *fort*	38 28n	66·07w
El Toro *mountain*	18 ·7n	65·50w
E Yunoue *mountain*	18·19n	65·50w
Este Purta *point*	18·08n	65·16w
Fort Allen army *installation*	18·01n	66·30w
Fosforescénte Bahía, bay	17·58n	67·01w
Grande de Añasco *river*	18·15n	67·12w
Grande de Arecibo *river*	18 29n	66 42w
Graide de Lo za *river*	18 27n	65 53w
Grande de Manati r *ver*	18 29n	66 32w
Guajataca *r* vtr	18 29n	66 57w
Guajataca Lago de *reservoir*	18 221	66 54w
Guana ibo *river*	18 '11	67 1lw
Guana ibo Punta *po nt*	18·10n	67·11w
Guánica Laguna de lake	18 00n	66 56w
Guayanés Punta p o nt	18·04n	65·48w
Guayaníl a Bah a de bav	18·00n	66·46w
Guaybal Lago *reservoir*	18 06n	6630%
Guilarte Monte *mountain*	18 091	6646%
H guero Punta *point*	18·22n	67·16w
Jobos Bahía de bay	17·59n	66·14w
La Cadena San Francisco, *mountain*..........	18·19n	67·10w
La Santa, Cerro *mountain*	18·07n	66·04w
Las Tunas Punta, *point*	18·30n	66·38w
La Torrecilla, *mountain*	18·12n	66·20w
Lima Punta point	18·10Q	65·42w
Lo za Ernbalse de *reservoir*	18 17n	66 00w
Luis Peia Cayo de *island*	18·18n	65·20w
Luquillo Sierra de mountains	18·17n	65·47w
Mala Pascua Cabo, cape	17·59n	65·55w
Mayaguez, Aeropuerto airport	18 15n	67·09w
Mayagüez Bahía de bav	18·12n	67·09w
Mercedita Aeropuerto *airport*	18·01n	66·34w
Mulas Pun a *point*	16 09n	65 27w
Norte Cayo island	18·20n	65·15w
Palom nos is a *island*	18·21n	65 34w
Petrona Punta *point*	37 58n	66 23%
Picúa Punta pornt	18·25n	65·46w
Piñeros Isla *island*	18·15n	65·35w
Pirata Monte h *ll*	18·06n	65 33w
Plata Río oe a *river*	18·29n	66·15w
Prieto *river*	18·15n	66·56w
Puerca Punta *point*	18·13n	65·36w
Puerto Nuevo Punta *point*	18·30n	66 24w
Puerto Rico International Airport of . .	18·27n	66·00w
Punta, Cerro de *mountain*	18.10n	66·35w
Ramey Air Force Ease	18·30n	67·08w
Rincón, Bahía de *bay*	17·59n	66·22w
Rojo Caoo cape	17·56n	67·12w
Rooseve ᵗ Roads Naval Station	18 15n	65 38w
San Jose Laguna *lagoon*	18·26n	66·01w
San Juan Bahía de bav	18·26n	66 07w
San Juan Cabeza de headland	18 23n	65 37w
San Juan Pasaje de *pasage*	18·23n	65·37w
San Juan Naval Station	18 28n	66·06w
Tabla Cerro de a *mounta n*	18 03n	66·08w
Tortuguero Laguna lagoon	'8 281	66 26w
Tortuguero Puerto del bay	'8 29n	66·28w
Uroyan Montanas de *mountains*	18 33n	67 00w
Vacía Talega Punta DO *nt*	.18·26n	65·53w
Vieques. Isla de *island*.............	.18·08n	65·25w
Vieques, Pasaje de cnannel	.18·10n	65·37w
Vieques Sonda de sound	.18·16n	65·23w
Virgin Passage	.18·20n	65·10w
Yauco rider	.17·57n	66·48w
Yauco Embalse de *reservoir*	.18·07n	66·50w



Size of symbol indicates relative size of town • ○ ◉ ▣ ▣

Elevations in metres

PUERTO RICO

executive, legislative, and judicial branches are spelled out in the constitution of 1952, which may be altered by the commonwealth so long as the articles are not in conflict with the U.S. Constitution or the legal stipulations of Puerto Rican–U.S. relations.

Executive, legislature, and judiciary

The governor, who is elected by direct popular vote to a four-year term, heads the executive branch. The legislature comprises the Senate and the House of Representatives, whose members are elected for four years. There are eight senatorial and 40 representative districts, and in addition, 11 senators and 11 representatives are elected at large. A complicated formula is used to assure proportional representation of minority parties.

Puerto Rico has a unified court system, which is administered by the island's Supreme Court, whose justices are appointed by the governor with the advice and consent of the commonwealth Senate. Civil law has been influenced largely by Spanish traditions and the code system of French law setting forth basic legal principles, though common law has influenced administrative law and many other areas of the legal system. A federal court has jurisdiction over the application of federal laws in Puerto Rico, and appeals may be carried to the Supreme Court in Washington, D.C.

Relations with the United States. Commonwealth voters elect a resident commissioner who has a voice, but no vote, in the U.S. House of Representatives. On the other hand, Puerto Ricans do not pay federal taxes on income received from island sources. Customs taxes on foreign goods imported into Puerto Rico and excise taxes on goods sold in the United States are collected by the federal treasury and returned to the commonwealth. Relations between Puerto Rico and the United States are defined in the Puerto Rico-Federal Relations Act, which retains many provisions of the Foraker (1900) and the Jones (1917) acts. Such matters as currency, defense, external relations, communications, and interstate commerce are within the province of the federal government. Local government—excluding San Juan, which has a city-management rule—is run by a mayor and council, both elected by popular vote.

Diversity of political parties

Politics. Elections are held every four years, supervised by an electoral board comprising representatives from majority and minority parties. There are four principal registered parties: the New Progressive, the Popular Democratic, the Puerto Rican Independence Party, and the Puerto Rican Socialist Party. The two leading parties are the pro-statehood New Progressive Party, which narrowly won the 1980 gubernatorial election with slightly more than 47 percent of the vote, and the Popular Democratic Party, which won 47 percent and supports the continuation of commonwealth status. Although the Popular Democratic Party was the strongest party from 1940 to 1968, the New Progressive Party won three of the four elections between 1968 and 1980. The Puerto Rican Independence Party, which polled third in the 1980 election, and the Puerto Rican Socialist Party advocate independence.

Education. Puerto Rico is deeply committed to the expansion of public education. About one-third of the overall governmental budget is allocated to education. Illiteracy has been reduced from more than 30 percent in 1940 to less than 10 percent by the late 1970s. Vocational and technical education has been stressed to combat the high rate of unemployment among the young.

The system of higher education is diversified. The University of Puerto Rico comprises a state university system, with three main campuses and a number of regional colleges. There are also two private universities, the Inter-American university of Puerto Rico and the Catholic University of Puerto Rico, as well as the Ana G. Mendez Educational Foundation, comprising two junior colleges, and other private colleges. Total higher education enrollment is about 125,000. Scholarships and aid have become important sources of income for universities and colleges, and private institutions especially are dependent on federal assistance to students. Although island universities generally are patterned after U.S. institutions, their atmosphere is distinctly Latin American. There are strong Spanish and Latin-American influences, especially in the humanities, whereas U.S. influence is strongest in the natural sciences and public administration.

Health. About 25 percent of the commonwealth budget is spent on health and welfare services. Health programs include malaria control and the building of pure water supplies and modern sewage disposal systems in cities and towns throughout the island. A network of urban clinics and rural health centres has been created to provide treatment for communicable diseases, basic medical care, and instruction in hygiene, nutrition, and prenatal and child care. The number of physicians rose from one per 1,000 population in 1969 to one per 640 by 1978. The rise in admissions to health centres indicates more widespread treatment.

Crises in housing and employment

Housing. Puerto Rico has made impressive strides in meeting the housing shortage, although the pressures continue to build. Home construction accounts for about 5 percent of the gross national product. The Urban Renewal and Housing Corporation is in charge of broad and diversified housing programs, with concentration on low-income projects. The corporation has taken advantage of assistance from several housing programs, but credit restrictions have affected both public and private housing. In spite of new construction, the proportion of housing that lacks essential services to the total supply of housing has remained more or less constant. As in other areas of the island's life, population growth seems to be the major contributing factor.

Employment and welfare. Despite industrialization, unemployment has risen dramatically; the number of jobless persons actively seeking work increased from about 10 percent of the labour force in the early 1970s to about 20 percent by the late 1970s. The problem is most acute among the young, one-fifth of whom are not in school, and among the rural workers, who constitute the bulk of the unskilled labour and work only four to six months of the year. This doubling of unemployment stems primarily from population pressures. Federal and commonwealth minimum wages cover every important industry and agricultural enterprise.

General price levels in Puerto Rico closely follow the levels on the U.S. mainland, and inflationary pressures severely affect the islanders. Though increased incomes have promoted social and economic–mobility, the growth has helped the urban areas at the expense of the rural. Most of the extreme poverty is concentrated there or in the city slums. Income security programs have been extended to islanders by mutual consent of the Congress and the insular legislature.

CULTURAL LIFE AND INSTITUTIONS

Cultural traditions

Puerto Rico's culture has strong roots in the Hispanic world. The language, the literature, the arts, and the surviving folklore link Puerto Rico with Latin America. The strong influence of the United States since 1898 has not deeply changed Puerto Rico's cultural expression. Though popular culture, strongly abetted by modern commercialism, may reflect some North American traits, the island's traditions are strong and have become the subject of concern and care by such institutions as the Puerto Rican Institute of Culture and the Ateneo Puertorriqueiio. A new generation of poets, novelists, short-story writers, and essayists keeps alive the traditions of such 19th-century forerunners as the novelists Alejandro Tapia and Manuel Zeno Gandia, the essayists and sociologists Eugenio Maria de Hostos and Salvador Brau, and the poets Jose Gautier Benitez and Lola Rodriguez de Tio. New playwrights and artists have also received considerable encouragement from the Institute of Culture, while there has been renewed interest in historical studies.

An annual drama festival has helped to promote the theatre. The Festival Casals, established in 1956 by Pablo Casals, became a major musical event that brought world-renowned musicians to the island each summer. The Institute of Culture sponsors many groups preserving popular folklore and conducting exhibits of local art.

As in many Caribbean countries, Puerto Ricans are searching for a definition of the island's cultural identity.

Some residents advocate cultural nationalism and a strong rejection of U.S. values, while others have a more eclectic position. There is, however, a general consensus that the island's culture is distinct from that of the United States, and assimilation that would obliterate Puerto Rico's Hispanic-American profile is neither possible nor desirable. Though these feelings are very strong in intellectual circles, the press, radio, and television are influenced heavily by U.S. modes and trends. Most Puerto Rican newspapers are published in Spanish and are served by U.S. press agencies.

PROSPECTS

Although Puerto Rico has had a dramatic economic boom since World War II, its close ties with the United States deeply affect its industrial and trading possibilities. Recession and inflation on the mainland create backlashes in Puerto Rico, and during the 1970s economic growth declined. A reduction in tourism and in manufacturing, amid continued population growth, could further depress the economy. This possibility is a subject of continuing concern to both government and private enterprise.

Sentiments on political status

Politically, the island remains divided among advocates of statehood, of outright independence, and of a fuller development of self-government under the commonwealth status. Statehooders argue that incorporation into the U.S. political system represents political equality and full participation in national life. Pro-independence groups insist that separation is the only way to attain cultural identity and to liberate Puerto Rico from U.S. economic and political domination. Commonwealth supporters hold that imposition of the federal tax system under statehood would adversely affect the island's economic development. They maintain further that the autonomous relation preserves the island's cultural heritage and yet keeps the advantageous political and economic ties with the United States. Although tensions and conflicts occur, there remains a strong commitment to voting and the democratic processes of government that is unmatched elsewhere in the Spanish-speaking Caribbean.

BIBLIOGRAPHY. For the history and cultural development of Puerto Rico, see ARTURO MORALES CARRION, *Puerto Rico and the Non-Hispanic Caribbean,* 2nd ed. (1971); and EUGENIO FERNANDEZ MENDEZ, *Historia cultural de Puerto Rico, 1493–*1968, 3rd ed. (1971); and for a cultural reference, see R. DEL ROSARIO, E. MELON DE DIAZ, and M. MASDEU, *Breve Enciclopedia de la Culturo Puertorriqueña* (1976). For the island's social development, consult JULIAN H. STEWARD (ed.), *The People of Puerto Rico* (1956). On political history, see REXFORD G. TUGWELL, *The Stricken Land* (1946, reprinted 1968); and THOMAS MATHEWS, *Puerto Rican Politics and the New Deal* (1960). Sympathetic accounts concerning the development of the commonwealth status are EARL PARKER HANSON, *Puerto Rico: Ally for Progress* (1962); and HENRY WELLS, *The Modernization of Puerto Rico* (1969). *A* more critical view is found in GORDON K. LEWIS, *Puerto Rico: Freedom and Power in the Caribbean* (1963). A radical view for independence is expressed in MANUEL MALDONADO-DENIS, *Puerto Rico: una interpretación histórico-social,* 5th ed. (1973; Eng. trans., 1972). The best overall studies are found in the PUERTO RICO COMMISSION ON THE STATUS OF PUERTO RICO, *Status of Puerto Rico: Selected Background Studies Prepared for the United States* (1966). For the geography and economics, see RAFAEL PICO, *The Geography of Puerto Rico* (1974); H.C. BARTON, *Puerto Rico's Industrial Development Program,* 1942–1960 (1959); and HARVEY S PERLOFF, *Puerto Rico's Economic Future* (1950, reprinted 1975). See also, JUANA GARCIA, *Panoramic History of Agriculture in Puerto Rico* (1979); and UNITED STATES DEPARTMENT OF COMMERCE, *Economic Study of Puerto Rico,* 2 vol. (1979). For a good description of the governmental structure, see CARMEN RAMOS DE SANTIAGO, *El gobierno de Puerto Rico* (annual). On the slums and the migration of Puerto Ricans to New York, see OSCAR LEWIS, *La Vida: A Puerto Rican Family in the Culture of Poverty* (1966); and CLARENCE SENIOR, *Our Citizens from the Caribbean* (1965).

(A.M.-C.)

Pufendorf, Samuel von

German publicist and jurist, Samuel von Pufendorf greatly influenced the philosophy of natural law in Germany and the rest of Europe from the latter part of the 17th century up to the French Revolution. His nontheological treatment of natural law still echoes in the opening lines of the U.S. Declaration of Independence.

Early life and works. Pufendorf was born in Dorfchemnitz near Thalheim, Saxony, on January 8, 1632, the son of a Lutheran pastor. Though the family was poor, financial help from a rich nobleman enabled his father to send both Samuel and his older brother Esaias to a well-known school in Grimma. There he acquired a sound classical education. He became a student of theology at Leipzig University, then a stronghold of Lutheran orthodoxy, but soon turned his attention to jurisprudence, philology, philosophy, and history. In 1656 he went to Jena, where he was introduced to the method of Descartes and also read the works of Hugo Grotius and Thomas Hobbes.

By courtesy of the Svenska Portrattarkivet Stockholm

Pufendorf, oil painting by Carl Peter Morth, 1735, after David Klocker Ehrenstrahl (1629–98). In the Nationalmuseum. Stockholm.

First work on natural law

In 1658 Pufendorf took employment as a tutor in the household of the Swedish ambassador in Copenhagen. When war broke out between Sweden and Denmark, he was imprisoned along with the rest of the ambassador's retinue. During the eight months of confinement he occupied himself by elaborating his first work on natural law, *Elementorum Jurisprudentiae Universalis Libri Duo,* published in 1660. In it he developed the ideas of Grotius and Hobbes. The elector palatine Karl Ludwig, to whom the work was dedicated, created a chair of natural law for Pufendorf in the arts faculty at Heidelberg, the first of its kind in Germany. Pufendorf taught there from 1661 to 1668, writing during this time his next work, *De Statu Imperii Germanici ad Laelium Fratrem Dominum Trezolani Liber Unus.* It was published in 1667 and took the form of a bitter attack, supposedly by a Veronese nobleman, on the constitution of the Holy Roman Empire and the house of Austria. It was based on wide reading in constitutional law and history and created an immediate sensation throughout Europe. The work was banned by the imperial censor; perhaps for that very reason it was translated into many languages and published abroad.

Career in Sweden. In 1668 Pufendorf left Heidelberg to accept the chair of natural law at the new University of Lund in Sweden. The 20 years he was to spend there proved to be his most fruitful ones. In 1672 he published his great work on the law of nature and of nations, *De Jure Naturae et Gentium Libri Octo*, and in 1673 an excerpt from it titled *De Officio Hominis et Civis Juxta Legem Naturalem Libri Duo.* In these works Pufendorf departed from the traditional approach of the medieval theologians to natural law and based it on man's existence as a social being *(socialitas).* Every individual, he held, on the basis of human dignity has a right to equality and freedom. He insisted that despite the teaching of Aristotle, there is no such creature as a natural slave; master–servant relationships can exist only on the basis of an agreement. Pufendorf's theory of civil, penal, and constitutional law also derived from the same principle of *socialitas.*

His great influence was not won without a struggle. His views were subjected to numerous attacks by conservative Protestant theologians in Sweden and Germany. The philosopher Leibniz dismissed him as "a man not a lawyer and scarcely a philosopher at all." The Swedish government, however, protected him. In the pamphlets signed *Eris Scandica,* which he published in 1686, he defended his beliefs very effectively.

Royal historiographer of Sweden

In 1667, after the Danish occupation of Lund, Pufendorf became the royal historiographer in Stockholm, where he devoted much of his time to writing the history of Sweden from Gustavus II Adolphus to Charles X Gustavus. In 1687 he also published *De Habitu Religionis Christianae ad Vitam Civilem,* setting forth the civil superiority of the state over the church but at the same time defending the church's power in ecclesiastical matters as well as the freedom of conscience of the individual. His approach became the basis of the collegial, or council, system of church government that was further developed in the 18th century to become the basis of church and state relations in Germany. The book also contained a justification of the idea of tolerance in general and in particular of the elector of Brandenburg, who had offered asylum to the Huguenots when they were driven out of France in 1685.

In 1688 Pufendorf, a baron, went to Berlin as historiographer to the elector of Brandenburg. He died there in October 1694. A posthumous work, *Jus Feciale Sive Dc Consensu et Dissensu Protestantium,* was published in 1695. It expounded more of his ideas on ecclesiastical law and argued for the formation of a united Protestant church from the Reformed and Lutheran churches.

The 18th century saw many editions of Pufendorf's works. John Locke and Jean Jacques Rousseau recommended them as reading for young people. The emperor Joseph II of Austria was instructed in *De officio* as a boy. One of Pufendorf's disciples was John Wise (1652–1725), an American clergyman and pamphleteer who greatly influenced American ecclesiastical law and the struggle for civil and religious liberty in the colonies.

BIBLIOGRAPHY. The best biography of Pufendorf is the article by HARRY BRESSLAU in the *Allgemeine Deutsche Biographie,* vol. 26 (1888); see also the brief biography by PAUL MEYER, *Samuel Pufendorf* (1894). G.H. VON TREITSCHKE, *Historische und politische Aufsätze,* vol. 4 (1897); J.G. DROYSEN, *Abhandlungen zur neueren Geschichte* (1876): and F. MEINECKE, *Die Idee der Staatsräson in der neueren Geschiclzte,* 3rd ed. (1929), all treat his achievements as historian. For a discussion of Pufendorf's teaching of natural history, see HANS WELZEL, *Die Naturrechtslehre Samuel Pufendorfs (1958).*

(Ha.We.)

Pulsar

The discovery of pulsars in 1967 ranks as an important milestone in the history of astrophysics. The name, an abbreviation of "pulsating radio star," derives from the fact that pulsars emit all their recorded radiation in the form of short pulses. These pulses typically last a few hundredths of a second and are usually emitted at intervals of somewhat less than one second. With one exception, the detectable radiation from pulsars occurs entirely in the radio wavelength region of the electromagnetic spectrum, and observation of pulsars is possible only by the use of radio telescopes. The pulse rate varies from source to source, but individual pulsars maintain their own periodicity with astonishing accuracy. If the pulses were used for timekeeping, such a clock would be correct to within a fraction of a second per year.

The first pulsar was discovered by chance by A. Hewish and S.J. Bell during a sky survey carried out at Cambridge, England, using an unusual type of radio telescope. This instrument consisted of a large array of 2,048 dipoles, which covered an area of 20,000 square metres (4½ acres). It was designed to distinguish ordinary radio galaxies from quasars using the phenomenon of interplanetary scintillation, a rapid intensity variation caused by ionized gas clouds in space that occurs only for sources that subtend a very small angle, such as the compact quasars; it thus provides a powerful method of discriminating between these and other types of radio source.

In the Cambridge survey a list was made of every source that scintillated. One rather weak source, which apparently showed scintillation, was unusual in that it did not always appear on repeated scans of the same region of sky. At first it was dismissed as a man-made radio interference, but a recorder of higher sensitivity revealed the first indication of pulsed emission on Nov. 28, 1967, and this result was confirmed in early December. A detailed investigation was carried out immediately, the results of which were published on Feb. 24, 1968. In this account the basic physical nature of the emitter was outlined, and it was suggested that only a collapsed star, such as a white dwarf, or the hypothetical neutron star could offer a plausible explanation.

News of the discovery initiated a period of intense activity among astronomers all over the world. Many questions about pulsars are still far from solved, but it is now generally agreed that pulsars must be neutron stars. The main observed features of pulsars and some theories regarding them will be outlined briefly below.

Approximate Positions and Periods of 55 Pulsars

key: CP—Cambridge, Eng.; MP, PSR—Molonglo, Austr.; JP—Jodrell Bank, Eng.; HP—Harvard, Mass.; NP—National Radio Astronomy Observatory, Green Bank, W.Va.; AP—Arecibo, P.R.; PP—Pushchino, U.S.S.R.

	right ascension			declination			period (in seconds)
	hours	minutes	seconds	degrees	minutes of arc	seconds of arc	
MP	00	31	37	−07	37	—	0.942951
MP	02	54	24	−54	—	—	0.448
CP	03	29	11.08	54	24	38.3	0.714518
MP	04	50	22	−18	—	—	0.5497
NP	05	25	45	21	58	—	3.745491
NP	05	31	31.46	21	58	54.8	0.033099
PSR	06	28	53	−28	33	—	1.244436
MP	07	36	51	−40	35	—	0.374918
CP	08	08	58	74	38	10	1.292241
MP	08	18	6	−15	—	—	1.237
AP	08	23	52	26	48	00.0	0.530659
PSR	08	33	39	−45	00	05	0.089209
CP	08	34	26.3	06	20	47.0	1.273763
MP	08	35	34	−40	—	—	0.765
HP	09	04	—	77	40	—	1.57905
MP	09	40	40	−56	—	—	0.662
PP	09	43	20	10	06	—	1.097707
CP	09	50	30.85	08	09	49.8	0.253065
MP	09	59	51	−54	37	—	1.436551
CP	11	33	27.39	16	07	30.4	1.187911
MP	11	54	45	−62	—	—	0.400
AP	12	37	17	25	09	30	1.382451
MP	12	40	21	−63	36	—	0.388
MP	13	59	43	−50	—	—	0.690
MP	14	26	35	−66	30	—	0.788
MP	14	49	22	−65	—	—	0.180
PSR	14	51	29	−68	32	—	0.263376
HP	15	08	03.27	55	42	50	0.739677
MP	15	30	23	−53	30	—	1.368852
AP	15	41	10	09	38	—	0.74845
MP	16	04	37	−03	—	—	0.421
MP	16	42	15	−03	12	36	0.387688
MP	17	06	35	−16	37	—	0.653050
MP	17	27	50	−47	40	—	0.829683
MP	17	47	56	−46	56	—	0.742349
PSR	17	49	49	−28	05	57	0.562553
MP	18	18	14	−04	25	—	0.598072
JP	18	45	10	−04	00	—	0.597731
MP	18	57	44	−25	—	—	0.6118
JP	18	58	40	03	24	—	0.655444
MP	19	11	15	−04	47	—	0.825933
CP	19	19	36.1	21	47	12.9	1.337301
PSR	19	29	52	10	53	04	0.226517
JP	19	33	31.9	16	09	58.8	0.358735
MP	19	44	38	17	—	—	0.440
JP	19	46	10	35	25	—	0.717306
JP	19	53	00	29	12	—	0.426676
JP	20	03	00	31	30	—	2.111206
AP	20	16	00.07	28	30	31	0.557953
JP	20	21	50	51	44	—	0.529195
PSR	20	45	47.6	−16	27	50	1.961566
JP	21	11	45	46	36	—	1.014686
HP	22	18	20	47	40	—	0.538467
AP	23	03	30	30	45	—	1.575869
JP	23	19	15	60	00	—	2.256483

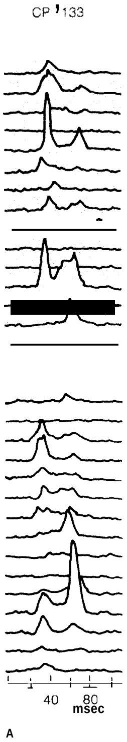
OBSERVED PROPERTIES

Periods. Searches for pulsars at several observatories led to the discovery of 49 sources by the end of 1969. The approximate positions and periods are listed in the Table. All of these sources exhibit similar behaviour, and the periods (intervals between successive pulses) range from 0.033 second to 3.7 seconds. The most rapid pulsar, NP 0532, lies in the Crab Nebula and was discovered at Green Bank, West Virginia. The most common period is close to one second, and approximately one half of the sources have periods between 0.5 and 1.0 second. The detection of pulsars becomes more difficult at the shortest periods, and it is possible that faster pulsars exist than can be detected at present, but there is no observational bias hindering the detection of slower pulsars, and if such sources exist, they must be exceedingly weak emitters.

Accurate timing of pulses indicates a systematic tendency for the period to increase with time. A useful measure of this trend is the number of years that must elapse before the period has doubled, assuming that the rate of slowing remains unchanged. For a typical pulsar about 10,000,000 years would be required, although the most rapid pulsar gives a measure of 2,400 years and one pulsar is so constant that over 100,000,000 years would be needed.

Timing measurements on those pulsars that emit the sharpest and strongest pulses enable the period to be measured with an error of less than one part in 1,000,000,000. Such studies have revealed occasional small, but sudden, changes of period in two cases. These results are of great interest and will be mentioned again (see below ***Theories***).

Pulse shapes. A typical record of pulses from the source CP 0950 is shown in Figure 1(B). Large variations of pulse intensity are clearly evident, and the pulse is sometimes undetectable. When records are made with a sufficiently sharp resolution in time it is found that the detailed shapes of the pulses also vary rapidly. Fine structure within individual pulses sometimes shows spikes that last for only a fraction of a millisecond. A series of consecutive pulses, shown in Figure 1(A), indicates the typically complicated shapes that are observed.



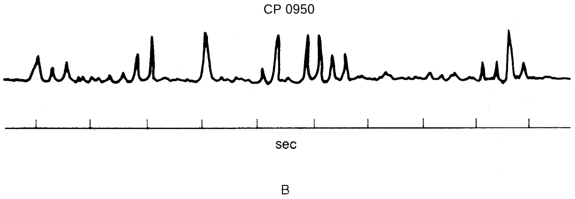


Figure 1: Pulse shapes.
(A) Consecutive pulses from CP 1133 indicating a rapidly changing fine structure within the pulses. (B) Train of pulses received from the pulsar CP 0950 at a wavelength of 3.j metres.

When many pulses from an individual source are superimposed and averaged, the mean pulse takes on a characteristic shape, the mean pulse envelope, which differs between one source and another and which is relatively unchanging for a given source. Some examples of characteristic mean pulse envelopes are given in Figure 2. For three of the known pulsars an interpulse occurs almost symmetrically placed in time between the main pulses. The interpulse is always weaker than the main pulse and has a different shape for the Crab Nebula pulsar, in which it is strong enough to be studied in some detail.

Pulsars have been observed at radio wavelengths ranging from eight metres to six centimetres, although the radiation is usually most intense near a wavelength of one to three metres. In one case (the pulsar in the Crab Nebula) both visible light and X-rays have been detected. It is important, for theoretical reasons, to see how the pulse shapes vary with wavelength, and the mean envelopes for the Crab Nebula pulsar are shown in Figure 2. Some differences are clearly evident, but the pulses at all wavelengths are emitted simultaneously from the source. In other cases a slight increase in width of the mean envelope is observed at the longest radio wavelengths, but the effect is usually small.

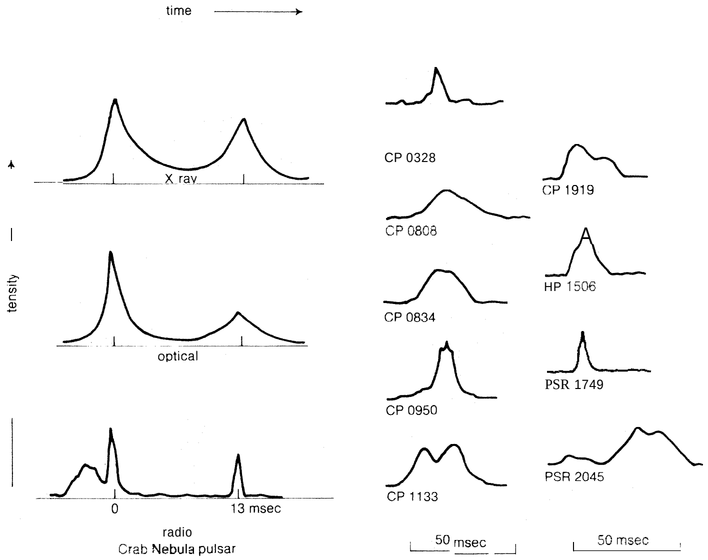


Figure 2: Examples of mean pulse envelopes (see text).

Intensity variations. One of the most remarkable features of pulsars is that their regular timekeeping is associated with extreme variability of emitted intensity. Not only do the pulses vary on a rapid time scale as shown in Figure 1; they also exhibit increases of intensity lasting for minutes, hours, days, and weeks. For the most part these effects seem to be entirely irregular, but careful studies of pulse trains have shown that certain pulsars exhibit slight intensity variations, which repeat at intervals corresponding to several pulse periods. This class II pulsation, as it has been called, is not commensurate with the main period and is related to an effect known as "marching sub-pulses" in which fine structure reappears in successive pulses and drifts steadily through the mean pulse envelope.

Intensity variations that last from several minutes to a few hours are probably caused by tenuous clouds of ionized gas in interstellar space. Such clouds will deviate the paths of radio waves as does the Earth's ionosphere, so that the radiation is received at slightly different angles. The mixing, or interference, of these waves gives rise to intensity variations analogous to the twinkling of visible stars. It is not possible, however, to explain such variations that persist longer than a few hours. To summarize, it appears that intensity fluctuations, which last for less than a few seconds or for longer than a day, must be genuine effects intrinsic to the pulsar emitter. Variations of an intermediate duration, on the other hand, can be ascribed to interstellar scintillation.

THE PHYSICAL NATURE OF PULSARS

Distribution of pulsars **in the Galaxy.** A map of the entire sky showing the positions of pulsars is given in Figure 3. This chart is plotted in galactic coordinates so that latitude 0° corresponds to the plane defined by the disk of the Galaxy; part of this disk is visible as the Milky Way.

The pulsars appear to be more or less randomly scattered over the sky, but there is a significant concentration

From Revista del Nuovo *Cimento* (1969)

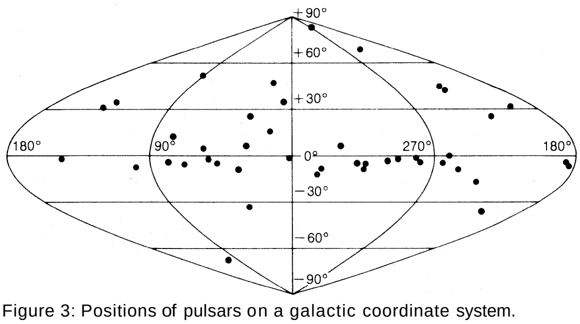


Figure 3: Positions of pulsars on a galactic coordinate system.

of sources toward the galactic plane. Pulsars are thus distributed similarly to the younger stars of the Galaxy and cannot be too distant. The Galaxy has a total extent of about 100,000 light-years, and the disk has a thickness of about 1,000 light-years. The known pulsars must be situated within a few thousand light-years of the solar system, or they would be expected to show a greater concentration within the Milky Way than is depicted in Figure 3. Thus, only the nearest pulsars have so far been found, and it is likely that some 10,000 pulsars are present in the entire Galaxy.

Pulsar concentrations

Distance measurement. Two methods have been used to find the distances of pulsars with greater accuracy than the statistical estimate discussed above. One of these relies upon the fact that the speed at which radio waves travel through an ionized gas depends upon the wavelength. It is known that interstellar space contains a trace of ionized hydrogen, and this means that receivers tuned to different wavelengths will record incoming pulses at slightly different instants, even though the pulses were emitted simultaneously at the source. In practice, the measurements are quite simple because the distance involved is so great that time delays of several seconds are usually observed. Unfortunately, the density of gas in interstellar space is not accurately known, and this uncertainty is the limiting factor.

Distance measurements using this technique, known as the method of pulse dispersion, usually assume a density of about 0.03 ions per cubic centimetre. On this basis, the nearest pulsar is at a distance of about 300 light-years, while the average distance is approximately ten times this value. It is encouraging that this method gives a result for the Crab Nebula pulsar that tallies well with the distance of the nebula determined by optical measurements.

The second way of estimating pulsar distances makes use of an absorption technique. Cold hydrogen gas in the form of atoms, as opposed to electrically charged ions, acts as a strong absorber of radiation having a wavelength of 21 centimetres. This fact has already been widely used in radio astronomy to study the spiral arms of the Galaxy (see RADIO SOURCES, ASTRONOMICAL). For a small number of intense pulsars it is possible to detect 21-centimetre absorption and hence to determine how many spiral arms lie between Earth and the source. This method also yields an estimated distance of about 3,000 light-years.

Size and energy. It has been obvious from the first detection of pulses that pulsars must be exceedingly small astronomical bodies. A large body cannot emit a sharp flash of radiation because of the different travel-time of radiation from different parts of its surface. Broadly speaking, the duration of a pulse multiplied by the velocity of light defines the greatest possible radius of the body. In the case of pulsars, the observed narrow pulses show that the emitter cannot have a radius exceeding a few thousand kilometres; that is, pulsars cannot be larger than small planets such as the Earth; and they could, of course, be smaller.

An estimate of the total power radiated by pulsars can be made because their distances are known with tolerable accuracy. The Crab Nebula pulsar turns out to be a more powerful radiator than the Sun. This is an exceptional case, and typical pulsars radiate less strongly. It is clear, however, that pulsars, although they are tiny objects, must be endowed with a prime source of energy that is comparable to that of a stellar body. The confinement of such energy within so small a volume is another remarkable feature of these sources.

Association with supernova **explosions.** Of great significance with regard to the physical nature of pulsars is the fact that two of them are found in the positions of old supernovas. The first evidence of this kind came from Australian astronomers who discovered the source PSR 0833 near the centre of a radio nebula called Vela X. This nebula emits no visible light, but its general characteristics are similar to those of other nebulae known to have resulted from stellar explosions that emit both light and radio waves.

The discovery of PSR 0833 was soon followed by that of NP 0532 at the centre of the Crab Nebula. This nebula is well known since the actual stellar explosion was witnessed and documented by the Chinese in AD 1054. The discoverers of NP 0532 were able to place only an upper limit on its pulse period, which was later determined by use of the large radio telescope at Arecibo. Finally, with the period established, optical astronomers were able to adopt pulse integration techniques, and a star at the centre of the nebula was found to be emitting light flashes synchronous with the radio pulses.

A photograph of the Crab Nebula, in which the flashing star is clearly visible, is shown in Figure 4. This star had been suspected for many years to be the remnant of the one that actually exploded, and its coincidence with the pulsar is most revealing as will be seen in the following section. The Table shows that PSR 0833 and NP 0532 possess the shortest known periods. Many unsuccessful attempts have been made to find rapid pulsars associated with nebulae resulting from other supernovas.

By courtesy of A. Hewish

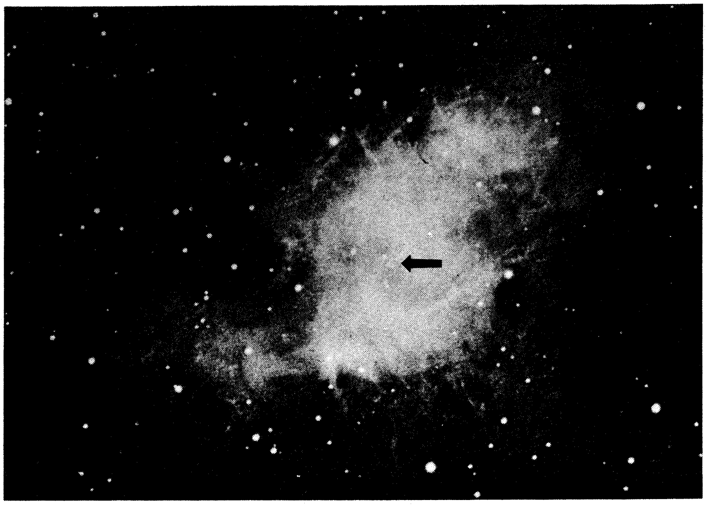


Figure 4: The Crab Nebula, which contains the only known sample of a visual pulsar.

THEORIES

No fully satisfactory theory of pulsars yet exists. Much speculation is still involved, and adequate solutions to many of the problems can be expected only when astrophysicists are more familiar with the extreme physical conditions encountered in and near these bodies. What follows is an outline of the concepts that currently have achieved general acceptance.

The **pulsar** clock. Any theory of the pulsars must concentrate, initially, on a mechanism for maintaining the basic period of the pulses. Three possibilities have been discussed, only one of which now appears to be reasonable. These possibilities are orbital motion, radial vibration, or rotation of a stellar body. In all three cases the periodicity is related very simply to the material density of the body or system concerned, and the observed periods require densities between 100 and 10,000 tons per cubic centimetre.

White dwarf stars are comparable in size to the Earth and have densities of about one ton per cubic centimetre. Unfortunately, they would disrupt and fly apart before producing periods as short as those observed. These stars are formed by the gravitational collapse of common stars, such as the Sun, when nuclear fusion processes within them come to a halt. White dwarfs cannot possibly exist at densities greater than about 100 tons per cubic centimetre, as gravity then crushes the matter to nuclear density in the region of 10,000,000 tons per cubic centimetre. Under these conditions the material is largely composed of neutrons; such stars have been postulated but never found.

Neutron stars—one solution

Neutron stars offer a possible solution to the pulsar problem, and rotation gives a simple explanation of the periods. Orbital motion, on the other hand, would give a period decreasing with time, which contradicts the observations (see above ***Periods),*** while radial vibration predicts periods that are far too short. The rotation theory, which was advanced soon after the initial discovery, gained weight from the observed slowing down of the periods and from the fact that two pulsars coincide with supernova debris. The latter point is significant since the gravitational collapse, which might lead to the formation of a neutron star, is a violent process and had been discussed for some years as an explanation of supernova explosions. Further, the principle of conservation of angular momentum requires the neutron star to be spinning rapidly.

Neutron star models. The understanding of some process for producing the pulsar clock is but a start to pulsar theory. The energy source that maintains the radiated pulses and the mechanism for producing the pulses themselves must also be determined.

Given a rotating body, the simplest method of producing pulses is to confine the radiation to a single narrow beam, the basic "lighthouse" model of pulsars. The observed pulse duration in relation to the period suggests that the beam typically has a width of about 15". Most of the explanations of such a beam—and why the radiation produced should be largely at radio wavelengths—relyon the intense magnetization that a neutron star should contain.

During the collapse of matter to the neutron star configuration, a great deal of gravitational energy is released, while the angular momentum, which depends on the product of the spin and the star's radius, remains constant. It has been estimated that immediately following collapse a neutron star will be spinning so rapidly (around 1,000 revolutions per second) that its energy of motion is comparable to that derived from nuclear fusion during the whole of its previous history as an ordinary radiating star. Collapse has therefore revitalized the star to a remarkable degree. In addition, the magnetization of the star will have been concentrated by a factor of about 10^{10}, since its radius has shrunk from about 1,000,000 kilometres to a mere 10 kilometres.

If the direction of the magnetic axis of the star does not coincide with the rotation axis, extremely large fluctuating magnetic fields will be produced in space. Even if this were not so, the voltages caused by dynamo action at the surface of the star will exert forces sufficient to tear charged particles from the star and fling them outward. These particles may either be swept around by magnetic forces or accelerated radially in the fluctuating fields until they are travelling at speeds close to that of light. When this happens the particles will generate radiation, and the theory of relativity shows that this will be cast forward into a narrow beam. Such processes should be sufficient to account for pulsar radiation, but detailed theories have not yet been formulated. Examples of the type of model currently under consideration are illustrated in Figure 5.

Outstanding problems. The association of pulsars with neutron stars and supernova explosions has focussed attention on the question of the final collapse of stars too heavy to reach equilibrium as white dwarfs. This process is not well understood; and, if pulsars had not been found, it is likely that recent theoretical work would have cast doubt on the possible creation of neutron stars according to earlier theories.

From *Annual Review of Astronomy and Astrophysics* (1970)

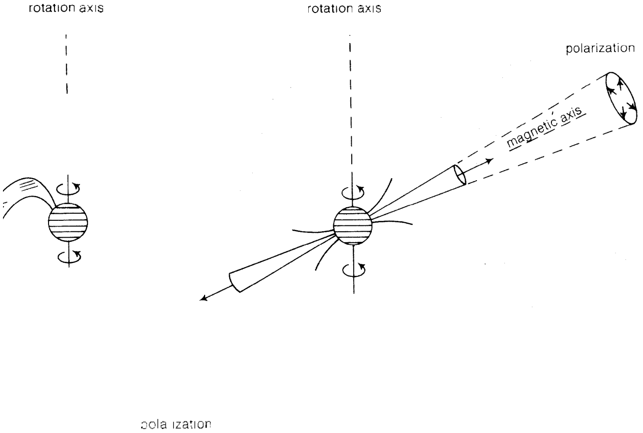


Figure 5: Two neutron star models that have been suggested to account for pulsed emission. (Left) Gold's model. (Right) Radhakrishnan's model.

It is also necessary to consider the sudden changes in period that have been observed in NP 0532 and PSR 0833. Superdense matter, such as that which must exist inside neutron stars. has peculiar properties in addition to its enormous density. It is possible that neutron stars are surrounded by a shell of more or less rigid material, within which a fluid core is contained. This fluid core may exist in a state similar to that of liquid helium at extremely low temperature, when the viscosity virtually disappears. The superfluidity of liquid helium is analogous to the superconductivity of certain metals in which, again at very low temperature, an electrical current once started will flow without ceasing for a great length of time.

Solid shell is possible

As the neutron star slows down, the shell undergoes a varying stress and will deform, at intervals, to achieve a less ellipsoidal shape. Such a momentary deformation will cause a sudden decrease of period; as is observed, but the fluid core will continue to spin at the original speed. Eventually viscous forces will cause the whole star to rotate at the same, slightly modified speed, and for PSR 0833 this process appears to take several years. Calculations show that the angular speeds of the core and the shell would equalize far more quickly if a superfluid state did not exist within the neutron star. A full understanding of the behaviour of neutron stars, therefore, involves a more accurate theory of this peculiar property of superdense matter than is currently available.

Still another problem is that of explaining the nonexistence of pulsars having periods longer than two or three seconds. It would be expected that the gradual loss of energy in the form of radiation would cause pulsars to spin more and more slowly. One suggestion is that the magnetization of the neutron star decays, causing the radiation to cease, before the spin rate has fallen much below one revolution per second. There are, however, divergent estimates of how long the magnetization will persist.

CONCLUSIONS

In the present state of knowledge it appears that pulsars can be explained only in terms of neutron stars. The discovery that such stars exist has opened a new chapter in astrophysics. The realization that gravitational collapse can cause a star to be reborn, at least from an energetic standpoint, may well have far-reaching consequences. The Crab Nebula, for example, has long been known to contain some source of energy by which it maintains the radiation that it emits at X-ray, light, and radio wavelengths. This radiation, from the whole of the nebula, contains considerably more power than that emitted from the pulsar NP 0532 itself. It now seems highly probable that the spinning neutron star provides this energy. It has been further postulated that spinning neutron stars may eject the fast-moving particles known as cosmic rays. The latter, which permeate the Galaxy, have been widely studied for many years, but their origin has remained a mystery.

The study of pulsars has also led to the resurrection of the question of the ultimate fate of stars too heavy to reach stability as neutron stars. According to present theory, matter at a density slightly greater than that inside neutron stars is unstable with respect to further gravitational collapse; that is, gravitational forces are so strong that the matter is crushed into an ever decreasing volume, and no known forces are available to resist the process.

In addition to providing what is, in a sense, a laboratory for the study of matter and physical laws under conditions that could not conceivably be attained on earth, pulsars may also be regarded as a new tool for astronomical purposes. They supply a form of clock that may prove to be sufficiently accurate for observational tests of general relativity. They are also being used to study the propagation of radio waves in interstellar space, which yields valuable information on such questions as gas density in interstellar space and the galactic magnetic field. There can be no question that pulsars have made, and will continue to make, a considerable impact on the development of astronomy.

BIBLIOGRAPHY. *Pulsating Stars,* 2 vol. (1968–70), contains a collection of research papers that originally appeared in *Nature,* a weekly journal. An extensive review of this subject by A. HEWISH may be found in the *A. Rev. Astron. Astrophys., 8*:265–296 (1970).

(A.He.)

Pump

A pump is a device that uses mechanical force and motion to raise, transport, or compress fluids. The earliest pumps were devices for raising water, such as the Persian and Roman waterwheels and, more technologically sophisticated, the famous Archimedean screw, invented by Archimedes of Syracuse (287–212 BC) to raise water from the hold of a ship. The screw turned in a tightly fitting cylinder whose inclination caused the water to be raised on the screw threads as the screw was rotated. Similar devices were used in dewatering operations in bridge building as late as the 18th century.

The mining operations of the Middle Ages led to development of the suction (piston) pump, many types of which are described by Agricola in *De re metallica* (1556). One such device consisted of three suction pumps driven by a single crank powered by a waterwheel. A suction pump works by atmospheric pressure; when the piston is raised, creating a partial vacuum, atmospheric pressure outside forces water into the cylinder, whence it is permitted to escape by an outlet valve. Atmospheric pressure alone can only force water to a maximum height of about 34 feet, so the force pump was developed to drain deeper mines. In the force pump the downward stroke of the piston forces water out through a side valve to a height that depends simply on the force applied to the piston. The search for bigger power sources for such mining pumps led to the invention of the steam engine.

GENERAL CONSIDERATIONS

Classification of pumps. Pumps are classified according to the way energy is imparted to the fluid. The basic methods are (1) volumetric displacement, (2) addition of kinetic energy, and (3) use of electromagnetic force.

The volumetric displacement of a fluid can be achieved either mechanically or by the use of another fluid. Kinetic energy may be added to a fluid either by rotating it at high speed or by providing an impulse in the direction of flow. In order to use electromagnetic force, the fluid being pumped must be a good electrical conductor.

Fluids include both liquids and gases. Pumps used to transport or pressurize gases are called compressors, blowers, or fans.

Pumps in which the volumetric displacement is accomplished mechanically are called positive displacement pumps. Volumetric displacement is also accomplished with other fluids, as in air lifts and blow cases.

Kinetic pumps

Kinetic pumps impart kinetic energy to the fluid by means of a rapidly rotating impeller.

Broadly speaking, positive displacement pumps provide relatively low volume at high pressure, while kinetic pumps produce relatively high volume at low pressure.

Uses. The pumping of fluids is an extremely common operation and the most basic operation in modern industry. The heart is a pump that causes blood to flow through the body. In the home, pumps are used to circulate water in hot-water heating systems and air in air-circulating heating and cooling systems. In automobiles, airplanes, and other vehicles, pumps circulate oil, water, and fuel.

Since fluids are so much more easily transported than solids, it is common industrial practice to convert solid materials into the fluid state by melting them, dissolving them in a liquid, or fluidizing them as a suspension of solid particles in a liquid or gas preparatory to movement. As early as 1850, gold-bearing gravels suspended in water were transported through pipelines in California. In 1913, coal suspended in water was transported through a pipeline in England. Typical of modern installations is a 72-mile pipeline that carries 700 tons of gilsonite per day from Utah to Colorado. Live fish, canned products, boxes, sewage, etc., suspended in water, are also pumped through pipelines.

Although very large pumps do exist, such as the ones used for irrigation service at the Hiwassee Dam (1,750,000 U.S. gallons per minute), the vast majority of pumps are relatively small, with capacities up to 2,000 gallons per minute and generated heads of between 5 and 300 feet. (One gallon is 3.785 litres.)

Efficiency of pumping. If a pump is installed at the wrong location in a pipeline, it cannot deliver its rated capacity. A certain amount of pressure is required to get the fluid to flow into the pump before additional pressure or velocity can be added. This head (energy per pound due to pressure, velocity, or elevation) is called the net positive suction head (NPSH) and is an integral part of the pump rating. A pump must be installed so that the head available at the intake is equal to or greater than the rated NPSH requirement. If the head available is less than the required NPSH, pumps will "cavitate." Cavitation is the formation of a vacuous space in the pump, which is normally occupied by liquid. Vaporization of liquid in the suction line is a common cause of cavitation. Vapour bubbles carried into the pump with the liquid collapse when they enter a region of higher pressure, resulting in excessive noise and vibration. Cavitation not only reduces the capacity of the pump but it leads to increased corrosion and erosion and a shorter life and may even destroy the pump.

Pump parameters

The important parameters of a pump are the required NPSH, the capacity against a given total head, and the overall percentage efficiency for pumping a particular fluid.

The dominant factor controlling the overall pumping efficiency is the viscosity of the fluid being pumped. Viscosity is that property of a fluid that resists any force tending to produce flow. Pumping efficiency is much higher for low viscosity liquids such as water than for a highly viscous fluid such as molasses. Since the viscosity of a liquid normally decreases as the temperature is increased, it is common industrial practice to heat very viscous liquids in order to pump them more efficiently.

Materials of construction. Pumps are manufactured in quantity in only a few standard materials such as cast iron, carbon steel, high-chrome steel, chrome-nickel-stainless steel, bronze, etc. In addition, pumps are fabricated from more specialized materials such as ceramics, glass, rubber, plastics, carbon, lead, aluminum and its alloys, titanium, zirconium, etc. Pumps are frequently fabricated with individual parts of different materials.

MAIN TYPES

Positive displacement. Positive displacement pumps (that is, pumps that lift a given volume for each cycle of operation) can be divided into two main classes, reciprocating and rotary. Reciprocating pumps include piston, plunger, and diaphragm types, while rotary pumps include gear, lobe, screw, vane, and cam pumps.

Piston and plunger pumps. The plunger pump is the oldest type in common use. Piston and plunger pumps consist of a cylinder in which a piston or plunger moves back and forth. In plunger pumps the plunger moves through a stationary packed seal and is pushed into the fluid, while in piston pumps the packed seal is carried on the piston which pushes the fluid out of the cylinder. The movement of the piston or plunger creates an alternating increase and decrease in volume. As the piston moves outward, the volume available in the cylinder increases, and fluid enters through the inlet one-way check valve. As the piston moves inward, the volume available in the cylinder decreases, the pressure of the fluid increases, and the fluid is forced out through the outlet one-way check valve. The pumping rate varies from zero at the point at which the piston changes direction to a maximum when the piston is approximately halfway through its stroke. The variation in pumping rate can be reduced by using both sides of the piston to pump fluid. Pumps of this type are called double acting. Fluctuations in pumping rate can be further reduced by using more than one cylinder.

Pumps with one, two, three, and four cylinders are called simplex, duplex, triplex, and quadruplex pumps respectively. An air chamber at the pump's discharge is sometimes used to smooth out the pulsations resulting from individual cylinder discharges. Overall pumping rates of piston pumps may be varied by changing either the reciprocating speed of the piston rod or the stroke length of the piston. The piston may either be driven directly by steam, compressed air, or hydraulic oil or through a mechanical linkage that transforms the rotary motion of a drive wheel to a reciprocating motion on the piston rod. The latter are called power-piston pumps and the former direct-acting piston pumps. Pumps that can be driven by air or steam are useful in remote locations where electricity may not be available.

Piston drives

Piston and plunger pumps are available with either vertical or horizontal cylinders. Although relatively expensive, they are extremely reliable and have an exceptionally long life. Piston pumps are known to have been running without repair or replacement for well over 100 years, Capacities range from as low as 0.15 gallon per minute for plunger pumps to 2,000 gallons per minute for steam-driven piston pumps. Most piston pumps are designed to generate discharge pressures up to 1,000 pounds force per square inch, and plunger pumps are available that are designed to generate discharge pressures up to 50,000 pounds force per square inch. If operated against a closed discharge, a power piston pump will continue to generate head until the pump bursts, unless provided with a relief valve.

Piston pumps can pump either liquids or gases or a mixture of both but are unsuitable for pumping fluids containing solid particles because of a tendency to clog.

Diaphragm pumps. The action of a diaphragm pump is similar to that of a piston pump in which the piston is replaced by a pulsating flexible diaphragm. This overcomes the disadvantage of having piston packings in contact with the fluid being pumped. As in the case of piston pumps, fluid enters and leaves the pump through one-way check valves. The diaphragm may be actuated mechanically by a piston directly attached to the diaphragm or by a fluid such as compressed air or oil. Oil may be supplied externally, or it may be pumped by a piston that forms an integral part of the pump.

Diaphragms are made from many materials, including rubber and elastomeric plastics. Diaphragm pumps can pump either liquids or gases or a mixture of both but share with piston pumps the disadvantage of a pulsating output. They are useful for pumping liquids that contain solid particles and for pumping expensive, toxic, or corrosive chemicals where leaks through packing cannot be tolerated.

Diaphragms normally last for six months to one year and can be speedily and easily replaced. Diaphragm pumps have the advantage that they can be run dry for an extended period of time. Furthermore, the overall pumping rate of most such pumps can be changed during operation. They normally range in capacity from 4 to 100 gallons per minute. Diaphragm metring pumps are available with capacities as low as one gallon per minute, designed to generate discharge pressures up to 3,500 pounds of force per square inch.

Pump capacities

***Gear** pumps.* The most common type of gear pump is the external gear pump illustrated in Figure 1. Two meshing gears of equal size are contained in a fixed casing. One of the gears is driven and the other (idler gear) normally runs free. Various types of gears are used. A partial vacuum, created by the unmeshing of the rotating gears, draws fluid into the pump. This fluid is then transferred to the other side of the pump between the rotating gear teeth and the fixed casing. As the rotating gears mesh together, they generate an increase in pressure that forces the fluid into the outlet line. If operated against a closed discharge, a gear pump will continue to generate head until the pump bursts. They are, therefore, normally provided with a relief valve. A basic external gear pump can discharge fluid in either direction, depending on the direction of the gear rotation. However, because of the position of the relief valve, the direction is fixed and usually clearly marked on the pump.

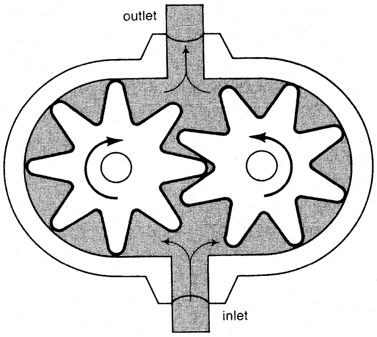


Figure 1: External gear pump.

An internal gear pump is shown in Figure 2. The driven gear is a rotor with internally cut gear teeth, which mesh with the teeth of an externally cut idler gear, set off-centre from the rotor. The crescent part of the fixed casing divides the fluid flow between the idler gear and the rotor. Gear pumps can pump liquids containing vapours or gases. Since they depend on the liquid pumped to lubricate the internal moving parts, they are not suitable for pumping gases alone. They deliver a constant output with negligible pulsations for a given rotor speed but have the disadvantage that a variable-speed drive is required to provide changes in pumping rate. Close clearances are essential between the moving parts, so that alignment is critical. Erosion and corrosion lead to an increase in the amount of liquid slipping back through the pump. Since gear pumps are subject to clogging, they are not suitable for pumping liquids containing solid particles. Since they do not need inlet and outlet one-way check valves, however, they can be used to pump very viscous liquids. External gear pumps normally range in capacity up to 200 gallons per minute and are designed to generate discharge pressures up to 500 pounds force per square inch, although pumps are available that operate at much higher capacities and discharge pressures. Internal gear pumps are available with capacities up to 1,100 gallons per minute, and discharge pressures are usually limited to 100 pounds force per square inch.

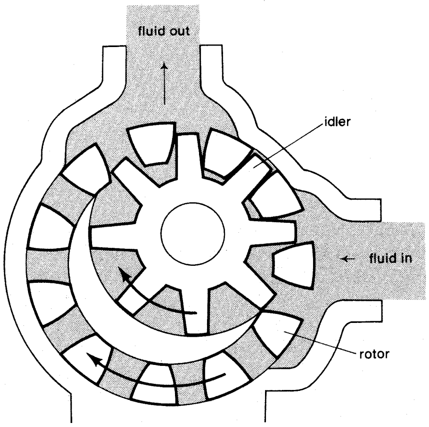


Figure 2: Internal gear pump.

Limitations of gear pumps

Lobe pumps. Lobe pumps operate on the same principle as external gear pumps. Rotors having two, three, or four lobes are used in place of gears. The two rotors are independently driven and usually have a small clearance between them. Lobe pumps have a more pulsating output than external gear pumps and are less subject to wear. Lobe pumps normally range in capacity up to 600 gallons per minute and are designed to generate discharge pressures up to 400 pounds force per square inch, although some have had capacities up to 2,000 gallons per minute. Lobe-type compressors are also used to pump gases. Each of the two rotors has two lobes.

Screw pumps. Screw pumps work on a different principle from the Archimedes screw, which rotates in a cylinder. A helical screw rotor revolves in a fixed casing that is shaped so that cavities formed at the intake move toward the discharge as the screw rotates. As a cavity forms, a partial vacuum is created, which draws fluid into the pump. This fluid is then transferred to the other side of the pump inside the progressing cavity. The shape of the fixed casing is such that at the discharge end of the pump the cavity closes, generating an increase in pressure that forces the fluid into the outlet line. The longer the length of the screw and the shorter the pitch, the greater the pressure developed. If operated against a closed discharge, a screw pump will continue to generate head until the pump bursts; relief valves are necessary to prevent this.

Screw pumps can pump liquids containing vapours or solid particles. They deliver a steady output with negligible pulsations for a given rotor speed but have the disadvantage that a variable-speed drive is required to provide changes in pumping rate. Since screw pumps do not need inlet and outlet one-way check valves, they can be used to pump very viscous liquids. Although screw pumps are bulky, heavy, and relatively expensive, they are robust, slow to wear, and have an exceptionally long life. Screw pumps normally range in capacity up to 3,000 gallons per minute and are designed to generate discharge pressures up to 1,000 pounds force per square inch.

Vane pumps. A sliding vane pump is illustrated in Figure **3.** The rotor is mounted off-centre. Rectangular vanes are positioned at regular intervals around the curved surface of the rotor. Each vane is free to move in a slot. The centrifugal force from rotation throws the vanes outward to form a seal against the fixed casing. As the eccentric rotor revolves, a partial vacuum is created at the suction side of the pump, drawing in fluid. This fluid is then transferred to the other side of the pump in the space between the rotor and the fixed casing. At the discharge side, the available volume is decreased, and the resultant increase in pressure forces the fluid into the outlet line. If operated against a closed discharge, a vane pump will continue to generate head until the pump bursts; relief valves will prevent this. Vane pumps have the advantage that the pumping rate can be varied by changing the degree of eccentricity of the rotor, making them useful in automatic-control systems. They do not need inlet and outlet one-way check valves. They can pump liquids containing vapours or gases but are not suitable for pumping liquids containing solid particles. Vane-type compressors are also used to pump gases.

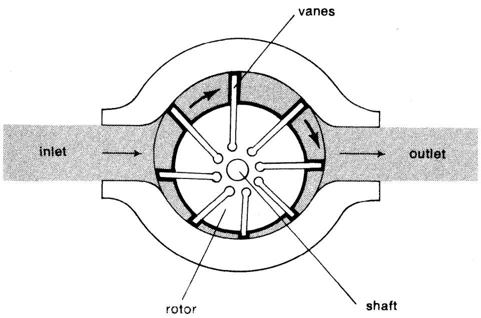


Figure 3: Vane pump.

Advantages of vane pumps

Vane pumps deliver a constant output with negligible pulsations for a given rotor speed. They are robust, and their vanes, easily replaced, are self-compensating for wear. Pumping capacity is not affected until the vanes are badly worn. Vane pumps normally range in capacity up to 2,000 gallons per minute and are designed to generate discharge pressures up to 150 pounds force per square inch. Specially made vane pumps can generate discharge pressures up to 2,000 pounds force per square inch.

Cam pumps. Cam pumps are sometimes called rotating piston or plunger pumps. A variety of cam pumps exists, but all operate on the basis of a rotating eccentric cam.

A rotary plunger pump differs from a true rotary pump in not having a rotating surface in contact with the fluid being pumped. The rotating eccentric cam moves inside a cylindrical plunger that is in direct contact with the fluid. The plunger creates a cavity at the suction side of the pump. This cavity progressively increases as the cam rotates, creating a partial vacuum that draws fluid into the pump. The volume available for the fluid progressively decreases during the cam cycle, and the resultant increase in pressure at the discharge side of the pump forces the fluid into the outlet line.

Cam pumps can pump liquids containing vapours or gases, and do not need inlet and outlet one-way check valves. They are light in weight and require only a small amount of space. They normally range in capacity up to 40 gallons per minute and are designed to generate discharge pressures up to 100 pounds force per square inch. Cam pumps are used only to pump liquids with a relatively low viscosity.

Kinetic. Kinetic pumps can be divided into two classes, centrifugal and regenerative. In kinetic pumps a velocity is imparted to the fluid. Most of this velocity head is then converted to pressure head. Kinetic pumps are a much more recent development than reciprocating positive displacement pumps. Even though the first centrifugal pump was introduced about 1680, the bulk of the development has occurred in this century.

Centrifugal pumps. These include radial, axial, and mixed flow units. 4 radial flow pump is commonly referred to as a straight centrifugal pump. The most common type is the volute pump. The principle is illustrated in Figure 4. Fluid enters the pump near the axis of an impeller rotating at high speed. The fluid is thrown radially outward into the pump casing. A partial vacuum is created that draws more fluid into the pump. The velocity head imparted to the fluid by the vanes of the impeller is converted into pressure head in a progressively widening spiral casing. Flow is continuous.

Volute pump

Volute centrifugal pumps are by far the most common type of pump. They are produced in standard designs for pumping liquids to provide a large capacity range using a relatively few interchangeable parts such as impellers, casings, motor-end frames, etc., with the advantage that industrial users require a smaller inventory of parts. Standard pumps range in capacity from 5 to 500 gallons per minute and generate total heads up to 250 feet, but centrifugal pumps are available in the capacity range 1 to 1,750,000 gallons per minute. Volute centrifugal pumps are robust and relatively inexpensive, quiet, and dependable, and their performance is relatively unaffected by corrosion and erosion. They are compact, simple in construction, and do not require inlet and outlet one-way check valves.

Volute centrifugal pumps can pump liquids containing solid particles, but when pumping liquids containing more than a small amount of vapour, they tend to lose prime; *i.e.*, their suction is broken owing to voids in the fluid. They have the advantage that they do not continue to

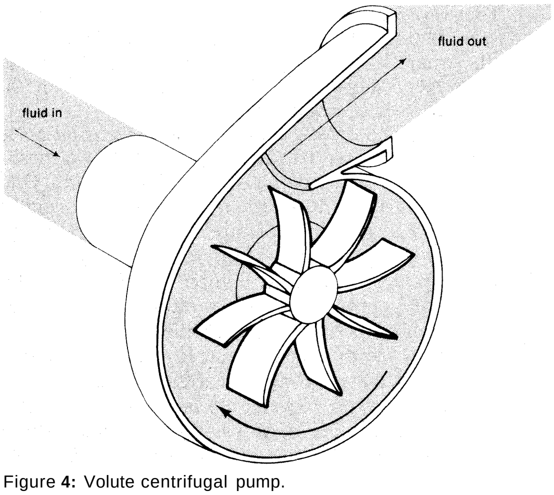


Figure 4: Volute centrifugal pump.

generate pressure when operated against a closed discharge. Volute centrifugal pumps operate best when pumping relatively nonviscous liquids, and their capacity is greatly reduced when used to pump viscous liquids.

The vide variety of volute centrifugal pumps available include vertical and horizontal units and also in-line units, which are designed to fit directly into pipelines. There are also specially designed self-priming pumps, pumps with long drive shafts that are designed to operate in wells, and pumps with several stages.

Another type of radial flow centrifugal pump is the diffuser pump, in which, after the fluid has left the impeller, it is passed through a ring of fixed vanes that diffuse the liquid, providing a more controlled flow and a more efficient conversion of velocity head into pressure head.

In axial flow centrifugal pumps the rotor is a propeller. Fluid flows parallel to the axis as illustrated in Figure 5, and head is generated by the lifting action of the propeller vanes. Diffusion vanes are located in the discharge port of the pump to eliminate the rotational velocity of the fluid imparted by the propeller. Axial flow propeller pumps normally range in capacity from 300 to 100,000 gallons per minute and are designed to generate total heads up to 40 feet. Axial flow compressors are also used to pump gases. In mixed flow pumps, fluid is discharged both radially and axially into a volute-type casing.

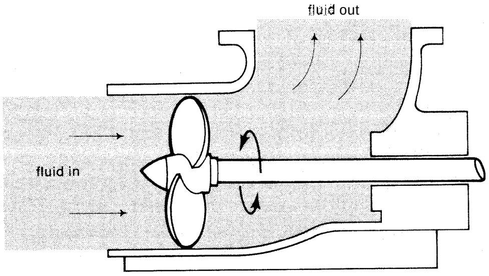


Figure 5. Axial flow centrifugal pump.

Regenerative pumps. A regenerative pump is also called a turbine or peripheral pump. The impeller has vanes on both sides of the rim that rotate in a ringlike channel in the pump's casing. The fluid does not discharge freely from the tip of the impeller but is recirculated back to a lower point on the impeller diameter. This recirculation or regeneration increases the head developed. Because of close clearances, regenerative pumps cannot be used to pump liquids containing solid particles. They can pump liquids containing vapours and gases, and in fact they can pump gases provided that they contain sufficient liquid to seal the close clearances. Regenerative pumps are only suitable for pumping liquids with a relatively low viscosity. Their life span is only about one quarter that of a volute centrifugal pump on comparable service. Regenerative pumps normally range in capacity from one to 200 gallons per minute, and they are designed to generate heads up to 500 feet. Multistage units are used to develop higher heads.

Other types. The air or gas lift pump works on the principle of the displacement of one fluid by a secondary fluid. Gas lifts are used to raise liquids from the bottoms of wells. Compressed gas is introduced into the liquid near the bottom of the well as in Figure 6. The resulting mixture of gas and liquid is lighter and more buoyant than the liquid alone so that the mixture rises and is discharged. Gas lifts have the advantage of no moving parts. Furthermore, they can be used to pump liquids containing solid particles. Capacities range up to 1,000 gallons per minute. Although air or gas lifts are now little used, air lifts were once widely used for pumping water, brine, and oil.

Capacity of gas lift pumps

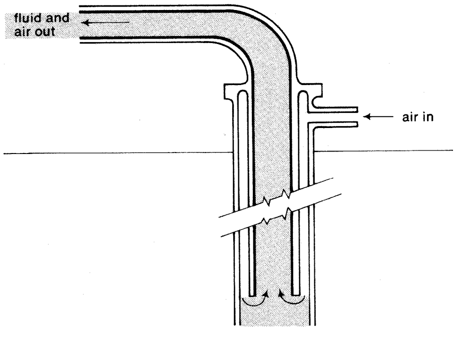


Figure 6: Air or gas lift pump.

Blow case pumps. The blow case or acid egg pump (so named because of its shape) also operates on the principle of the displacement of one fluid by a secondary fluid. The liquid to be pumped is admitted to the egg-shaped container through a one-way check valve. Compressed gas is then admitted to the container and the liquid is forced out through a discharge pipe. The valves can be controlled automatically. In this case two blow cases are used alternately to provide a smoother output. Although blow cases are now little used, they were once widely used for pumping acids.

Jet ejector pumps. in this type of pump, fluid passes through a venturi nozzle (a nozzle with an opening smaller than the pipe to which it is attached) and develops a suction that causes a second stream of fluid to be entrained. (Siphons are jet ejectors.) They are still commonly used to transfer a liquid from one tank to another. In the aspirator pump, water flows through a venturi nozzle and develops a suction for drawing in air. Steam ejectors are widely used for pumping large volumes of vapours and gases at low pressures. Steam at high velocity enters the main body of the pump. The steam transfers some of its momentum to the gas, which is sucked in from the inlet line. A mixture of steam and gas enters the main venturi nozzle known as the diffuser. Kinetic energy is converted to pressure energy and the mixture of steam and gas is compressed. Thus energy in the steam is used to compress gas from a low to a higher pressure. Jet ejector pumps have the advantage of no moving parts. They have been used since about 1850.

Steam ejectors

Hydraulic ram pumps. This type of pump uses the energy of a downward flowing stream of water to lift a proportion of the water to a higher level. Flowing water in the inlet pipe causes a one-way check valve to close. As in a water hammer (in which a flow of water is suddenly stopped, producing a hammering action), kinetic energy is converted to pressure energy, and a second one-way check valve is opened to force some water into the air chamber and up the discharge pipe. The pressure falls in the inlet water pipe, and the first one-way check valve reopens. The compressed air closes the one-way check valve to the air chamber, and the whole cycle is repeated. Approximately 15 percent of the water flowing in the inlet pipe may be raised to a height of five times the fall in the inlet pipe. Hydraulic ram pumps were developed in the late 18th century and are still used in some domestic water systems.

Electromagnetic pumps. These can only be used to pump fluids that are good electrical conductors. The pipe carrying the fluid is placed in a magnetic field and a current passed crosswise through the fluid, so that it is subjected to an electromagnetic force in the direction of flow. The current and the field can be produced in a variety of ways. The principle of the electromagnetic pump is the same as that of the electric motor. Electromagnetic pumps are used for pumping liquid metals, which are used for cooling nuclear reactors.

Applications. Pump manufacturers can provide invaluable help in selecting the right pump for a particular application. The cheapest pump is not the best if it breaks down after a relatively short life. On the other hand, there is little point in buying an expensive pump that will outlast its intended period of use. The properties of the fluid to be pumped and the availability of services such as electricity, steam, compressed air, and water are important factors in deciding the pump to be used. The volute centrifugal pump is easily the most common type of pump for general use. It is versatile, reliable, relatively inexpensive, and well suited to pumping most liquids.

If the liquid being pumped has a tendency to vaporize, as is often the case with warm liquids, it is better to use a positive displacement pump. But if the temperature is too high, pumps with close clearances such as gear pumps run the risk of seizing. Again, a volute centrifugal pump would not be used for a very viscous liquid. If the viscous liquid also contains solid particles, many positive displacement pumps would not be suitable. For this case either a screw or a diaphragm pump would be used.

Vacuum pumps

Vacuum. The atmosphere at sea level normally exerts a pressure of about 14.7 pounds force per square inch. Lower pressures than atmospheric are referred to as vacuum. Vacuum pumps are simply compressors that take in gas at a pressure lower than atmospheric pressure, compress it, and discharge the gas at atmospheric pressure. Since gas at low pressures has a large volume, vacuum pumps tend to be bulky. Steam jet ejectors are extensively used industrially for creating vacuum. Reciprocating piston and rotary-vane-type pumps are also widely used for producing vacuum (see also VACUUM TECHNOLOGY).

Oil and chemical industry. Most types of positive displacement and kinetic pumps can be found in the oil and chemical industry. However, by far the most common type is the volute centrifugal pump. Specially made volute centrifugal chemical pumps are available that have an extra erosion and corrosion allowance and maximum interchangeability of parts. Provision can also be made for leak collection and the water cooling of bearings. Leakage of chemicals may be expensive or even dangerous. Materials of construction must be carefully chosen. In addition to possible erosion and corrosion of the pump, the chemical being pumped may become contaminated. Glass and stoneware volute centrifugal pumps are also available in addition to those made of corrosion resistant alloys. Jacketed pumps for heating with steam or hot water are also used in the chemical industry. Canned pumps in which rotating members are enclosed are used to pump hazardous chemicals. These pumps do not require packed or mechanical seals to prevent leakage. A combined rotor and impeller assembly is driven by the rotating magnetic field of an induction motor that acts, without contact, through the pump housing.

Irrigation and drainage. Irrigation and drainage pumps are of high capacity and normally generate heads up to about 30 feet. Volute centrifugal, axial flow propeller, and mixed flow centrifugal pumps are all used in this work. The most common materials of construction are cast iron for pump casings, bronze for impellers and shaft sleeves, and carbon steel for shafts. In the case of highly polluted water, more corrosion resistant materials may be used. The pumped water usually contains solid particles of silt and mud. Clean water is used to lubricate the bearings on the pump shaft. These pumps are required to stand long continuous service.

Hydrokinetic transmissions

Hydraulic power transmission. Power is transmitted hydraulically by combining a pump and a turbine. Hydraulic power transmission is more flexible than mechanical systems and cheaper and more compact than electrical systems. The pump and turbine can either be housed in a single casing or interconnected by flow and return pipes. The latter system allows the output turbine to be installed in inaccessible locations. By controlling either the pump or the turbine, a continuous variation of speed ratio between the input and output units can be obtained. Centrifugal units are used for systems housed in a single casing. These are called hydrokinetic transmissions and are used in transport vehicles. Interconnected pumps and turbines are usually positive displacement units. Both external gear and reciprocating piston units are in common use.

Deep wells. A special type of volute centrifugal pump is used to raise water from deep, narrow wells. This is a bottom suction pump with a vertical shaft, to which one or more impellers are attached. These pumps are also sometimes called turbine pumps. Discharge occurs along the shaft axis, and they can either be water or oil lubricated. In the latter, a tube enclosing the drive shaft is filled with oil. Long shafts can be avoided by using a close-coupled, totally submersible motor located beneath the pump chamber.

Handling of sewage, wastes, and solids. Specially designed volute centrifugal pumps are used to pump liquids containing solids. Sewerage and industrial waste pumps have large suction and discharge ports that are free from obstructions. They have nonclogging impellers with rounded edges, which are specially designed to be self-cleaning. Some of these pumps are rubber lined. Others have replaceable casing liners.

Condensate pumps. Condensate pumps are usually large volute centrifugal pumps with a number of stages. They are used to extract water at a very low pressure from the hot well of a steam condenser, the container in which condensed steam collects. Condensate pumps have casings of cast iron, impellers of bronze; and shafts of chrome steel. Usually the bearings are lubricated with water to avoid any possible contamination of the feed water with oil. Condensate pumps are generally automatically controlled by the level in the hot well.

Barrel pump

Feed pumps. Multistage volute centrifugal pumps are used to feed hot water to boilers. The impellers are operated at very high rotational speeds, and control of the impeller speed is used to regulate the flow. The pumps must be heated to the temperature of the hot water before starting the feed. Initially the hot water cools and tends to accumulate in the lower part of the pump. The uneven thermal expansion tends to distort the pump casing and shaft. The barrel pump was developed to overcome this problem. It consists of an outer and inner casing. The latter is continually subjected to the discharge pressure of the pump to keep it leak tight. The pumps are made of chrome steel.

Metring pumps. Positive displacement pumps are used to supply a constant rate of liquid irrespective of the discharge pressure. Most metring pumps are of the reciprocating plunger type, although diaphragm metring pumps are common. The rate of liquid is varied by altering the length of the stroke of the plunger. Two or more pumps can be driven by the same motor to provide a constant ratio of two or more rates. To a lesser extent gear pumps are also used to metre liquids.

Statistics on production and use. The manufacture of pumps is a major industry. The most commonly pumped liquid is water. Petrochemical liquids are the second most commonly pumped liquids. The following gives the water requirement in various industries:

1 ton of steel production requires 65,000 U.S. gallons of water

1 ton of paper production requires 70,000 U.S. gallons of water

1 ton of rubber production requires 30,000 U.S. gallons of water

1 ton of petroleum production requires 35,000 U.S. gallons of water

All these industries and many others require pumps.

BIBLIOGRAPHY. F.A. HOLLAND and F.S. CHAPMAN, *Pumping of Liquids* (1966), a text dealing with all aspects of pumping liquids through pipes, with extensive descriptions of centrifugal and positive displacement pumps and their uses; AMERICAN INSTITUTE OF CHEMICAL ENGINEERS, *Pump Manual* (1960), abbreviated but authoritative descriptions of all types of pumps, including their advantages and disadvantages; A. KOVATS, *Design and Performance of Centrifugal and Axial Flow Pumps and Compressors* (1964), a monograph with a basically theoretical approach to the subject, with some information on industrial applications; J.H. PERRY (ed.), *Chemical Engineers' Handbook*, 4th ed. (1963), a concise description of most common pump types together with theoretical and practical information on the pumping of both liquids and gases; S.J. PEERLESS, *Basic Fluid Mechanics* (1967), an engineering degree course dealing with all aspects of fluid mechanics including hydraulic power transmission.

(F.A.H.)

Punctuation

Punctuation is the use of spacing, conventional signs, and certain typographical devices as aids to the understanding and correct reading, both silently and aloud, of handwritten and printed texts. The word is derived from the Latin *punctus*, "point." From the 15th century to the early 18th the subject was known in English as pointing; and the term punctuation, first recorded in the middle of the 16th century, was reserved for the insertion of vowel points (marks placed near consonants to indicate preceding or following vowels) in Hebrew texts. The two words exchanged meanings between 1650 and 1750.

Since the late 16th century the theory and practice of punctuation have varied between two main schools of thought: the elocutionary school, following late medieval practice, treated points or stops as indications of the pauses of various lengths that might be observed by a reader, particularly when he was reading aloud to an audience; the syntactical school, which had won the argument by the end of the 17th century, saw them as something less arbitrary, namely, as guides to the grammatical construction of sentences. Pauses in speech and breaks in syntax tend in any case to coincide; and although English-speaking writers are now agreed that the main purpose of punctuation is to clarify the grammar of a text, they also require it to take account of the speed and rhythm of actual speech.

Syntactical punctuation is, by definition, bad when it obscures rather than clarifies the construction of sentences. Good punctuation, however, may be of many kinds: to take two extreme examples—Henry James would be unintelligible without his numerous commas, but Ernest Hemingway seldom needs any stop but the full point. In poetry, in which the elocutionary aspect of punctuation is still important, and to a lesser degree in fiction, especially when the style is close to actual speech, punctuation is much at the author's discretion. In nonfictional writing there is less room for experiment. Stimulating variant models for general use might be the light punctuation of George Bernard Shaw's prefaces to his plays and the heavier punctuation of T.S. Eliot's literary and political essays.

Ancient Greek practices

Punctuation in Greek and Latin to 1600. The punctuation now used with English and other western European languages is derived ultimately from the punctuation used with Greek and Latin during the Classical period. Much work remains to be done on the history of the subject, but the outlines are clear enough. Greek inscriptions were normally written continuously, with no divisions between words or sentences; but in a few inscriptions earlier than the 5th century BC, phrases were sometimes separated by a vertical row of two or three points. In the oldest Greek literary texts, written on papyrus during the 4th century BC, a horizontal line called the paragraphos was placed under the beginning of a line in which a new topic was introduced. This is the only form of punctuation mentioned by Aristotle. Aristophanes of Byzantium, who became librarian of the museum at Alexandria about 200 BC, is usually credited with the invention of the critical signs, marks of quantity, accents, breathings, and so on, still employed in Greek texts, and with the beginnings of the Greek system of punctuation. Rhetorical theory divided discourse into sections of different lengths. Aristophanes marked the end of the short section (called a comma) by a point after the middle of its last letter, that of the longer section (*colon*) by a point after the bottom of the letter, and that of the longest section (periodos) by a point after the top of the letter. Since books were still being written in tall majuscule letters, like those used in inscriptions and like modern capital letters, the three positions were easily distinguishable. Aristophanes' system was seldom actually used, except in a degenerated version involving only two points. In the 8th or 9th century it was supplemented by the Greek form of question mark (;). The modern system of punctuating Greek texts was established by the Italian and French printers of the Renaissance, whose practice was incorporated in the Greek types cut by Claude Garamond for Francis I of France between 1541 and 1550. The colon is not used in Greek, and the semicolon is represented by a high point. Quotation marks and the exclamation mark were added more recently.

Roman practice

In almost all Roman inscriptions points were used to separate words. In the oldest Latin documents and books, dating from the end of the 1st century BC to the beginning of the 2nd century AD, words were divided by points, and a change of topic was sometimes indicated by paragraphing: the first letter or two of the new paragraph projected into the margin instead of being indented, as they have been since the 17th century. Roman scholars, including the 4th-century grammarian Donatus and the 6th-century patron of monastic learning Cassiodorus, recommended the three-point system of Aristophanes, which was perfectly workable with the majuscule Latin scripts then in use. In practice, however, Latin books in their period were written continuously—the point between words had been abandoned. The ends of sentences were marked, if at all, only by a gap (which might be followed by an enlarged letter) or by an occasional point. The only books that were well punctuated at that time were copies of the Vulgate Bible, for which its translator, St. Jerome (died 419/420), devised punctuation per cola et *commata* ("by phrases"), a rhetorical system, based on manuscripts of Demosthenes and Cicero, which was especially designed to assist reading aloud. Each phrase began with a letter projecting into the margin and was in fact treated as a minute paragraph, before which the reader was expected to take a new breath.

Early medieval practice

During the 7th and 8th centuries, which saw the transition from majuscule to minuscule handwriting (minuscule scripts were usually smaller than majuscule and had projections above and below the body of the letters, as in modern lowercase letters), scribes to whom the Latin language was no longer as well-known as it had been, especially Irish, Anglo-Saxon, and German scribes, to whom it was a foreign language, began to separate words. It was only in the 13th century that monosyllables, especially prepositions, were finally detached from the word following them. To mark sentences, a space at the end became the rule; and an enlarged letter, often a majuscule, generally stood at the beginning of sentences and paragraphs alike. The use of points was somewhat confused by St. Isidore of Seville (died 636), whose encyclopaedia recommended an aberrant version of the three-point system; but a point, high or low, was still used within or after sentences. The ends of sentences were often marked by a group of two or three marks, one of which might be a comma and not a simple point.

St. Jerome's concern for the punctuation of sacred texts was shared by Charlemagne, king of the Franks and Holy Roman emperor, and his Anglo-Saxon adviser Alcuin, who directed the palace school at Aachen from 782 to 796. An important element in the educational revival over which they presided was the improvement of spelling and punctuation in biblical and liturgical manuscripts. It is in the earliest specimens of the new Caroline minuscule script, written at Corbie and Aachen (now in northern France and West Germany, respectively), about 780–800, that the first evidence for a new system of punctuation appears. It soon spread, with the script itself, throughout Europe, reaching its perfection in the 12th century. Single interior stops in the form of points or commas and final groups of stops continued in use; but they were joined by the mark later known as punctus *elevatus* (⹎) and by the question mark (punctus *interrogativus*), of much the same shape as the modern one but inclined to the right. The source of these two new marks was apparently the system of musical notation, called neums, which is known to have been used for Gregorian chant from at least the beginning of the 9th century. *Punctus* elevatus and punctus interrogativus indicated not only a pause and a syntactical break but also an appropriate inflection of the voice. By the 12th century another mark, punctus *circumflexus* (7), had been added to *elevatus* to indicate a rising inflection at the end of a subordinate clause, especially when the grammatical sense of the sentence was still not complete. Liturgical manuscripts in particular, between the 10th and the 13th century, made full use of this inflectional system: it is the origin of the "colon" still used to divide verses of the Psalms in breviaries and prayer books. In the later Middle Ages it was especially the Cistercian, Dominican, and Carthusian orders and the members of religious communities such as the Brethren of the Common Life who troubled to preserve a mode of punctuation admirably adapted to the constant reading aloud, in church and refectory, that characterized the religious life. The hyphen, to mark words divided at the ends of lines, appears late in the 10th century; single at first, it was often doubled in the period between the 14th and 18th centuries.

Late medieval practice

Most late medieval punctuation was haphazard by comparison with 12th-century work—notably in the university textbooks produced at Paris, Bologna, and Oxford in the 13th and 14th centuries. In them, as elsewhere, a form of paragraph mark representing *c* for *capitulum* ("chapter") is freely used at the beginning of sentences. Within the same period the plain point and *punctus elevatus* are joined by the virgule (/), as an alternative form of light stop. Vernacular literature followed the less formal types of Latin literature; and the printers, as usual, followed the scribes. The first printed texts of the Bible and the liturgy are, as a rule, carefully punctuated on the inflectional principle. The profusion of points and virgules in the English books of the printer William Caxton pays remarkably little attention to syntax. Parentheses, used in the same way as now, appear by about 1500. During the 15th century some English legal documents were already being written without punctuation; and British and American lawyers still use extremely light punctuation in the hope of avoiding possible ambiguities.

The beginnings of postmedieval punctuation can be traced to the excellent manuscripts of classical and contemporary Latin texts copied in the new humanistic scripts by Italian scribes of the 15th century. To about 1450 the point and the *punctus elevatus* seem to have been preferred for minor pauses; after that date they are often replaced by the virgule and what is now called the colon (:). The virgule, originally placed high, sank to the base line and developed a curve—turned, in fact, into a modern comma. The Venetian editor and printer Aldus Manutius (Aldo Manuzio; died 1515) made improvements in the humanistic system, and in 1566 his grandson of the same name expounded a similar system in his *Orthographiae ratio* ("System of Orthography"); it included, under different names, the modern comma, semicolon, colon, and full point, or period. Most importantly, the younger Aldo stated plainly for the first time the view that clarification of syntax is the main object of punctuation. By the end of the 17th century the various marks had received their modern names, and the exclamation mark, quotation marks, and the dash had been added to the system.

Punctuation in English since 1600. By the end of the 16th century writers of English were using most of the marks described by the younger Aldo in 1566; but their purpose was elocutionary, not syntactical. When George Puttenham, in his treatise *The Arte* of *English Poesie* (1589), and Simon Daines, in *Orthoepia Anglicana* (1640), specified *a* pause of one unit for a comma, of two units for a semicolon, and of three for a colon, they were no doubt trying to bring some sort of order into a basically confused and unsatisfactory situation. The punctuation of Elizabethan drama, of the devotional prose of John Donne or of Richard Hooker, and indeed of Bunyan's *Pilgrim's Progress* (1678) was almost wholly elocutionary; and it lacked the inflectional element that had been the making of 12th-century punctuation. It was Ben Jonson, in his *English Grammar*, a work composed about 1617 and published posthumously in 1640, who first recommended syntactical punctuation in England. An early example is the 1625 edition of Francis Bacon's *Essayes;* and from the Restoration onward syntactical punctuation was in general use. Influential treatises on syntactical punctuation were published by Robert Monteith in 1704 and Joseph Robertson in 1795. Excessive punctuation was common in the 18th century: at its worst it used commas with every subordinate clause and separable phrase. Vestiges of this attitude are found in a handbook published in London as late as 1880. It was the lexicographers Henry Watson Fowler and Francis George Fowler, in *The King's English,* published in 1906, who established the current British practice of light punctuation. Punctuation in the United States has followed much the same path as in Britain, but the rules laid down by American authorities have in general been more rigid than the British rules.

Shift to syntactical punctuation

The system of punctuation now used by writers of English has been complete since the 17th century. Three of its most important components are the space left blank between words; the indentation of the first line of a new paragraph; and the uppercase, or capital, letter written at the beginning of a sentence and at the beginning of a proper name or a title. The marks of punctuation, also known as points or stops, and the chief parts that they play in the system are as follows.

The end of a grammatically complete sentence is marked by a full point, full stop, or period. The period may also be used to mark abbreviations. The colon (:), which was once used like a full point and was followed by an uppercase letter, now serves mainly to indicate the beginning of a list, summary, or quotation. The semicolon (;) ranks halfway between a comma and a full point. It may be substituted for a period between two grammatically complete sentences that are closely connected in sense; in a long or complicated sentence, it may precede a coordinate conjunction (such as "or," "and," or "but"). A comma (,) is the "lightest" of the four basic stops. As the most usual means of indicating the syntactical turning points in a sentence, it is exposed to abuse. It may be used to separate the elements of a series, before a relative clause that does not limit or define its antecedent, in pairs to set off or isolate words or phrases, or in combination with coordinating conjunctions.

Other punctuation marks used in modern English include parentheses, which serve, like a pair of commas, to isolate a word or phrase; question, exclamation, and quotation marks; the hyphen; and the apostrophe.

Punctuation in French, Spanish, German, and Russian. Since the modern punctuation of all the western European languages stems from the practice of the great Italian and French printers of the 15th and 16th centuries, national differences are not considerable. In French, guillemets (« ») or dashes are used to mark quotations. In Spanish, since the middle of the 18th century, an inverted mark of interrogation or exclamation has stood at the beginning of sentences as well as the normal mark at the end; and quotations may be marked either as in French or as in English. German punctuation, which is still based on rules propounded in 1781, is more rigorously syntactical than the rest: all relative clauses and all clauses beginning with *dass* ("that") must be preceded by a comma. Quotations are marked either by pairs of commas („“) or by reversed guillemets (» «). Letter spacing, as well as italic type, is used for emphasis. Early Russian punctuation was based on Greek practice, since the Cyrillic alphabet is derived from the Greek; and by the 17th century several quite elaborate systems had evolved in different areas. Since the 18th century Russia has used a form of western European punctuation that has much in common with German practice: notably an even wider obligatory use of commas with subordinate and indeed coordinate clauses, and letter spacing (as well as italics) for emphasis. German quotation marks, French guillemets, and dashes may be used for direct speech.

Characteristics of German punctuation

Punctuation in Oriental and African languages. In Hebrew manuscripts written since the 9th century the main use of points is to indicate the vowel sounds, the alphabet being consonantal only. In Biblical texts points and commas are used to mark the middle and end of verses; and in the commentaries points mark the end of sentences. Since the late 18th century, when Jews in Germany began to compose secular texts in Hebrew, the punctuation of such texts has been based on German practice. Early Arabic manuscripts had no punctuation, since the structure of the language ensured that the main and subordinate clauses were readily distinguishable without it. After Arabic began to be printed, European punctuation marks were gradually adopted. The first such mark was the reversed comma; it is now the commonest and indicates a suitable point at which to pause and draw breath.

In Sanskrit, prose texts use one vertical stroke to mark the end of the sentence, and verse texts use one vertical stroke for the end of a line, two for the end of a couplet. In Bengali, Gujarati, Hindi, and Marathi, the vertical stroke is used as in Sanskrit, in conjunction with other marks borrowed from English. The diacritical signs and elements of punctuation found in Tamil were introduced early in the 18th century by a Jesuit missionary.

Before the modern period, the grammatical structure of written Chinese was such that no punctuation was required; but in the 19th century editors of texts began to add hollow circles, intended either to mark the ends of phrases or to emphasize particular passages. Since 1912 some of the European punctuation marks have been adopted, notably the marks of interrogation and exclamation and the comma (the hollow circle serves as full point). Direct speech is indicated either by double inverted commas or by an L-shaped mark placed at a corner of the first and last characters. Characters are capitalized by the addition of a straight or wavy line underneath or at the side, according to whether the text is written horizontally or vertically.

In Japan a complicated system of *kaeriten* and *kunten* marks was used from the 8th century onward to clarify the meaning and grammatical construction of texts in Chinese. As a result of contact with Europeans in the 15th and 16th centuries, a hollow point (O)and a reversed virgule (') were used during the Edo period (1603–1868) as equivalents of the European full point and comma. Since 1868 they have been joined by the solid point (to separate items in a list), by the dash used as in English, and, finally, by the European marks of exclamation and interrogation.

The history of punctuation in Africa is part of the history of the scripts used in different parts of the continent: the Coptic script, based on the Greek alphabet with some additions from demotic writing, for the ancient language of Egypt; a derivative of South Semitic script, known as Ethiopic, for the languages of Ethiopia; Arabic script for speakers of Arabic, Berber, and Swahili; Latin—*i.e.,* European — script for the languages first recorded during and since the 19th century.

BIBLIOGRAPHY, G.V. CAREY, *Mind the Stop,* rev. ed. (1958); and E.H. PARTRIDGE, *You Have a Point There* (1953), are the best guides to modern punctuation as practiced in Britain. The former is short and stimulating; the latter is more exhaustive and includes a chapter on American practice by J.W. CLARK. A comparable American work is the chapter on punctuation in P.G. PERRIN, *Writer's Guide and Index to English,* 4th ed. (1966). The following describe the practice of two famous presses, Oxford University Press and the University of Chicago Press: HORACE HART, *Rules for Compositors and Readers at the University Press,* 37th ed. rev. (1967); and *A Manual of Style,* 12th ed. rev. (1969). For punctuation in antiquity and the Middle Ages, see E.M. THOMPSON, *An Introduction to Greek and Latin Palaeography* (1912); FRANZ STEFFENS, *Lateinische Palaographie,* 2nd ed. (1909); PETER CLEMOES, *Liturgicnl Influence on Punctuation in Late Old English and Early Middle English Manuscripts* (1952); and R.W. SOUTHERN, *The Life of St. Anselm by Eadmer,* pp. 25–34 (1962). For punctuation in and since the Renaissance, especially in Britain, see A.C. PARTRIDGE, *Orthography* in *Shakespeare and Elizabetharz Drama* (1964), of which ch. 14 and 15 and appendix VIII are on punctuation; T.F. and M.F.A. HUSBAND, *Punctuation* (1905); and R.A. SKELTON, *Modern English Punctuation,* 2nd ed. (1950). ALEXANDER BIELING, *Das Princip der deutschen Interpunktion, nebst einer übersichtlichen Darstellung ihrer Geschiclzte* (1880), is useful not only for German practice bat for European punctuation in general since the 15th century.

(T.J.Br.)

Punic Wars

The Punic (Carthaginian) Wars is a name specially given to the wars between Rome and Carthage in the 3rd and 2nd centuries BC. The origin of these conflicts is to be found in the position that Rome acquired about 275 BC as leader and protector of all Italy. Rome's subsequent desire to safeguard the peninsula against foreign interference made it necessary to prevent the neighbouring island of Sicily from falling into the hands of a strong and expansive power. Carthage, on the other hand, had long been anxious to conquer Sicily and to complete the chain of island posts by which it controlled the western Mediterranean from its base in North Africa.

First Punic War (264–241 BC). The proximate cause of the first outbreak was a crisis in the city of Messana (Messina), commanding the straits between Italy and Sicily. A band of Campanian mercenaries, the Mamertini, who had forcibly established themselves within the town and were being hard-pressed in 264 by Hieron II of Syracuse, applied for help to both Rome and Carthage. The Carthaginians, arriving first, occupied Messana and effected a reconciliation with Hieron. The Roman commander, nevertheless, persisted in throwing troops into the city and by seizing the Carthaginian admiral during a parley induced him to withdraw. This aggression involved Rome in war with Carthage and Syracuse.

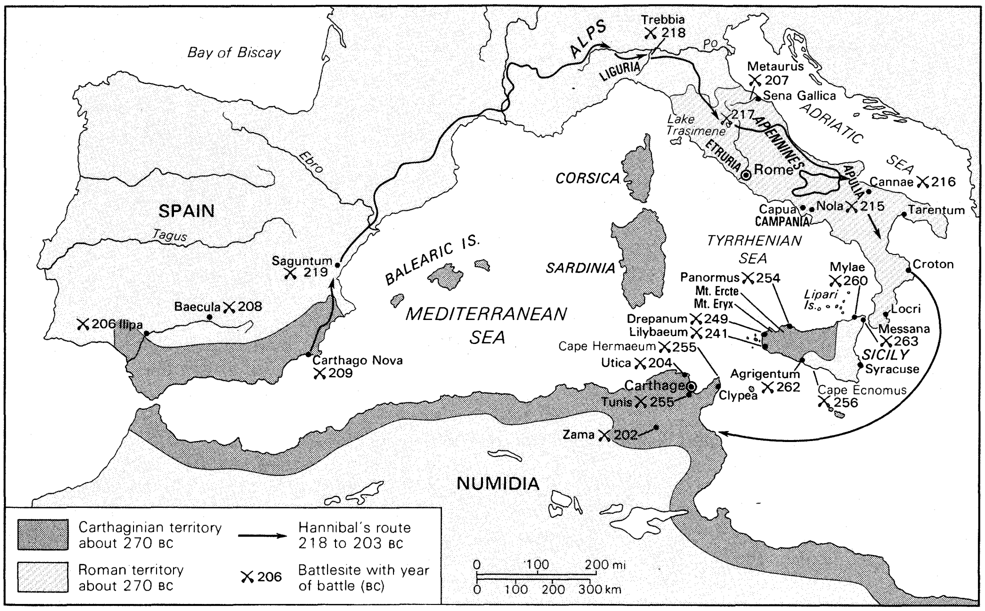
Operations began with a joint attack upon Messana, which the Romans easily repelled. In 263 the Romans advanced with a considerabl: force into Hieron's territory and induced him to seek peace and alliance with them. In 262 they besieged and captured the enemy's hase at Agrigentum. But they made little impression upon the Carthaginian fortresses in the west of the island and upon the towns of the interior.

In 260 the Romans built their first large fleet of standard battleships. At Mylae (Milazzo), off the north Sicilian coast, their admiral Gaius Duilius defeated a Carthaginian squadron of superior manoeuvring capacity by grappling and boarding. This left Rome free to land a force on Corsica (259) and expel the Carthaginians but did not suffice to loosen their grasp on Sicily. A large Roman fleet sailed out in 256, repelled the entire Carthaginian fleet off Cape Ecnomus (near modern Licata), and established a fortified camp on African soil at Clypea (Kélibia in Tunisia). The Carthaginians, whose citizen levy was utterly disorganized, could neither keep the field against the invaders nor prevent their subjects from revolting. After one campaign they were ready to sue for peace, but the terms that the Roman commander M. Atilius Regulus offered were intolerably harsh. Accordingly, the Carthaginians equipped a new army in which, by the advice of a Greek captain of mercenaries named Xanthippus, cavalry and elephants formed the strongest arm. In 255, under Xanthippus' command, they offered battle to Regulus, who had taken up position with an inadequate force near Tunis, outnianoeuvred him, and destroyed the bulk of his army. A second Roman fleet, which reached Africa after defeating the full Carthaginian fleet off Cape Hermaeum (Cape Bon), withdrew all the remaining troops.

The campaign against Sicily

The Romans now directed their efforts once more against Sicily. In 254 they captured the important fortress of Panormus (Palermo), but when Carthage threw reinforcements into the island the war again came to a standstill. In 251 or 250 the Roman general Caecilus Metellus at last brought about a pitched battle near Panormus, in which the enemy's force was effectively crippled. This victory was followed by an investment of the chief Punic base at Lilybaeum (Marsala), together with Drepanum (Trapani), by land and sea. The besiegers met with a gallant resistance and in 249 were compelled to withdraw by the loss of their fleet in a surprise attack upon Drepanum, in which the admiral P. Claudius Pulcher was repulsed with a loss of 93 ships. This was the Romans' only naval defeat in the war, but their fleet had suffered a series of grievous losses by storm, and now it was so reduced that the attack upon Sicily had to be suspended. At the same time, the Carthaginians, who felt no less severely the financial strain of the prolonged struggle, reduced their forces and made no attempt to deliver a counterattack. The only noteworthy feature of the ensuing campaigns is the skillful guerrilla war waged by a new Carthaginian commander, Hamilcar Barca, from his strong positions on Mt. Ercte (247/246–244) and Mt. Eryx (modern Erice; 244–242) in western Sicily, by which he effectually screened Lilybaeum from any attempt on it by the Roman land army.

In 242 Rome resumed operations at sea. By a magnificent effort on the part of private citizens, a fleet of 200 warships was equipped and sent out to renew the blockade of Lilybaeum. The Carthaginians hastily collected a relief force but in a battle fought off the Aegates, or Aegusae (Aegadian) Islands, west of Drepana, their fleet was



The western Mediterranean during the Punic Wars.
Adapted from *Westermann* Grosser *Atlas zur Weltgeschichte;* Georg Westermann Verlag, Braunschweig

caught at a disadvantage and mostly sunk or captured (March 10, 241). This victory, by giving the Romans undisputed command of the sea, rendered certain the ultimate fall of the Punic strongholds in Sicily. The Carthaginians accordingly opened negotiations and consented to a peace by which they ceded Sicily and the Lipari Islands to Rome and paid an indemnity of 3,200 talents.

The interval between the First and Second wars (241–218 BC). The loss of naval supremacy not only deprived the Carthaginians of their predominance in the western Mediterranean but exposed their overseas empire to disintegration under renewed attacks by Rome. The temper of the Roman people was soon made manifest during a conflict that broke out between the Carthaginians and their discontented mercenaries. A gross breach of the treaty was perpetrated when a Roman force was sent to occupy Sardinia, whose insurgent garrison had offeredto surrender the island (238). To the remonstrances of Carthage the Romans replied with a declaration of war and only withheld their attack upon the cession of Sardinia and Corsica and the payment of a further indemnity.

From this episode it became clear that Rome intended to use the victory to the utmost. To avoid complete humiliation Carthage had no resource but to humiliate its adversary. The recent complications of foreign and internal strife had indeed so weakened the Punic power that the prospect of renewing the war under favourable circumstances seemed remote. But the scheme of preparing for a fresh conflict found a worthy champion in Hamilcar Barca, who sought to compensate for the loss of Sicily by acquiring a dominion in Spain where Carthage might gain new wealth and form a fresh base of operations against Rome. Invested with an unrestricted foreign command, he spent the rest of his life in founding a Spanish empire (237–228). His work was continued by his son-in-law Hasdrubal and his son Hannibal, who was placed at the head of the army in 221. These conquests aroused the suspicions of Rome, which in a treaty with Hasdrubal confined the Carthaginians to the south of the Ebro. At some point Rome also entered into relations with Saguntum (Sagunto), a town on the east coast, south of the Ebro. In 219 Hannibal laid siege to Saguntum and carried the town in spite of a stubborn defense.

It has always been a debatable point whether or not his attack contravened the new treaty. The Romans certainly took this view and sent to Carthage to demand Hannibal's surrender. But his defiant policy was too popular to be disavowed; the Carthaginian council upheld Hannibal's action and drew upon itself a declaration of war.

Second Punic War (218–201 BC). It seemed as though the superiority of the Romans at sea must enable them to choose the field of battle. They decided to embark one army for Spain and another for Sicily and Africa. But before their preparations were complete, Hannibal began that series of operations by which he dictated the course of the war for the greater part of its duration. Realizing that, so long as the Romans commanded the resources of an undivided Italian confederacy, no foreign attack could beat them down beyond recovery, he conceived the plan of cutting off their supply of strength at the source by carrying the war into Italy and causing a disruption of the league. His chances of ever reaching Italy seemed small, for the sea was guarded by the Roman fleets and the land route was long and arduous.

Hannibal's strategy

But the very boldness of his enterprise contributed to its success; after a six months' march through Spain and Gaul and over the Alps, which the Romans were nowhere in time to oppose, Hannibal arrived (autumn 218) in the plain of the Po with 20,000 foot and 6,000 horses, the pick of his African and Spanish levies.

His further advance was here disputed by some Roman troops, but the superiority of the Carthaginian cavalry and the spread of insurrection among the Gaulish inhabitants forced the defenders to fall back upon the Apennines. At the end of the year the Roman army was reinforced by the division from Sicily and led out to battle on the banks of the Trebbia. Hannibal, by superior tactics, repelled the assailants with heavy loss and thus made his position in north Italy secure.

In 217 the campaign opened in Etruria, into which the invading army, largely reinforced by Gauls, penetrated by an unguarded pass. A rash pursuit by the Roman field force led to its being entrapped on the shore of Lake Trasimene (Trasimeno) and destroyed with a loss of at least 15,000 men. This catastrophe left Rome completely uncovered; but Hannibal, having resolved not to attack the capital before he could collect a more overwhelming force, directed his march toward the south of Italy, where he hoped to stir up the peoples who had formerly been the most stubborn enemies of Rome. The Italians, however, were slow everywhere to join the Carthaginians; and a new Roman army under the dictator Q. Fabius Maximus ("Cunctator") that, without ever daring to close with Hannibal, dogged his steps on his forays through Apulia and Campania, prevented his acquiring a permanent base of operations.

The Battle of Cannae

The eventful campaign of 216 was begun by a new, aggressive move on the part of Rome. An exceptionally strong field army, variously estimated at between 48,000 and 85,000 men, was sent to crush the Carthaginians in open battle. On a level plain near Cannae in Apulia, which Hannibal had chosen for his battleground, the Roman legions delivered their attack. Hannibal deliberately allowed his centre to be driven in by their superior numbers, while Hasdrubal's cavalry wheeled round so as to take the enemy in flank and rear. The Romans, surrounded on all sides and so cramped that their superior numbers aggravated their flight, were practically annihilated; and the loss of citizens was perhaps greater than in any other defeat that befell the republic.

The moral effect of the battle was no less momentous. The south Italian peoples at least found courage to secede from Rome, the leaders of the movement being the people of Capua, at the time the second greatest town of Italy. Reinforcements were sent from Carthage, and several neutral powers prepared to throw their weight into the scale on Hannibal's behalf. At first sight it seems strange that the Battle of Cannae did not decide the war. But the great resources of Rome, though terribly reduced in respect to both men and money, were not yet exhausted. In north and central Italy the insurrection spread but little and could be sufficiently guarded against with small detachments. In the south, the Greek towns of the coast remained loyal, and the numerous Latin colonies continued to render important service by interrupting free communication between the rebels and detaining part of their forces.

In Rome itself the quarrels between the nobles and commons, which had previously unsettled Roman policy, gave way to a unanimity unparalleled in the annals of the republic. The guidance of operations was henceforth left to the Senate, which, by maintaining a persistent policy until the conflict was brought to a successful end, earned its greatest title to fame.

The subsequent campaigns of the war in Italy assumed a new character. Though the Romans contrived at times to raise 200,000 men, they could spare only a moderate force for field operations. Their generals, among whom the veterans Fabius and M. Claudius Marcellus frequently held the most important commands, rarely ventured to engage Hannibal in the open and contented themselves with observing him or skirmishing against his detachments. Hannibal, whose recent accessions of strength were largely discounted by the necessity of assigning troops to protect his new allies or secure their wavering loyalty, was still too weak to undertake a vigorous offensive. In the ensuing years the war resolved itself into a multiplicity of minor engagements, which need not be followed out in detail. In 216 and 215 the chief seat of war was Campania, where Hannibal, vainly attempting to establish himself on the coast, experienced a severe repulse at Nola.

In 214 the main Carthaginian force was transferred from Apulia in hopes of capturing Tarentum (Taranto). Though Croton and Locri on the southern coast had fallen into his hands, Hannibal still lacked a suitable harbour by which he might have secured his overseas communications. For two years he watched in vain for an opportunity to surprise the town, while the Romans narrowed down the sphere of revolt in Campania and defeated other Carthaginian commanders.

In 213–212 the greater part of Tarentum and other cities of the southern seaboard at last came into Hannibal's power. But in 212 the Romans found themselves strong enough to place Capua under blockade. They severely defeated a Carthaginian relief force and could not be permanently dislodged even by Hannibal himself. In 211 Hannibal made a last effort to relieve his allies by a feint upon Rome itself, but the besiegers refused to be drawn away from their entrenchments, and eventually Capua was starved into surrender.

Its fall was a sign that no power could in the long run uphold a rival Italian coalition against Rome. After a year of desultory fighting, the Romans in 209 gained a further important success by recovering Tarentum. Though Hannibal still won isolated engagements, he was slowly being driven back into the extreme south of the peninsula.

The campaign of 207

In 207 the arrival of a fresh invading force produced a new crisis. Hasdrubal, who in 208–207 had marched overland from Spain, appeared in northern Italy with a force scarcely inferior to the army that his brother had brought in 218. After levying contingents of Gauls and Ligurians, he marched down the east coast with the object of joining his brother in central Italy for a direct attack upon Rome itself. By this time the steady drain of men and money was telling so severely upon the confederacy that some of the most loyal allies protested their inability to render further help. Yet by exerting a supreme effort the Romans raised their war establishment to the highest total yet attained and sent a strong field army against each Carthaginian leader.

The danger to Rome was chiefly averted by the prompt insight and enterprise of the consul Gaius Nero, who commanded the main army in the south. Having discovered that Hannibal would not advance beyond Apulia until his brother had established communications with him, Nero slipped away with part of his troops and arrived in time to reinforce his colleague Livius, whose force had recently got into touch with Hasdrubal near Sena Gallica (Senigallia).

The combined Roman army frustrated an attempt of Hasdrubal to elude it and forced him to fight on the banks of the Metaurus (Metauro). The battle was evenly contested until Nero, by a dexterous flanking movement, cut the enemy's retreat. Hasdrubal himself fell, and the bulk of his army was destroyed.

The campaign of 207 decided the war in Italy. Though Hannibal still maintained himself for some years in southern Italy, this was chiefly due to the exhaustion of Rome after the prodigious strain of past years and the consequent reduction of Roman forces. In 203 Italy was finally cleared of Carthaginian troops. Hannibal, in accordance with orders received from home, sailed back to Africa; and another expedition under his brother, Mago, which had sailed to Liguria in 205 and endeavoured to rouse the slumbering discontent of the people in Cisalpine Gaul and Etruria, was forced to withdraw.

Campaigns in Sicily and Spain. Concurrently with the great struggle in Italy, the Second Punic War was fought out on several other fields. It will suffice merely to allude to the First Macedonian War (214–205), which King Philip V commenced when the Roman power seemed to be breaking up after Cannae. This compelled the Romans to stretch their already severely strained resources still further by sending troops to Greece, but the diversions that Roman diplomacy provided for Philip in Greece, and the maintenance of a patrol squadron in the Adriatic Sea, prevented any effective cooperation on his part with Hannibal.

In view of the complete stagnation of agriculture in Italy the Romans had to look to Sardinia and Sicily for their food supply. Sardinia was attacked by Carthaginians in 215, but a small Roman force sufficed to repel the invasion. In Sicily a more serious conflict broke out. Some isolated attacks by Punic squadrons were easily frustrated by the strong Roman fleet. But in 215 internal complications arose. The death of Hieron II, Rome's steadfast friend, left the realm of Syracuse to his inexperienced grandson, Hieronymus. Flattered by the promises of Carthaginian emissaries, the young prince abruptly broke with the Romans, but before hostilities commenced he was assassinated. The Syracusan people now repudiated the monarchy and resumed their republican constitution; but, by threats of terrible punishment at the hands of Rome, they were led to cooperate with the Carthaginians.

Roman attack on Syracuse

The attacks of a Roman army and fleet under Marcellus, which speedily appeared before the town, were completely baffled by the mechanical contrivances of the Syracusan mathematician Archimedes (213). Meantime, the revolt against Rome spread in the interior, and a Carthaginian fleet gained control of towns on the south coast.

In 212 Marcellus at last broke through the defense of Syracuse and, in spite of the arrival of a Carthaginian relief force, mastered the whole town in 211. A guerrilla warfare followed in which the Carthaginians maintained the upper hand until in 210 they lost their base at Agrigentum. They were dislodged from their remaining positions, and by the end of the year Sicily was wholly under the power of Rome.

The conflict in Spain was second in importance only to the Italian war. From this country the Carthaginians drew large supplies of troops and money that might serve to reinforce Hannibal; hence it was in the interest of the Romans to challenge their enemy within Spain. Though the force that Rome at first spared for this war was small in numbers and rested entirely upon its own resources, the generals Pnblius and Gnaeus Scipio, by skillful strategy and diplomacy, not only won over the peoples north of the Ebro and defeated the Carthaginian leader Hasdrubal Barca in his attempts to restore communication with Italy but carried their arms along the east coast into the heart of the enemy's domain.

But eventually their successes were nullified by a rash advance. Deserted by their native contingents and cut off by Carthaginian cavalry, among which the Numidian prince Masinissa rendered conspicuous service, the Roman generals were killed and their troops destroyed (211).

The Battle of Ilipa

Disturbances in Africa prevented the Punic commanders from exploiting their success. Before long the fall of Capua enabled Rome to transfer troops from Italy to Spain; and in 210 the best Roman general of the day, the young son and namesake of P. Scipio, was placed in command. He signalized his arrival by a bold and successful *coup de main* upon the great arsenal of Carthago Nova (Cartagena) in 209. Though after an engagement at Baecula (Bailen; 208) he was unable to prevent Hasdrubal Barca from marching away to Italy, Scipio profited by his departure to push back the remaining hostile forces the more rapidly. A last effort by the Carthaginians to retrieve their losses with a fresh army was frustrated by a great victory at Ilipa, near Seville, and by the end of the year 206 they had been driven out of Spain.

The war in Africa. In 205 Scipio, who had returned to Rome to hold the consulship, proposed to follow up his victories by an attack upon the home territory of Carthage. Though the presence of Hannibal in Italy deterrred Fabius and other senators from sanctioning this policy, Scipio gradually overbore all resistance. He built up a force, which he organized and supplemented in Sicily, and in 204 sailed across to Africa. He was met there by a combined levy of Carthage and King Syphax of Numidia and for a time was penned to the shore near Utica. But in the spring he extricated himself by a surprise attack upon the enemy's camp, which resulted in the total loss of the allied force by sword or fire.

In the campaign of 203 a new Carthaginian force was destroyed by Scipio on the Great Plains 75 miles from Utica, their ally Syphax was captured, and the renegade Masinissa reinstated in the kingdom from which Syphax had recently expelled him. These disasters induced the Carthaginians to sue for peace; but before the very moderate terms that Scipio offered could be definitely accepted, a sudden reversal of opinion caused them to recall Hannibal's army for a final trial of war and to break off negotiations. In 202 Hannibal assumed command of a composite force of citizen and mercenary levies stiffened with a corps of his veteran Italian troops.

After an abortive conference with Scipio he prepared for a decisive battle near Zama. Scipio's force was smaller in numbers but well trained throughout and greatly superior in cavalry. His infantry, after evading an attack by the Carthaginian elephants, cut through the first two lines of the enemy but was unable to break the reserve corps of Hannibal's veterans. The battle was ultimately decided by the cavalry of the Romans and their new ally Masinissa, which by a manoeuvre recalling the tactics of Cannae took Hannibal's line in the rear and destroyed it.

The Carthaginians having thus lost their last army again applied for peace and accepted the terms that Scipio offered. They were compelled to cede Spain and the Mediterranean islands still in their hands, to surrender their warships, to pay an indemnity of 10,000 talents within 50 years, and to forfeit their independence in affairs of war and foreign policy.

The Second Punic War, by far the greatest struggle in which either power engaged, had thus ended in the complete triumph of Rome. This triumph is not to be explained in the main by any faultiness in the Carthaginians' method of attack. The history of the First Punic War and that of the Second outside of Italy prove that the Romans were irresistible on neutral or Carthaginian ground. Carthage could only hope to win by invading Italy and using the enemy's home resources against him. The failure of Hannibal's brilliant endeavour to realize these conditions was not due to any strategical mistakes on his part. It was caused by the indomitable strength of will of the Romans, whose character during this period appears at its best, and to the compactness of their Italian confederacy, which no shock of defeat or strain of war could entirely disintegrate.

It is this spectacle of individual genius overborne by corporate and persevering effort that lends to the Second Punic War its peculiar interest.

Cato's influence

The Third Punic War (149–146 BC). The political power of Carthage henceforth remained quite insignificant, but its commerce and material resources revived in the 2nd century with such rapidity as to excite the jealousy of the growing mercantile population of Rome and the alarm of its more timid statesmen. Under the influence of these feelings the conviction—sedulously fostered by Cato the Censor—that "Carthage must be destroyed" overbore the scruples of more clear-sighted statesmen. A *casus belli* was readily found in a formal breach of the treaty, committed by the Carthaginians in 150 when they resisted Masinissa's aggressions by force of arms. A Roman army was dispatched to Africa; and although the Carthaginians consented to make reparation by giving hostages and surrendering their arms, they were goaded into revolt by the further stipulation that they must emigrate to some inland site where they would be debarred from commerce.

By a desperate effort they created new war equipment and prepared their city for a siege (149). The Roman attack for two years completely miscarried until in 147 the command was given to a young officer who had distinguished himself in the early operations of the war—Scipio Aemilianus, the adopted grandson of the former conqueror of Carthage. Scipio made the blockade stringent by walling off the isthmus on which the town lay and by cutting off its sources of supplies from overseas. His main attack was delivered on the harbour side, where he effected an entrance in the face of a determined and ingenious resistance. The struggle did not cease until he had captured house by house the streets that led up to the citadel.

Of a city population which, according to Strabo, had exceeded a quarter of a million, only 50,000 remained at the final surrender. The survivors were sold into slavery; the city was razed to the ground and its site condemned by solemn imprecations to lie desolate forever. The territory of Carthage, which had recently been much narrowed by Masinissa's encroachments, was converted into a Roman province under the name of "Africa."

BIBLIOGRAPHY

Ancient sources: The two main surviving authorities are Polybius and Livy, namely Polybius bk. 1 (First War), bks. 2 and 3, and fragmentary accounts in books 7–15 (Second War), and fragments of bks. 36–38 (Third War); and Livy bks. 21–30 (Second War), based partly on Polybius and partly on less reliable Roman annalists. Apart from Appian's account of the Third War (*Libyca*, 67–135, based partly upon Polybius), the subsidiary authors add little of value.

Modern works: Two general accounts are given by TENNEY FRANK *et al.*, *Cambridge Ancient History*, vol. 7, ch. 21, and vol. 8, ch. 2–5, 15 (1928–30); and by H.H. SCULLARD, *A History of the Roman World, 753 to 146 B.C.*, 3rd ed. (1961), while ARNOLD J. TOYNBEE, *Hannibal's Legacy*, 2 vols. (1965), provides much background information. An indispensable work is F.W. WALBANK, *Historical Commerztary on Polybius*, vol. 1 (1957), vol. 2 (1967), for full discussion of all problems, G. DE SANCTIS, *Storia dei Romani*, vol. 3, pt. 1–2 (1916–17), and vol. 4, pt. 3 (1964), is fundamental. U. KAHRSTEDT, *Geschichte der Karthager von 218–146* (1913), provides detailed source-criticism. J. KROMAYER and G. VEITH, *Antike Schlnchtfelder*, vol. 3–4 (1912–31), and *Schlachten-Atlas zur antiken Kriegsgeschichte*, Rom. Abt. 1–2 (1922), are the standard works on purely military aspects, while the role of sea-power is handled by J.H. THIEL, *History of Roman Sea-Power Before the Second Punic War* (1954), and *Studies on the History of Roman Sea-Power in Republican Times* (1946),

ch. 2 for the Second Punic War. S. GSELL, *Histoire ancienne de l'Afrique du Nord,* vol. 1–4, esp. vol. 3 (1913–20), deals fully with the Carthaginian side. On the Spanish and African campaigns, see H.H. SCULLARD, *Scipio Africanus in the Second Punic War* (1930), and *Scipio Africanus: Soldier and Politician* (1970). For the Third Punic War, see A.E. ASTIN, *Scipio Aemilianus* (1967). On Hannibal, see E. GROAG, *Hannibal als Politiker* (1929). On the coinage of the Second Punic War, see E.S.G. ROBINSON, *Numismatic Chronicle* (1964).

(M.Car./H.H.S.)

Punishment

Punishment may be defined as the infliction of some pain, suffering, loss, or social disability as a direct consequence of some action or omission on the part of the person punished. The punishment may consist of death, physical assault, detention, loss of civil and political rights, or banishment. There must be a perceived relationship of legitimacy between the punisher and the punished: the agent of punishment must be in a position of legitimate authority over the punished, and the action or omission must be seen to merit the punishment by reference to a set of pre-existing criteria by which such acts or omissions may be judged.

The legitimacy of the actions of the punisher is the primary distinction between punishment and other forms of coercion and constraint, for unless the agent who inflicts the pain or deprivation has authority to do so, his action constitutes a form of assault or civil wrong, both of which are themselves transgressions. Thus, for example, in Western society, a parent is the lawful guardian of a minor and is expected to constrain the activities of his child, if necessary by punishment. But whereas the law permits the parent to slap a child for misbehaviour in the street, it does not accord such a right to each and every adult who may be present, and a stranger who intervened might well be held accountable in the courts for assault.

Punishment is normally personalized, in that it is applied to particular individuals on the basis of their perceived wrongdoing. It is occasionally applied in a collective context without reference to the extent of individual responsibility or accountability. So-called collective punishments are generally regarded as of dubious or marginal legality, and their use, apart from that in the schoolroom, is to all intents and purposes confined to situations in which the state or an occupying army has resort to extreme methods of coercion. The Nazi army of occupation in Czechoslovakia thus put the entire adult male population of the village of Lidice to death as a punishment for the shooting of Reinhard Heydrich, the country's Gestapo ruler; in the 19th century the Austrian general Julius Haynau ordered the wholesale flogging of women after putting down the Hungarian revolt against Austrian rule in 1849. Such actions are nowadays defined as war crimes.

THE ROLE OF PUNISHMENT IN SOCIETY

Punishment in the family. Punishment may be inflicted in a variety of social contexts, as, for example, in the family and the school and by such corporate bodies as professional associations and trade unions.

In all human societies the family is seen as having authority over its members. In the simpler societies of hunters and food-gatherers the social and economic unit of society is usually so small that all social authority is vested within the family. In larger scale but technologically undeveloped peasant societies, the family is still vested with considerable authority over its members, notwithstanding that the state may possess a comparatively well-developed political structure. In the Rome of classical antiquity, the law thus underwrote the authority of the paterfamilias (male head of the family), who exercised control over all the members of his household, including those who had married into it, as yell as slaves and free servants. Similarly, in the later Middle Ages and well into the preindustrial era in the English-speaking world, young men who were apprenticed to master craftsmen were not only regarded as part of the household but were subject to the discipline of the master. A parallel situation obtained in the universities, where the masters of colleges were seen as having both the right and the duty to control the behaviour of students, if necessary by the infliction of punishment.

Analogies of the family in other social structures

In understanding the role of the family as a legitimate agent of social control through punishment, it is important to bear in mind the connection between the family and other social institutions, such as religious organizations and the state. Many primitive religions, and many sophisticated ones, emphasize a connection between the structure of the family and the transcendental social order. Among primitive peoples, reverence for ancestors is related to the part played by the spirits of the dead in controlling the living. Thus, for example, among the people of Ontong Java, British Solomon Islands, the dead will uncomfortably haunt the living who fail to honour their obligations to kin. In the strongly anthropomorphic Judeo-Christian tradition the human family is seen as mirroring in greater or less degree the divine family.

The role of punishment within the family in the modern period may best be considered by reference to Victorian attitudes and the reaction to them. Within what may be loosely termed the Victorian family there is reason to believe that the punishment of children was both frequent and harsh, inflicted for transgressions against a moral code that was itself starkly puritanical. Punishments covered a range of behaviour. Children were required to be not only honest and truthful but also industrious, obedient, and respectful toward their elders. They were also constrained in their attempts to acquire sexual information; and as sex was regarded as wholly sinful outside marriage and a guilty necessity within it, such an activity as masturbation was discouraged not merely by punishment but by the assertion that it resulted in blindness and insanity.

Insofar as the social attitudes of most sections of the community tend to reflect those of the most dominant and prestigious groups, the puritanical authoritarianism that characterized the Victorian upper and middle class family was mirrored throughout society. What was true of Victorian England was also true to a greater or lesser extent among the well-to-do in the United States and in continental Europe. It was not until the publication of the early work of Sigmund Freud that doubt began seriously to be cast upon child-rearing practices that leaned so heavily upon repression and physical punishments. Perhaps the most important conclusion to be drawn from Freud's work in this context is that punishment, particularly of a physical nature, is often related at an unconscious level with disturbances and anxieties within the punisher. Through the mechanism of projection the parent may thus punish what he feels guilty about himself; a parent who is himself sexually repressed may, for example, punish a child for looking at representations of the female body or for masturbation. Punishment may also be inspired by the sadistic sexual tendencies of the parent. Equally, it is now generally accepted among experts in the field of child care that physical punishments, especially when applied in the anal region, may result in erotic stimulation of an unhealthy nature.

The effect of Freudian psychology on the character of punishment within the family in the last 50 years has been considerable, as it has been on the character of punishment in educational institutions and, to a lesser extent, within the penal system. The writings of Dr. Benjamin Spock have had a very wide influence on middle class parents, especially in the United States, in establishing a concept of normality regarding discipline that is functional rather than specifically moral in content.

Punishment in moral and religious thought. As indicated above, the connection between the family and religious beliefs is a long and complex one. This is especially the case with regard to religions of a distinctly anthropomorphic character in which ancestors (often the recently dead) play an important part in sanctioning the activities of the living or in which the social organization of the gods is perceived as being essentially similar to that of earthly society.

Among primitive peoples it is, broadly speaking, the case that the sanctions of ancestors are directed toward the maintenance of the status quo, which in turn depends upon the adequate performance of those kinship obligations upon which so much of social and economic life in primitive society revolves. Punishments may take a variety of forms, among which haunting by ghosts is extremely common. The incest taboo is also sometimes supported by religious sanctions, and the failure of crops, death of cattle, human sickness, or the visitation of natural disaster is perceived as a consequential punishment for unlawful intercourse or failure to marry according to the prescribed rules of selection.

Divine sanction in the Judeo-Christian tradition

In more sophisticated cultures there tends to be a greater distinction between ethical systems maintained by punishments imposed by men themselves and those maintained by religious beliefs. In classical mythology the gods were often guilty of actions that, if committed by men, would justify sanctions. Thus, if a local chief carried away the nubile daughter of a neighbour, a violent reaction might follow; for a god to carry away a maiden while disguised as a swan or a bull could, on the other hand, be classified as a form of divine right. In the Judeo-Christian tradition, however, the connection between ethical beliefs and divine sanctions is extremely clear. Yahweh, the tribal god of the ancient Israelites, was distinct from the tribal deities of other pastoral peoples in that he existed within a distinctly monotheistic context. The Israelites were told unambiguously, "You shall have no other gods before me." Yahweh not merely created the world and all that was in it but in the Ten Commandments set out a comprehensive code of social as well as ritual regulations. These were designed to enforce both tribal order and tribal identity. Defiance of these divinely instituted rules, as well as of specific injunctions, was followed by dramatic and severe punishment. Adam and Eve were cast from the garden, women were turned into salt pillars, and cities destroyed. Those, such as the Egyptians, who failed to accord with divine requirements, ran the risk of being visited by plague and death.

It has been suggested by psychoanalytic writers that the vengeful and sometimes unpredictable God of the Old Testament, who indulged in such sadistic testing of loyalty and faith as the projected sacrifice of Isaac or the endless trials of Job, represents the projection of the powerful and feared father. In *Totem and Taboo*, Freud argued that the fear of incest derives from the guilt attendant upon the slaying of the father (the elder of the tribe) by his sons in order that they might have sexual access to the nubile women whom the father monopolized. This Oedipal theme is, in fact, recurrent in the mythology of a wide range of cultures.

The coming of Christianity marked a change in the pattern of divine–human relations in that the death of Christ, the Messiah, was the sacrificial reconciliation between God and man. Henceforth, the doctrine of the Atonement guaranteed that man could be absolved from his sin. He had a clear choice: to obey the law of God and enjoy eternal bliss or defy it at the cost of certain and eternal damnation.

An essential component of punishment in the Judeo-Christian tradition is the guilt of the recipient. In the Old Testament the punitive situations described are all paradigmatic examples of the fate awaiting transgressors. Christian teaching, however, modified the position by introducing the concept of mercy. Jesus, when the woman taken in adultery is brought before him, thus not only suggests to the crowd that the most blameless among them should cast the first stone of the traditional punishment but merely tells the woman to "sin no more." At an early stage the church thus developed the rituals of confession and penance. After the Resurrection the disciples are told, "Whose sins you shall forgive, they are forgiven them." A condition of absolution from guilt was the performance of penance, and in the early church this was often severe, involving fasting, physical degradation, and often self-inflicted corporal punishment. By the later Middle Ages, penances had become more symbolic and in the case of the rich could be performed by the donation of land or precious goods to the church.

The relationship between guilt and punishment so clearly worked out in medieval theology has had a profound influence on the position of punishment within the criminal law. The legal doctrine of *mens rea* ("guilty mind"), holding that there can be no crime without intent, derives directly from it and has for centuries enabled the English common law to protect those whose actions, although intrinsically unlawful, were committed without malice. The increase in administrative legislation has considerably eroded the traditional concept of guilt based upon *mens rea.*

An equally important feature of Christian thinking concerned with punishment is that it is related to the offender's responsibility. Roman law had always regarded a child of less than seven years as being incapable of guilt on the grounds that he was too young to comprehend the distinctions between right and wrong, and the church similarly did not make confession available to children below this age on the grounds that Baptism, having absolved them of the original sin of Adam, was sufficient to ensure their salvation in case of death. A characteristic of penance was that it was tailored, so to speak, to the degree of responsibility of the penitent.

A young man sorely tempted by the wiles of women to commit fornication might thus be given a lesser penance than an experienced lecher. A condition of penance or expiation was, however, the honest resolve to sin no more, any mental reservations on the subject by the penitent automatically invalidating the absolution. Although this latter feature was never adopted by the criminal law, one may observe in penance the early origins of the individualized sentence, which is nowadays regarded as an important part of sentencing under the criminal law.

Secularization of penal philosophy in the Enlightenment

The destruction of the unity of Christendom and its universal theology by the Reformation was followed by the increasing secularization of the European nation-state, not least in the administration of punishment. Although the churches still preached the doctrine of divine punishment, it was that imposed by the state that came to assume immediate importance in society. Increasingly, the infliction of punishment by the state through the courts became dominated by the pragmatic necessity of maintaining order in the face of social change. It was the philosophers of the Enlightenment, rather than theologians, who came to offer–new views on the subject. According to the 13th-century philosopher St. Thomas Aquinas, the order of society was the result of a divine plan; according to the 17th-century English philosopher Thomas Hobbes, man had voluntarily abdicated the right of self-help against transgressors of his rights to the sovereign in order to avoid the chaos of the state of nature. Both theories justified the existence of temporal power, the right of the state or sovereign to make laws and punish offenders. The thinkers of the Enlightenment were, however, concerned to emphasize both the rationality of human action and the extent to which the state oppressed the liberty of the subject through harsh and essentially unreasonable laws, resulting from a combination of political tyranny and judicial corruption. By the middle of the 18th century an intellectual climate had developed that produced not only the American and French revolutions but a fundamental re-examination of penal philosophy.

THEORIES OF PUNISHMENT

The most influential writer of the period was the Italian Cesare Beccaria (1738–94), whose *Dei deletti e delle pene (An Essay on Crimes and Punishments)* was published in Italian in 1764 and in English three years later. It had the public approval of Voltaire and was one of the most widely read polemics of the time. Beccaria was committed to a contractual theory of society as being necessary to avoid war and social anarchy. He argued that the rightful source of law was in the legislature, and that its interpretation ought not to be left to the judiciary. Beccaria tended to equate flexibility on the bench with judicial tyranny. He took the view that the function of the judges was the determination of guilt, else the spirit of the law would represent no more than the capricious logic of the judiciary. He argued that the state had a right to punish, and that there should be set out a scale of crimes and punishments ranging from acts that would bring about a dissolution of the state to those that caused the smallest harm to an individual citizen, between these poles lying all actions that, being contrary to the common good, were punishable. The act, and not the intention, was the measure of harm done. The end of punishment, he reasoned, was not to make the offender miserable, nor to compensate for the harm that had been done, but was solely preventive. Punishments ought therefore to make the greatest deterrent impact consistent with the infliction of the least pain on the offender. To be effective, punishment should be certain, speedy, and uniform for particular crimes. When punishments were seen to have no preventive value, they should be abolished. Prisons should be used as a means of training offenders to be useful citizens. Above all, the laws should be clear and unambiguous and be widely known to all citizens.

The weaknesses of some of Beccaria's otherwise enlightened and humane thinking became evident when some of his proposals were incorporated into the French Revolutionary Penal Code of 1791. In ignoring individual social differences between offenders and assuming that the pursuit of pleasure and the avoidance of pain were the mainsprings of human behaviour, he had placed reliance on a deceptively simple concept of crime. The importance of Beccaria's *Essay,* however, lies less in the substance of his particular views and more in the fact of its galvanic effect upon the penal reformers of the time. In his critical challenging of existing modes cf punishment, he paved the way for a radical re-appraisal of many aspects of criminal law.

Bentham's penal philosophy

The writings of the English philosopher and legal reformer Jeremy Bentham brought some of Beccaria's ideas into closer focus. His most important ideas are contained in *An Introduction to the Principles of Morals and Legislation* (1789). As a philosopher, Bentham belonged to the Utilitarian school, which believed that men seek pleasure and avoid pain and that society ought to be so organized as to secure the greatest happiness of the greatest number. By constructing a great "felicific calculus" that enabled him to demonstrate the goodness or badness of particular acts, Bentham evolved an enormously complicated catalog of crimes and punishments. In essence, his penal philosophy was based upon the notion that punishment ought to be sufficient to prevent crime but no more. Like Beccaria before him, he thought that the ineffectiveness of savage penal laws was self-evident, and none could deny that the penal systems of 18th-century Europe were brutal in the extreme. The strength of Bentham's view lay in the usefulness of the idea that excessive punishment might be counter-productive—men argued that they might as well be hanged for a sheep as for a lamb. The prime weakness was not unlike Beccaria's; that is, a reliance on hedonistic psychology and an assumption that all men act rationally.

Toward the end of the 19th century the study of crime, which had been developing along sociological lines, shifted its emphasis as a result of the work of the Italian Cesare Lombroso, a surgeon who had been influenced not only by the positivism of the French sociologist Auguste Comte but also by the work of Charles Darwin. In *L'uomo delinquente* (1876) he argued that the criminal was an atavistic biological specimen who could be identified by a number of physical stigmata; and although the theory of atavism was to undergo considerable modification both by Lombroso and his disciples, the views of this school of criminology remained consistently positivist. In positivist terms, moral responsibility is of no account, only social responsibility. There could be no degrees of responsibility, for either the act was injurious to the public welfare or it was not. If a public building had been destroyed by arson, for example, the character of the arsonist was in that context irrelevant. Hence, punishment, which some philosophers had argued ought to embody retribution, should consist only of those measures that, taking into account the dangerousness of the offender, needed to be taken to protect society.

Functions of punishment in society. *Deterrence and social control.* Theories of deterrence rely on two assumptions: first, that the potential offender will act rationally in his own best interest and seek to avoid pain; second, that he will both remember past experiences and anticipate the consequences of his intended future actions. For this reason, penal systems from early times have frequently been characterized by the public infliction of punishment, often of a brutal nature. The death penalty has been almost universally used, although less among the simpler peoples than the more advanced. It has taken a variety of forms, among which decapitation by sword or ax and hanging are the most common. In imperial Rome, crucifixion was commonly employed for thieves and slaves, and being hurled alive from the l'arpeian Rock was a penalty in Rome itself, as was being fed to the wild animals in the Colosseum. Crucifixion was also used in Japan until the 19th century. In medieval Europe, hanging was used for offenders of low status and decapitation reserved for persons of quality. Heretics were burned, as were witches. In England the penalty of hanging, drawing, and quartering was especially brutal. The victim was first partly strangled and then while still aiive drawn or disemboweled; his entrails were burned and his body then literally butchered into four pieces. This punishment was normally reserved for treason. In France, traitors were dispatched either by being pulled apart by draft horses, or by being broked upon a large wheel to which they were attached.

Apart from death the physical maiming of offenders was commonplace. In the Middle Ages and well into the 17th century, branding with hot irons, blinding, and the amputation of ears, hands, and tongues were practiced, along with flogging. The effect of such punishments was often the death of the offender. Placing the offender in the stocks or pillory enabled him to be pelted by the crowd and to sustain serious physical injury. Stoning appears to have been employed from ancient times among the pastoral peoples of the Middle East and has frequent biblical mention.

Seamen and soldiers were, until recent times, also subjected to a considerable degree of corporal punishment as well as the death sentence for disciplinary offenses, invariably inflicted before an assembled company. During the period of colonial settlement, transportation to plantations or dispatch to serve as galley rowers was used as an alternative to capital punishment.

Rationale of publicly inflicted punishment

Publicly inflicted punishment has the effect of encouraging the spectators to participate in a situation in which the values that have been violated are reaffirmed; alternatively, it demonstrates the power of the political authority and the price to be paid for disobedience. The deterrent theory varies in its effectiveness. Where the potential offender is rational and the penalty not so severe as to be counterproductive, it may be reasonably successful as a medium of social control. It is least effective in dealing with mentally abnormal offenders. Its success also relies very heavily on certainty of detection and conviction, and many criminologists and police officers argue that certainty of detection is more important than severity of sentence in the prevention of crime.

Retribution. The lex talionis, typified in the doctrine of "an eye for an eye," formed part of the law of many cultures and may be seen in residual form in many popular expressions of opinion about crime and punishment in present-day society. The notion that the offender ought to suffer in proportion to his wickedness has support from a wide range of philosophical and theological sources. At a psychological level, suffering on the part of the criminal may give some comfort to the victim, but it is less easy to match one injury for another than is always realized. Although death may be matched by death, a sexual murder or a rape cannot be replicated. A further difficulty lies in estimating the degree of wickedness of the offender as distinct from the intrinsic evil of the crime. Thus, the rape-murders of young children may be classified as wicked, and acts from which reasonable men ought properly to shrink, but such acts are usually committed by persons of distinctly unsound mind. Another weakness of the retributive view is that the provision of punishment that matches the putative wickedness of the offender may be so severe as to be of no use in deterring crime. It may lead to a point at which punishment is inflicted whether it has any deterrent value for others or not, and for this reason penologists and judges are cautious not to accord retribution too great an emphasis in sentencing. The exception to this is situations in which sentences are passed in order to express the horror of society at the crime, as, for example, at incest or homosexual offenses in certain cultures.

Reformation. It is important in this context to distinguish between constructive penal treatment, aimed at bringing about a change in the offender's behaviour by whatever means, and the infliction of punishment, which must, by definition, involve some pain or suffering by him. Some penologists argue that it is not possible to punish and reform simultaneously, as punishment is essentially damaging to the offender, and if retributive, totally at variance with the course of action most likely to achieve reform. Others would prefer to strip punishment of its moral overtones and regard it as a form of conditioning, which properly used may bring about change. Thus, the deprivation of liberty by imprisonment may produce a degree of shock such as to enable the offender to examine realistically the consequences of his behaviour.

While many legislatures and most penologists have supported the idea that reform ought to take priority in dealing with offenders, many judges—especially in Britain and the United States, where rising crime rates are the source of much public concern—have expressed grave doubts about the wisdom of this view. They have argued that the courts must reflect a public abhorrence of crime and that justice demands that some attempt be made to impose punishment fitting to the crime. There is, of course, no reason why reform and deterrence should not coexist, providing there is adequate information about the offender's mental state and social competence and that commonsense distinctions are made between types of crime. It is thus reasonable to assume that traffic violations may be largely controlled by deterrent sentences, but unreasonable to assume that sex offenses can be similarly contained.

Difficulties of preventive sentencing

The protection of society. Attitudes toward punishment are shaped to some extent by the assumed social danger that some offenders present to society. In some cases this danger may be removed by treating the offender either by medical means or by social retraining. In others the prognosis may be very poor, such that after a period of incarceration the offender may return again to be a danger in the community. For such offenders it may be argued that a preventive sentence—their removal for an indefinite period or for natural life, for example—is the proper course. While many societies have accepted life imprisonment as the just penalty for certain crimes such as murder, notwithstanding that many murderers can be safely returned to the community, most have been reluctant to use long or indeterminate sentences to deal with those who, though their prognosis may be poor, have not yet committed the most serious crimes. An example of this is the violent sex offender who may be a potential killer but about whose future criminal career there may still be uncertainty. It can be argued that to condemn a man in advance is essentially unjust. On the other hand, many legislatures have evolved statutes enabling persistent offenders of all kinds to be detained for longer periods than a particular offense would merit in terms of the tariff system. Both the effectiveness and the social acceptability of preventive sentences must rest upon the accuracy of predictive techniques. Although prediction methods in criminology have improved considerably over the last three decades, they are still insufficiently exact to be acceptable as the sole basis of preventive sentencing.

The effectiveness of punishment. To be effective, punishment, whether in the familial, educational, or penal context, must be certain and relatively swift, else its conditioning characteristics are diluted or lost. It must not be excessive, or it will be subject to the law of diminishing returns. It must be of itself morally acceptable—that is, not seen as cruel or unreasonable either by society or the recipient, who must ideally accept the legitimacy of its imposition. It must also be limited to actions that are by consensus defined as unacceptable, for the application of penal sanctions for infringements of laws that have little or no public support brings the law into contempt. Punishment, or the application of the criminal sanction, must also bring about some general social results collectively as well as individually.

Importance of prisons in modern penal systems

The core of modern penal systems is imprisonment, although it is increasingly true that the proportion of offenders sentenced to noncustodial disposal is producing a situation in which it is becoming more rare for first offenders to be imprisoned. Evidence from the Western world suggests that the frequency of convictions is a better statistical predictor of future crime than the type of sentence imposed. Among noncustodial measures, fines and probation have a high rate of success except among those who have previous convictions, although the picture differs in detail among different countries. By and large, juveniles who have served custodial sentences have a poorer prognosis than those who have not. Of males who are imprisoned for the first time, between 65 and 85 percent are not reconvicted, but among those who are, the subsequent recidivism rate tends to increase progressively. Thus, most long-term prisons and penitentiaries are filled with men whose prognosis is poor, and the fact tends to be reflected in rates of parole violation. The ineffectiveness of imprisonment may be ascribed to various factors, including overcrowding, understaffing, and lack of resources, which result in the regime's becoming negatively custodial rather than constructively therapeutic; contact with other inmates that may reinforce rather than diminish criminal attitudes; criminogenic factors in the outside world, which on the prisoner's release may be more dominant than the attempts made by the prison at resocialization; the presence of a group of highly successful professional criminals who perceive prison as an occupational hazard; and the deleterious effects of institutionalization on socially inadequate offenders.

A major difficulty in assessing the effectiveness of penal sanctions stems from the fact that criminological data are very rarely gathered in the form of experimental observations. Thus, for example, the fact that in England and Wales the imposition of fines appears to be the most successful sanction may be related to the possibility that courts use fines in the cases of offenders who are least likely to persist in crime. They are also used for minor offenses as well as major ones. Without the possibility of controlled experiment, which might take a group of comparable offenders, fining some and dealing with others in alternative ways, evaluation is complex and fraught with technical difficulties.

BIBLIOGRAPHY. The literature on crime and punishment is considerable, and the subject has engaged the attention of many great writers, including Defoe, Dickens, Hugo, and Dostoevsky. The following references are largely confined to the more philosophical and academic writings readily available in English.

A.C. EWING, *The Morality of Punishment* (1929); F.A.P. LONGFORD, *The Idea of Punishment* (1961); B.F. WOOTTON, *Crime and the Criminal Law* (1963); and H.L.A. HART, *Punishment and Responsibility* (1968), contain broad discussions of the ethical problems involved in punishment and their application within the context of the criminal law. A.C. MacINTYRE, *A Short History of Ethics* (1966), may also be consulted. GEORG RUSCHE and OTTO KIRCHHEIMER, *Punishment and Social Structure* (1939, reprinted 1968), argue an interesting thesis, namely, that types of punishments are related to the prevailing conditions of the economic and social structure and less to abstract ethical principles. J.C. FLUGEL, *Man, Morals and Society* (1945); HENRY WEIHOFEN, *The Urge to Punish: New Approaches to the Problem of Mental Irresponsibility for Crime* (1956); and GILES PLAYFAIR, *The Punitive Obsession* (1971), consider the problem from a viewpoint that derives from the psychoanalytic approach to the study of crime and deviant behaviour, while the educational innovator A.S. NEILL, *Hearts Not Heads in the School* (1945), considers the problem in the context of the progressive school that relies on pupil self-government. F.G. JACOBS, *Criminal Responsibility* (1971), deals with the question of criminal

responsibility in law; while D.A. THOMAS, *Principles of Sentencing* (1970); NIGEL WALKER, *Crime and Punishment in Britain,* 2nd rev. ed. (1968), and *Sentencing in a Rational Society* (1969); and C.H. ROLPH, *Common Sense About Crime and Punishment* (1961), are concerned with relating criminal punishments to the needs and policy of a liberal society. MARC ANCEL, *La Défense sociale nouvelle* (1954; Eng. trans., *Social Defence: A Modern Approach to Criminal Problems,* 1965), is an important work, placing the matter in a broad European perspective. HERMANN MANNHEIM, *Group Problems in Crime and Punishment* (1955); EDWIN H. SUTHERLAND and DONALD R. CRESSEY, *Criminology,* 8th ed. (1970); H.E. BARNES and N.K. TEETERS, *New Horizons in Criminology,* 3rd ed. (1959); and NORMAN JOHNSTON, LEONARD SAVITZ, and MARVIN E. WOLFGANG (eds.), *The Sociology of Punishment and Correction,* 2nd ed. (1970), are primarily criminological texts from which much information on theories of punishment and penal systems may be derived. CHRISTOPHER HIBBERT, *The Roots of Evil* (1963), is a social history of crime and punishment; and J.P. CONRAD, *Crime and Its Correction* (1965), provides a wide international perspective on punishment within the penal context. D.C. GIBBONS, *Changing the Lawbreaker: The Treatment of Delinquents and Criminals* (1965); and DANIEL GLASER, *The Effectiveness of a Prison and Parole System* (1964), are examples of analyses of the effectiveness of various penal methods. L.W. FOX, *The English Prison and Borstal Systems* (1952), although primarily about the English system, has useful material on 19th-century penal thinking in both Britain and the United States.

(T.P.M.)

Punjab (India)

Punjab, a state in the Indian Union, is situated in northwest India. It is bounded on the north by Jammu and Kashmir, on the east by Himachal Pradesh, on the south by Haryāna and Rājasthān, and on the west by Pakistan.

Land of the five rivers

The word Punjab is a compound of two Persian words, *panj* ("five) and *āb* ("water"), thus signifying the land of five waters, or rivers. The first known use of it occurs in the writings of the Muslim traveller Ibn Baṭṭūṭah (*q.v.*), who visited India in the 14th century. The origin can perhaps be traced to *pañca nada,* Sanskrit for "five rivers," the word used before the advent of Muslims with a knowledge of Persian to describe the meeting point of the Jhelum, Chenāb, Rāvi, Beās, and Sutlej rivers, before they joined the Indus. The name Punjab subsequently came to be used for the land through which these rivers flow. As applied to the present Indian state of Punjab, it is strictly a misnomer, for, since the partition of India in 1947, only two rivers, the Sutlej and the Beās, lie within its territory.

The state in its present form came into existence on November 1, 1966, as a result of a territorial reorganization carried out on the basis of language. The capital city is Chandigarh, which is also the headquarters of the government of Haryāna, a new state comprising most of the residuary, non-Punjabi-speaking areas of the older unit. Chandigarh, planned by the Swiss architect and city planner Le Corbusier, was built as a capital city for what became the Indian Punjab after the partition.

Punjab covers an area of 19,495 square miles (50,491 square kilometres); its population in 1971 was 13,731,000. (For coverage of an associated physical feature, see INDUS RIVER.)

History. The foundations of the present Punjab (historical Pañjāb) may be said to have been laid by Bandā Singh Bahādur, a hermit who became a military leader and who, with his fighting band of Sikhs, temporarily liberated the eastern part of the province from Mughal rule in 1709–10. Bandā Singh's defeat and execution in 1716 were followed by a long, drawn-out struggle between the Sikhs on one side and the Mughals and Afghans on the other. In 1764–65 the Sikhs established their sway in the region. Ranjit Singh (1780–1839) built up the Punjab into a powerful kingdom and attached to it the provinces of Multān, Kashmir, and Peshāwar. In 1849 the Punjab fell to the troops of the British East India Company and came under British rule. When independence was achieved in 1947, the Punjab was split between the new sovereign states of India and Pakistan, the smaller, eastern portion becoming part of India.

Since independence, the history of the Indian sector of the Punjab has been dominated by Sikh agitation for a Punjabi-speaking state, first led by Master Tara Singh and later by his political successor, Sant Fateh Singh. In November 1956 the Punjab was enlarged by the merger with it of the state of Pepsu, short for Patiāla and East Punjab States Union, which was formed by the amalgamation of the princely territories of Patiāla, Jind, Nābha, Faridkot, Kapūrthala, Kalsia, Mālerkotla, and Nālagarh. Political and administrative leadership to the new Punjab was provided by Sardar Partap Singh Kairon, Congress chief minister of the state from 1956 to 1964. Demands for the formation of a new state by the consolidation of the Punjabi-speaking areas and the elimination of the Hindi-speaking portions were eventually agreed to by the government of India. On November 1, 1966, the state of Punjab was divided on the basis of language into the two new states of Punjab and Haryāna (the Hindi-speaking area).

The absorption of the princely states

Physical geography. More than 95 percent of the total area of Punjab is a flat plain, sloping gently from about 900 feet above sea level in the northeast to about 550 feet in the southwest. Physiographically, it is divisible into three parts: (1) the Siwālik Hills, in the northeast, about 900 to 3,000 feet high (covering a small fraction of the state's area); (2) the narrow, undulating foothill zone dissected by closely spaced seasonal torrents, locally known as *cos,* several of which terminate in the plain below without joining any stream; and (3) the flat tract, with fertile alluvial soils. The low-lying floodplains along the rivers and the slightly elevated flat uplands between them are distinguishable within the plain. Sand dunes occur in the southwest and six to nine miles west of the Sutlej River.

The climate. Punjab has an inland subtropical location, and its climate is continental, being semi-arid to subhumid. Summers are very hot, the mean temperature during June being 93" F (34" C), rising above 113" F (45" C) on exceptionally hot days. Winters are fairly cold, the January mean temperature being 55° F (13" C), with night temperatures occasionally touching the freezing point. Annual rainfall is highest in the Siwālik Hills in the northeast, where it is about 49 inches, and decreases gradually to about 14 inches in the southwest. More than 70 percent of the annual rainfall is concentrated in the months of the monsoon winds, July to September. Winter rains from the western cyclones that occur from December to March account for nearly 15 percent of the total rainfall.

Vegetation. With the growth of human settlement over the centuries, Punjab has been largely cleared of its vegetal cover. The existing vegetation is tropical and deciduous. The *shisham* or *sissoo* tree (of which the wood is used for construction and furniture making), the chinaberry, the *peepal* (which bears a small fig), and jujube tree, and *kikar* (gum arabic) are among the most common trees. Over large parts of the Siwālik Hills, bush vegetation has succeeded tree vegetation as a result of extensive deforestation. New types of trees have been planted on the hillsides and along roads in recent years.

Animal life. Wildlife in Punjab includes the blue bull (also called the nilgai, a large bluish-gray animal), the wild boar, rabbits, jackals, foxes, and deer of various kinds. The commonest birds are crows, sparrows, doves, pigeons, parrots, partridges and peacock. The koel (cuckoo) is a summer visitor and the mountain crane a winter one. Waterfowl include cranes, herons, and geese. Snakes abound in summer, being especially in evidence in the rainy months; the cobra, viper, and krait (an extremely venomous kind of snake) are poisonous. Land in Punjab is precious, and agriculture impinges heavily on wildlife.

Population. The people of Punjab are mainly descendants of Aryan tribes that entered India from the northwest about 1500 BC. Relics of a highly developed pre-Aryan civilization have been discovered at Rūpar, near Chandigarh. Successive waves of invaders—Greeks, Parthians, Kushans, and Huns—became assimilated with the early Aryans. The Jats, Rājputs, and other allied peoples are the products of this intermixture. Some members of the scheduled castes—formerly called the untouchables but named Harijan (people of God) by Gandhi—and of aboriginal tribes have traces of the physical features of the Dravidian people.

At the 1961 census, the population of Punjab was 11,255,056, and it is estimated that it had risen to almost 14,000,000 by April 1971. The annual growth rate rose from 1.6 percent in 1951 to 2.16 percent in 1961 and was 2.20 percent for the period 1961–71. The earlier increase was due to the rapid decline in mortality, the expectation of life at birth having gone up from an average of about 50 years (males) and 45 years (females) in 1951–60 to about 60 years (males) and 55 years (females) in 1966–70. Punjab's birthrate, about 42 per thousand (1966–70), is higher than the rate for all India (38.6), despite its unfavourable sex ratio of 871 females to 1,000 males.

The language pattern

In 1961 Punjabi was recorded as being the mother tongue of about two-thirds of the people and Hindi of one-third. Punjab is the only state in India with a majority of Sikhs, who account for 60 percent of the population. The Sikhs are followers of an Indian religion that had its origin in the teachings of Nānak, a 15th-century Gurū (teacher; see SIKHISM). Hindus account for 38 percent and Muslims and Christians for 1 percent each, while the remaining 1 percent are Jains, Buddhists, and followers of other religions.

Punjab is comparatively urbanized. Twenty-five percent of its population live in cities and towns, as compared with about 20 percent for the whole of India. It has four cities with populations of over 100,000 (Amritsar, 433,000; Ludhiāna, 401,000; Jullundur, 296,000; and Patiāla, 152,000). Though the cities are fast-growing centres of industry and manufacture, more than 70 percent of the state's population is still dependent on agriculture.

Administration. As in other states of the Indian Union, the governor, the constitutional head of the administration, is appointed by the president of India on the advice of the Union cabinet and acts through a council of ministers responsible to a legislative assembly of 104 members, elected every five years on the basis of adult franchise. The legislature was bicameral until January 1970, when the upper chamber (the legislative council) was abolished. The council of ministers is headed by a chief minister, who chooses its members from the legislative assembly.

The executive officials belong to a permanent, nonpolitical secretariat, members of which are selected by union or state public-service commissions. The administration is assisted in policy making and expertise both by the secretariat and by a number of standing boards and committees, such as the Planning Board, the Rural Development Board, and the Electricity Board.

The deputy commissioner is the chief executive of the district (which forms the basic unit of administration), and the state is divided into 11 such districts, grouped into the two divisions of Patiāla and Jullundur, each headed by a divisional commissioner, whose functions are chiefly supervisory. The 11 districts are Gurdāspur, Amritsar, Bhatinda, Jullundur, Hoshiārpur, Firozpur, Ludhiāna, Kapūrthala, Sangriir, Rūpar, and Patiāla.

The judiciary

The judiciary is independent of the executive. At its head is the high court, which, as constituted at present, is common to Punjab and Haryāna. Justice at the district level is administered by a district and sessions judge, assisted by subjudges and magistrates. Appeals from decisions of the high court are directed to the supreme court of the Union.

Social conditions. ***Education.*** In 1969 there were about 2,000,000 students in schools in the state. Of these, about 185,000 were in high schools. There were more than 7,000 primary schools, 900 middle schools, and 1,100 high schools. The number of teachers in primary and middle schools was about 31,000 and in high schools, 23,000. Schools are maintained largely by the state. Education is compulsory and free for pupils aged 6 to 11. Secondary education is also free in state schools.

According to the census of 1971, the rate of literacy among the population was about 34 percent (41 percent for males and 26 for females), as compared with the all-India percentage of 29 (40 percent for males and 18 for females).

There are three universities in the state: the Punjabi University at Patiāla, the Gurū Nānak University at Amritsar, and the Punjab Agricultural University at Ludhiāna. A fourth university, which serves three of the Punjab districts, is located at Chandigarh. Of 121 colleges in the state, 29 are professional, 6 medical, and 2 engineering. There are 106 technical institutes providing industrial and occupational courses. There is a state department to encourage and organize sports, and Punjab youth is well represented in Indian athletic teams.

Expenditure on education amounted to about 26 percent of the state budget in 1970–71.

Health. Punjab enjoys higher survival prospects and better health conditions than other states. The death rate has dropped from almost 19 per thousand in 1951–60 to about 8 per thousand in 1966–71. The number of hospitals is 510, with a total of 11,000 beds. There is one qualified doctor for every 2,600 persons. The state has undertaken an extensive family-planning program, especially in rural areas.

Welfare services. The state has schemes for the uplift of the scheduled castes—who form 22 percent of the population—through scholarships, recruitment to public services, and help in finding employment.

Developments that have had a revolutionary effect on the lives of the people are the spread of electrification in rural areas and the dissemination of vocational and cultural education through broadcasting.

Numerous social services are provided by government and voluntary organizations. These include the care of infirm or handicapped persons, neglected children, widows, and destitute persons. The government also provides pensions for the aged who have no means of subsistence. There is a network of employment exchanges to assist the unemployed.

Economy. The economy of the state is characterized by a dynamic agriculture; a developing small-scale industry; and a growing per capita income despite a rapid rise in population.

Although Punjab claims no more than 2.9 percent of India's total cultivated area, in 1969–70 it accounted for almost 7 percent of its total output of food grains and 24 percent of its total output of wheat. During the six years ending June 1970, the production of wheat in the state more than doubled, rising from 2,535,300 to 5,291,000 tons. This phenomenal progress has been described as the Green Revolution. In 1970 the state supplied 2,400,000 tons of wheat to the union government. The principal food-grain crops are wheat, maize (corn), rice, *bājrā* (millet), barley, and pulses (the edible seeds of various leguminous crops, such as peas, beans, and lentils). Other crops include cotton, sugarcane, potatoes, and oilseeds.

The Bhākra–Nāngal project

The increase in agricultural production is due largely to the successful application of modern agricultural technology. By June 1969, 67 percent of the net area sown had been brought under irrigation—the highest percentage of any Indian state. Canals watered about 49 percent of the irrigated area and wells, tube wells, and pumping sets the remaining 51 percent. The Bhākra–Nāngal multipurpose river-valley project, designed to irrigate 550,000 hectares (1,359,000 acres) of land in Punjab, 680,000 hectares (1,680,300 acres) in Haryāna, and 230,000 hectares (568,300 acres) in Rājasthān, is the largest of its kind in Asia. Other significant changes are the adoption of high-yielding varieties of semidwarf Mexican wheats and hybrids, the use of chemical fertilizers and pesticides, and increased mechanization.

The benefits of improved agriculture, augmented by co-operative societies and community-development programs, have tended to give comparatively more help to the bigger farmers, with the result that the disparity in farm-family incomes has increased considerably.

Punjab has practically no mineral or fossil-fuel resources, and only 4 percent of its total area is under forest. Consequently, industrial development in the state is associated with agriculture or devoted to consumer goods. In January 1970 there were about 23,000 small and 120 medium and large industrial units, employing 186,000 workers. The principal industries produce cotton, woollen, and silk textiles, sugar, machine tools, agricultural implements, cycles, sewing machines, sports goods, electrical goods, automobile parts, flour, and milk products or engage in cotton ginning and pressing.

The total revenue of the state for 1970–71 was estimated at 1,359,000,000 rupees (Rs. 7.28 = $1 U.S.; Rs. 19 = £1 sterling, in May 1972), two-thirds of which was tax revenue, with indirect taxes contributing the larger share. The total expenditure was estimated at Rs. 1,164,700,000, about 57 percent of which constituted development expenditure.

Increasing agricultural production

Since the introduction of planning in 1951, Punjab has maintained a growth rate higher than that of any other Indian state in agriculture and has achieved the highest per capita income. In 1967–68, its net industrial output was valued at Rs. 11,265,000,000, and per capita income was Rs. 828. Owing to the failure of the economy fully to utilize the growing manpower available, however, this represented an increase of only about 22 percent in real per capita income since 1960–61. The number of unemployed persons in 1969 was estimated at almost 500,000.

The state's fourth five-year plan (1969–74) postulated an investment of Rs. 2,935,600,000 and an annual growth rate in the total income of 6.7 percent.

Transportation and communication. Extension of road transport has taken priority in development plans. A network of roads has been constructed to link the remotest parts of the state with railway stations, marketing centres, and industrial towns, thus opening the countryside to commerce and facilitating the interflow of farm products and consumer and capital goods. The total road length in 1968 was 5,776 miles, of which 4,959 miles was paved. More than 40,000 motor vehicles were in use, of which 2,700 were public-service vehicles and almost 8,000 goods vehicles. The state, which has been following a policy of gradual nationalization of passenger transport, in 1969 operated a fleet of more than 1,000 buses.

Long-distance transport is confined largely to the railways. Punjab is served by a state-operated railway, which has its headquarters in Delhi. Regular air passenger service is available between Delhi and some of the Punjab towns, such as Amritsar, Ludhiāna, and Patiāla. Delhi and Chandigarh are similarly connected. Water transport is practically nonexistent. Like the railways, the posts and telegraph service and broadcasting service are owned and controlled by the Union government.

Cultural life and tradition. Folklore, ballads of love and war, fairs and festivals, dancing, music, and Punjabi literature are characteristic expressions of the state's cultural life. Old romances, such as *Heer Riinjhii, Sassi Punnūn*, and *Mirzii Ṣāḥibān*, have been told in verse many times over, and the names of their authors—especially Warris and Hasham, both Muslims—are household names in Punjab. Older than these folk romances is the mystical and religious verse of the 13th-century Muslim Ṣūfī (mystic) Shaikh Farīd and of the 15th–16th-century Gurū (spiritual teacher) Nānak, founder of the Sikh faith, who for the first time used Punjabi extensively as a medium of poetic expression. These are the origins of Punjabi literature, which, at the beginning of the 20th century, entered its modem phase with the writings of the poet and author Bhai Vir Singh, and the poets Puran Singh and Dhani Ram Chatrik.

The whole populace joins in religious and seasonal festivals, such as Dussehra, Dīwālī, Baisākhi, and in the anniversary celebrations in honour of gurus and saints. Secular dances were introduced by immigrants after partition. The *bhangra, jhumar*, and *sammi* are the popular forms. The *giddha*, a native Punjabi form, is a women's hilarious song-dance. In addition to Sikh religious music, semi-classical Mughal forms, such as the *kāfi, khayāl, ṭhumrī, ghazal*, and *qawwālī*, continue to be popular.

The outstanding architectural monument is the Golden Temple at Amritsar, blending Indian and Saracenic styles. Its chief motifs, such as the dome and the geometrical design, are repeated in most of the Sikh gurdwaras, or shrines. Mural paintings in some of the gurdwaras preserve specimens of Sikh art. The Golden Temple itself is rich in gold filigree work of the most delicate kind and in panels with floral designs and marble facings inlaid with coloured stones. Among other important buildings may be mentioned the Hindu Temple of Durgiāna (also at Amritsar), a Moorish mosque at Kapiirthala, and old forts at Bhatinda and Bahādurgarh.

A Punjabi painter who made a great impact on modern art in the first half of the 20th century was Amrita Sher-Gil (1913–41). Some of the leading present-day Punjabi painters are Satish Gujral, Sobha Singh, S.G. Thakur Singh, and Kirpal Singh.

BIBLIOGRAPHY

History: R.C. MAJUMDAR, A.D. PUSALKER, and A.K. MAJUMDAR (eds.), *The History and Culture of the Indian People*, 11 vol. (1951–69), contains chapters on the history and culture of Punjab; S.M. LATIF, *History of the Punjab* (1891, reprinted 1964), the first comprehensive account in English by an Indian; INDUBHUSAN BANERJEE, *Evolution of the Khalsa*, 2nd ed., 2 vol. (1963), a modern history of the Sikhs, and thus of Punjab, by an Indian historian using original Punjabi sources; KHUSHWANT SINGH, *A History of the Sikhs*, 2 vol. (1963–66), a popular history by a Sikh scholar; HARBANS SINGH, *The Heritage of the Sikhs* (1964), an account of the Sikhs from 1469 to 1947.

Physical geography: L. DUDLEY STAMP, *India, Pakistan, Ceylon and Burma* (1957), with a section on natural regions; R.L. ANAND, B.S. OJHA, and GURDEV SINGH GOSAL, *Punjab Census Atlas* (1966), contains a cartographic portrayal of the relief, demography, and economy of Punjab; GURDEV SINGH GOSAL and GOPAL KRISHNA, "Upper Bari Doab," in R.L. SINGH (ed.), *India: Regional Studies* (1968), deals with physical, social, and economic phenomena in a specific tract—Upper Bari Doab; P.D. STRACEY, *Wild Life in India: Its Conservation and Control* (1963), includes charts on regional distribution.

Administration: Report of the Punjab Administrative Reforms Commission (1966); EDWARD NIRMAL MANGAT RAI, *Civil Administration in the Punjab: An Analysis of State Government in India* (1963).

Cultural life and traditions: W.G. ARCHER, *Paintings of the Sikhs* (HMSO, 1966), the only work to date on the paintings of the Sikhs by a Western art critic; RICHARD C. TEMPLE, *The Legends of the Punjab*, 3 vol. (1884–1900, reprinted 1962–63), a well-known work on the folklore of the Punjab, consisting of original texts in Roman script with English translation; MOHAN SINGH, *A History of Panjabi Literature, 1100–1932*, 2nd ed. (1956), the first history of Punjabi literature in English by a Punjabi scholar.

(H.K.M.S.)

Punjab (Pakistan)

The province of the Punjab (Pakistan) lies between the Himalayan foothills and the Indian state of Rājasthān, comprising an area of 79,704 square miles (206,432 square kilometres). Punjab means "five waters" or "five rivers" and signifies the land drained by the Jhelum, Chenāb, Rāvī, Beās, and Sutlej.

The land was once unfavourable for settlement, but its character changed after the building of an extensive irrigation network at the beginning of the 20th century. The area of settlement, which had formerly been limited to the north and northeast, was enlarged to include the whole province. The irrigated areas are laid out in rectangular blocks, or *chaks*. The more recently established villages and towns have a gridiron pattern, with an open square in the centre, contrasting with the older settlements of the province.

History. Archaeological excavations indicate that an urban civilization existed in this area from about 2500 BC to 1500 BC, when, it is believed, Aryan incursions brought it to an end. Of the following 1,000 years little is known. The early recorded history of the region begins with the annexation of Punjab and Sind to the Persian Empire by Darius I (*c.* 518 BC). Alexander descended on the Punjab in the spring of 326 BC to establish his transient rule; the Greek withdrawal was completed by about 317 BC. Candragupta incorporated the states of the Punjab into his Indian empire, which leached its zenith in the reign of his grandson Aśoka (ruled *c.* 265–238 BC). The Greeks of Bactria (northern Afghanistan) extended their rule to parts of the Punjab in the last decade of the 2nd century BC.

The 1st century BC and the first two centuries AD were marked by political chaos resulting from the incursions of the Sakas, the Parthians, and the Kushans. In the first half of the 3rd century AD, the Kushans yielded to the Sāsānians. Punjab formed a part of the Gupta domain, which was established in the middle of the 4th century and which in turn was shattered by the invasions of the Hephtalites in the third quarter of the 5th century. During a long period of anarchy, the Punjab changed hands between Kashmiri, Kābulī, and Hindu Shāhīs rulers.

Period of Mughal rule

The first Muslims to penetrate into northern India were the Arabs, who in 712 conquered the lower Punjab. The rest of the Punjab was conquered (1007–27) by Maḥmūd of Ghazna. From 1027 until the victories of Muʿizz-ud-Din Muhammad of Ghūr between 1176 and 1193, this part of the Indian subcontinent remained fragmented. In 1206 Punjab came under the Sultanate of Delhi. It was then successively ruled by the Khaljis (1290–1320), the Tughluqs (1320–98), the Sayyids (1414–50), and the Lodis (1451–1526). The Mughals made their entry with the victory of Bābur at Pānīpat on April 21, 1526. Under the Mughals the province enjoyed peace and prosperity for more than 200 years. The Mughals, who had strong artistic and cultural traditions, also made wide social reforms. Their power declined after 1738, and in 1747 Lahore fell to Afghan troops. The Afghan hold remained quite weak, giving rise to lawlessness and disorder. The religious sect called the Sikhs rose to power in the latter part of the 18th century.

The Punjab came under British occupation in 1849, after the British victory over Sikhs in the battles of Chiliānwāla and Gujrāt. After the British annexation of the territory, the Punjab was incorporated into a province that included areas northwest of the Jumna River extending to the Indo-Afghan border. There were later territorial adjustments. The Northwest Frontier Province was separated from the Punjab in 1901, as was the Delhi enclave in 1902. The province of Punjab was given autonomy, together with other provinces, in accordance with the Government of India Act of 1935.

When the Indian subcontinent received its independence in 1947, the British Indian province of Punjab was divided into West and East Punjab, which later became known, respectively, as Punjab (Pakistan) and Punjab (India). The boundary was drawn in such a way as to achieve, among other things, contiguous Muslim and non-Muslim majority areas; thus it lacked any physical or geographical basis. Preservation of some irrigation and communication systems was attempted, at a loss to Punjab (Pakistan) of areas in which there were Muslim majorities. Punjab (Pakistan) was part of the single province of West Pakistan from 1955 to 1970, when it was reconstituted as a separate province; it also included the former princely state of Bahāwalpur.

The natural environment. The province of Punjab lies on an alluvial plain formed by the Indus system of rivers. To the north are the hills of Murree and Rāwalpindi and the Pabbi hills of Gujrāt, forming part of the sub-Himalayas. The highest of these hills, Murree, has an altitude of 7,445 feet (2,269 metres). Potwar Plateau, in the far north, is a maze of uplands and small, alluvial, loessial flats, ranging in height from 1,000 to 2,000 feet above sea level. It is drained by the Haro and Soān rivers.

The plain of the Punjab has a compound slope. The general tilt of the land is from northeast to southwest, with an average gradient of one foot to the mile, but it rises in the areas between rivers. The plain has a diversity of landforms: an active floodplain, which is inundated by a river almost every rainy season and contains changing river channels; a meander floodplain, adjacent to the active floodplain and containing old channels; a covered floodplain, with deposits resulting from sheetflooding; a scalloped interfluve highland area between rivers comprised of older alluvium and with practically no relief. The deserts are studded with sand dunes.

Climatic variation

Punjab lies on the margin of the monsoon climate. The temperature is generally hot, with marked variations between summer and winter. The hottest month is June; the coldest is January. The average maximum and minimum temperatures of the hill station of Murree for June are 81° F (27" C) and 60" F (16" C), respectively; those for January are 45" F (7° C) and 31° F (—0.7" C). In the plain the mean June temperature is 97° F (36" C) at Multan and 93" F (34" C) at Lahore. The mean January temperature of those stations is, respectively, 56° F (13° C) and 54" F (12" C). Average annual rainfall is low, except in the sub-Himalayan and northern areas. It decreases markedly from north to south or southwest: 32 inches at Siālkot, 23 inches at Lahore, and seven inches at Multān.

The Murree hills carry subtropical and temperate forests. In the plain the natural vegetation consists of tough, wiry grass or dry stunted bushes, with few large trees. About 6.4 percent of the uncultivated area of the province is under forests. Wild animal life is scanty.

Population. Punjab is the most populous province of Pakistan, having in the early 1970s an estimated population of 37,000,000. According to the 1961 census, it had a population of 25,619,000, which constituted about 60 percent of the population of what was then West Pakistan. There were 875 females for every 1,000 males. While the density of population in Pakistan is 138 persons per square mile, that of the Punjab is 322. The density ranges from 724 persons per square mile in Lahore Division to 147 in Bahāwalpur Division. The urban population constitutes 21 percent of the total. There are seven cities: Lahore (population over 1,900,000); Lyallpur (425,000); Multān (358,000); Rāwalpindi (340,000); Gujrānwāla (196,000); Siālkot (164,000); and Sargodha (129,000). There is considerable rural–urban migration, particularly to the larger centres.

In religion, the province is almost entirely Muslim. About two percent of the population are Christians, and only 0.1 percent practice other religions.

Punjabi is the mother tongue of 90 percent of the population in all districts. The main written language is Urdu, followed by English. The dominant ethnic groups, which have inhabited the Punjab throughout recorded history, are Jats, Rājputs, Arain, Gūjars, and Awan. The caste system is gradually becoming blurred as a result of increasing social mobility, intercaste marriages, and changing opinion.

Administrationand social conditions. The provincial capital is Lahore. The chief executive is a governor, who is appointed by the president of Pakistan. The province prepares its own budget, based on provincial receipts and central grants. The governor is assisted by the provincial secretariat, headed by a chief secretary. Civil servants are recruited on the basis of merit.

The judicial function is vested in the High Court of the province. Its decisions can be appealed to the Supreme Court of Pakistan.

The province is divided for administrative purposes into five divisions, 19 districts, and 73 tahsils. A division is headed by a commissioner, a district by a deputy commissioner, and a *tahsil* by a *tahsildār*.

Literacy and education

The literacy rate in the Punjab in 1961 was about 16 percent. The percentage of females who were literate was much lower than that of males. Since 1961 the number of educational institutions has increased considerably, particularly at primary, middle, and secondary levels. A number of vocational and commercial institutes and professional colleges have been added. The total number of educational institutions in 1969–70 was 26,261 with an enrollment of 3,200,000.

Health facilities, though undergoing improvement, are still inadequate. There were only four hospital beds per 10,000 of the population in 1971. It was estimated that there were 14 doctors and seven nurses per 100,000 of the population. There was an acute housing shortage in urban centres. The number of persons per habitable room was 3.2 in Lahore, and similar figures obtained in other cities.

The economy. It was estimated in 1968–69 that agriculture accounted for about 38 percent of the Punjab's gross provincial product, manufacturing for 17 percent, wholesale and retail trade for 14 percent, services for 11 percent, and transport and communications for about 5 percent.

Agriculture. The Punjab's chief crops are wheat and cotton. Other crops include rice, grain, sugarcane, millet, corn (maize), oilseeds, pulses, fruits, and vegetables. About three-quarters of the cultivated land is irrigated. Livestock and poultry products make up about one-quarter of the output of the agricultural sector.

Manufacturing. The Punjab is one of the more industrialized regions of Pakistan. The more important industries are textiles, machinery, electrical appliances, surgical instruments, metal industries, bicycles and rickshas, floor-covering, and food industries.

Transportation. Road transport has been glowing at a rapid rate. The total number of motor vehicles in 1968–69 was 69,359, of which 10,998 were trucks and 5,481 were buses. In 1970, the province had 5,923 miles of surfaced roads. The growth of road transport was at the expense of the railroads, which increased their traffic very little in the 1960s.

Cultural life. Martial traditions, rural romanticism, and religion form the basis of Punjabi culture. These are reflected in Punjabi literature, particularly in the folklore. The oft recited folk romances *Heer Rānjhā*, *Sohni Māhīwal,* and *Mirzd Ṣāḥibān* form the basis of Punjabi mystic poetry.

Marriages are generally arranged by parents. While marriage is a social contract between the bride and groom, it is also considered a tie between families and a source of prestige. Punjabi parents, therefore, attach high importance to making good marriages for their children. The custom of dowry is also important, particularly in the opulent sections of the urban society.

The practice of *vartan bhanji,* the exchange of gifts, favours, and services, is widespread on ceremonial occasions. A woman's rights of *vartan bhanji,* or visits to her parental home for receiving gifts, are almost continuous. The system is essentially a device for distributing gifts and money at times when they are most needed.

Dress commonly consists of a long shirt and *shalwār*, trousers or *lungī* (unstitched cloth tied around the waist in place of trousers), and a turban. Women generally wear a veil.

BIBLIOGRAPHY. General works on the Indo-Pakistan subcontinent and the undivided British province of the Punjab include: *Gazetteer of the Punjab*, "Provincial Series" (1889); S.M. LATIF, *History of the Punjab* (1891, reprinted 1964); H.K. TREVASKIS, *The Land of Five Rivers* (1928) and *The Punjab of Today: An Economic Survey of the Punjab in Recent Years (1890–1925)*, 2 vol. (1931–32); D.N. WADIA, *Geology of India,* 3rd ed. rev. (1961); M.L. DARLING, *The Punjab Peasant in Prosperity and Debt,* 2nd ed. (1928); K. DAVIS, *The Population of Zndia and Pakistan* (1951); and O.H.K. SPATE and A.T.A. LEARMONTH, *India and Pakistan,* 3rd ed. rev. (1967). Some of the more important publications on Pakistan and West Pakistan, containing useful information relating to the Punjab, are: the *West Pakistan Year Books; Twenty Years of Pakistan* (1967); R.R. PLATT (ed.), *Pakistan: A Compendium* (1961); the *Five Year Plans of Pakistan;* and the *Food and Agriculture Commissions Report* (1960). Z.S. ELGAR, *A Punjabi Village in Pakistan* (1960), is a study in the cultural life of a village community of the Punjab. The periodical, *Pakistan Geographical Review,* contains useful articles on the subject.

Puppetry

A puppet is an inanimate object moved by human agency in some kind of theatrical show, and the puppet theatre includes any kind of theatrical show that is presented through the medium of puppets. These definitions are wide enough to include an enormous variety of shows and an enormous variety of puppet types, but they do exclude certain related activities and figures. A doll, for instance, is not a puppet, and a girl playing with her doll as if it were a living baby is not giving a puppet show; but, if before an audience of her mother and father she makes the doll walk along the top of a table and act the part of a baby, she is then presenting a primitive puppet show. Similarly, automaton figures moved by clockwork that appear when a clock strikes are not puppets, and such elaborate displays of automatons as those that perform at the cathedral clock in Strasbourg or the town hall clock in Munich must be excluded from consideration, but a dancer gyrating by clockwork that is introduced by the hand of a manipulator into a show *is* a puppet.

Puppet shows seem to have existed in almost all civilizations and in almost all periods. In Europe, written records of them go back to the 5th century BC (*e.g.,* the *Symposium* of the Greek historian Xenophon). Written records in other civilizations are less ancient, but in China, in India, in Java, and elsewhere in Asia there are ancient traditions of puppet theatre, the origins of which cannot now be determined. Among the American Indians, there are traditions of puppet-like figures used in ritual magic. In Africa, records of puppets are meagre, but the mask is an important feature in almost all African magical ceremonies, and the dividing line between the puppet and the masked actor, as will be seen, is not always easily drawn. It may certainly be said that puppet theatre has everywhere antedated written drama and, indeed, writing of any kind. It represents one of the most primitive instincts of the human race.

CHARACTER OF PUPPET THEATRE

It may well be asked why such an artificial and often complicated form of dramatic art should possess a universal appeal. The claim has, indeed, been made that puppet theatre is the most ancient form of theatre, the origin of the drama itself. Claims of this nature cannot be substantiated, nor can they be refuted; it is improbable that all human dramatic forms were directly inspired by puppets, but it seems certain that from a very early period in man's development puppet theatre and human theatre grew side by side, each perhaps influencing the other. Both find their origins in sympathetic magic, in fertility rituals, in the human instinct to act out that which one wishes to take place in reality. As it has developed, these magical origins of the puppet theatre have been forgotten, to be replaced by a mere childlike sense of wonder or by more sophisticated theories of art and drama, but the appeal of the puppet even for modern audiences lies nearer a primitive sense of magic than most spectators realize.

Appeal of puppetry

Granted the common origin of human and puppet theatre, one may still wonder about the particular features of puppet theatre that have given it its special appeal and that have ensured its survival over so many centuries. It is not, for instance, simpler to perform than human theatre; it is more complicated, less direct, and more expensive in time and labour to create. Once a show has been created, however, it can provide the advantage of economy in personnel and of portability; one man can carry a whole theatre (of certain types of puppet) on his back, and a cast of puppet actors will survive almost indefinitely. These are clear advantages, but it would be a mistake to imagine that they can explain the whole popularity of puppet theatre. They do not apply to every kind of puppet—some puppets need two or even three manipulators for each figure, and many puppets need one manipulator for each figure. The company employed by a major puppet theatre, whether it be a traditional puppet theatre from Japan or a modern one from eastern Europe, will not be fewer than for an equivalent human theatre. The appeal of the puppet must be sought at a deeper level.

The essence of a puppet is its impersonality. It is a type rather than a person. It shares this characteristic with masked actors or with actors whose make-up is so heavy that it constitutes a mask. Thus, the puppets have an affinity with the stock characters of ancient Greek and Roman drama, with the masked characters of the Renaissance commedia dell'arte (*q.v.*), with the circus clown, with the ballerina, with the mummers, and with the witch doctor and the priest.

In an impersonal theatre, where the projection of an actor's personality is lacking, the essential rapport between the player and his audience must be established by other means. The audience must work harder. The spectators must no longer be mere spectators; they must bring

An English Punch and Judy show. Detail from "Punch or May Day," oil painting by Benjamin Robert Haydon, 1829. In the Tate Gallery, London.

By courtesy of the trustees of the Tate Gallery, London: photograph, A.C. Cooper Ltd.

their sympathetic imagination to bear and project upon the impersonal mask of the player the emotions of the drama. Spectators at a puppet show will often swear that they saw the expression of a puppet change. They saw nothing of the kind; but they were so wrapped up in the passion of the piece that their imaginations lent to the puppets their own fears and laughter and tears. The union between the actor and the audience is the very heart and soul of the theatre, and this union is possible in a very special way, indeed in a specially heightened way, when the actor is a puppet.

The unreality and universality of puppets

The impersonality of the puppet carries other characteristics. There is the sense of unreality. In the traditional English Punch and Judy puppet shows, for instance, no one minds when Punch throws the Baby out of the window or beats Judy till she is dead; everyone knows that it is not real and laughs at things that would horrify him if they were enacted by human actors. Psychologists agree that the effect is cathartic—one's innate aggressive instincts are released through the medium of these little inanimate figures.

The puppet also carries a sense of universality. This, too, springs from its impersonality. A puppet Charlemagne in a Sicilian puppet theatre is not merely an 8th-century Frankish king but a symbol of royal nobility; and the leader of his rear guard dying on the pass of Roncesvalles is not merely a petty knight ambushed in a skirmish but a type representing heroism and chivalry. Similarly, in the Javanese puppet theatre, a grotesque giant is a personification of the destructive principle, while an elegantly elongated local deity is a personification of the constructive principle. Here the puppet theatre reveals its close relationship with the whole spirit of folklore and legend.

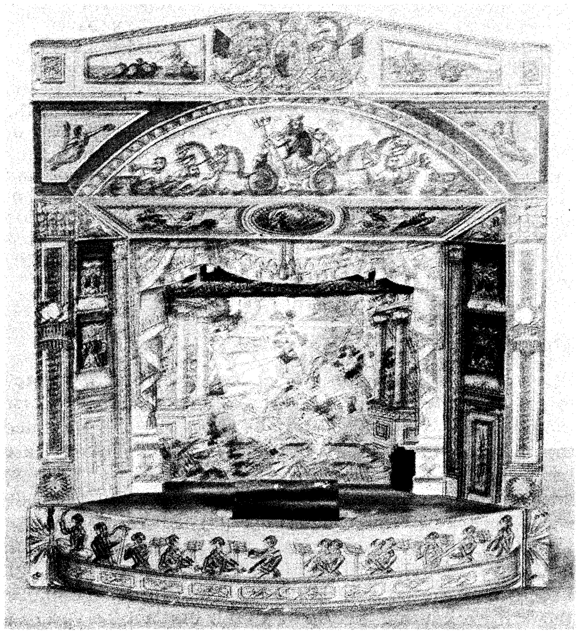
The puppet achieves its elemental qualities of impersonality, unreality, and universality through the stylizations imposed upon it by its own limitations. It is a mistake to imagine that the more lifelike or natural a puppet can be, the more effective it is. Indeed, the opposite is often the case. A puppet that merely imitates nature inevitably fails to equal nature; the puppet only justifies itself when it adds something to nature—by selection, by elimination, or by caricature. Some of the most effective puppets are the crudest: at Liège, Belgium, for instance, there is a tradition of puppets whose arm and leg movements are not controlled but purely accidental. The Rajasthani puppets of India have no legs at all. Even less naturalistic and even more stylized are the hunchbacked grotesques of the European tradition, the birdlike profiles of the Indonesian shadow figures, and the intricately shaped leather cutouts of Thailand, but it is precisely among these most highly stylized types of puppets that the art reaches its highest manifestations.

While admiring these puppets that exist furthest from nature, it cannot be denied that there is a charm and a fascination in the miniaturization of life. Much of the appeal of the puppet theatre has come from the spectators' delight in watching a world in miniature. This can be appreciated best of all in a toy theatre, in which a tiny stage on a drawing room table can be filled with choruses of peasants, troops of banditti, or armies locked in combat, while the scenery behind them depicts far vistas of beetling cliffs or winding rivers. A toy theatre, more than any other type of puppet theatre, is the human theatre on a miniature scale, but even here there is something more than a mere arithmetical reduction in scale; there is a something special that belongs to the world of the puppet.

Manipulation—seen and unseen

And to the appreciation, often instinctive, of these characteristics that mark the puppet theatre, there must be added admiration for the sheer human skill that has gone into the making and manipulation of the figures. The manipulator is usually unseen; his art lies in hiding his art, but the audience is aware of it, and this knowledge adds an element to the dramatic whole. In some kinds of presentation—for instance, in a type of cabaret floor show that became popular in the mid-20th century—the manipulator works in full view of the audience, who may, if they wish, study his methods of manipulation. This is a far cry from the philosophy of the traditional European puppet players of earlier generations, who guarded the secrets of their craft as if they were conjuring tricks. It is, indeed, fair to say that any presentation that deliberately draws attention to the mechanics of how it is done is distorting the art of puppetry, but the realization, nevertheless, of the expertise involved in a performance and some knowl-

By courtesy of Pollock's Toy Museum. London; photograph, A.C. Cooper Ltd.



An English toy theatre, 1850. In Pollock's Toy Museum, London.

edge of the technical means by which it is achieved do add an extra dimension to the appreciation of this difficult and highly skilled art.

TYPES OF PUPPET:

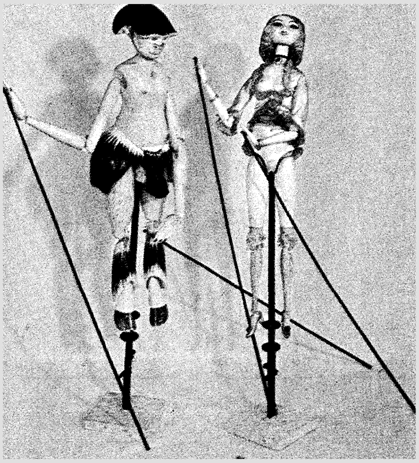
There are many different types of puppets. Each type has its own individual characteristics, and for each there are certain kinds of suitable dramatic material. Certain types have developed only under specific cultural or geographical conditions. The most important types may be classified as follows:

Hand or glove puppets. These have a hollow cloth body that fits over the manipulator's hand; his fingers fit into the head and the arms and give them motion. The figure is seen from the waist upward, and there are normally no legs. The head is usually of wood, papier mâché, or rubber material, the hands of wood or felt. The most common way to fit the puppet on the hand is for the first finger to go into the head, and the thumb and second finger to go into the arms. There are, however, many variants of this. In the puppets of Catalufia, Spain, for instance, the first three fingers fit into a wooden shouider piece, carved in one with the head, while the thumb and the little finger fit into extensions running to the arms. The usual "first-two-fingers-and-thumb" method is used for Punch-type figures; it allows the puppet to pick up and grasp small props very well and is obviously useful when wielding the stick that plays a big part in the show, but it tends to produce a lopsided effect, with one arm higher than the other. The Catalan method produces a larger and more impressive figure. The performer normally holds his hands above his head and stands in a narrow booth with an opening just above head height. Most of the traditional puppet folk heroes of Europe are hand puppets; the booth is fairly easily portable, and the entire show can be presented by one person. This is the typical kind of puppet show presented in the open air all over Europe and also found in China. But it need not be limited to one manipulator; large booths with three or four manipulators provide excellent scope for the use of these figures. The virtue of the hand puppet is its agility and quickness; the limitation is small size and ineffective arm gestures.

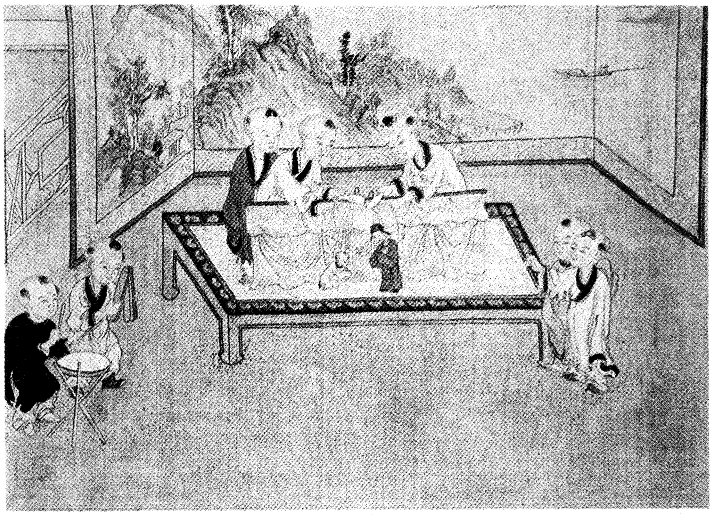
Rod puppets. These figures are also manipulated from below, but they are full-length, supported by a rod running inside the body to the head. Separate thin rods may move the hands and, if necessary, the legs. Figures of this type are traditional on the Indonesian islands of Java and Bali, where they are known as ***wayang golek***. In Europe, they were for a long time confined to the Rhineland; but in the early 20th century Richard Teschner in Vienna developed the artistic potentialities of this type of figure. In Moscow, Nina Efimova carried out similar experimental productions, and these may have inspired the Central State Puppet Theatre in Moscow, directed by Sergep Obraztsov, to develop this type of puppet during the 1930s. After World War II Obraztsov's theatre made many tours, especially in eastern Europe, and a number of puppet theatres using rod puppets were founded as a result. Today, the rod puppet is the usual type of figure in the large state-supported puppet theatres of the U.S.S.R., Poland, Bulgaria, Romania, Hungary, and East Germany. In a similar movement in the United States, largely inspired by Marjorie Batchelder, the use of rod puppets was greatly developed in school and college theatres, and the hand-rod puppet was found to be of particular value. In this figure the hand passes inside the puppet's body to grasp a short rod to the head, the arms being manipulated by rods in the usual way. One great advantage of this technique is that it permits bending of the body, the manipulators' wrist corresponding to the puppet's waist. Although, in general, the rod puppet is suitable for slow and dignified types of drama, its potentialities are many and of great variety. It is, however, extravagant in its demands on manipulators, requiring always one person, and sometimes two or three, for each figure on stage.

Prevalence in eastern Europe

By courtesy of the Puppentheatersammlung, Munich



"Faun" an3 [illegible] ymp[illegible], [illegible] d pupp[illegible] y Richard Teschner, 1914. In the Puppentheatersammlung, Munich.



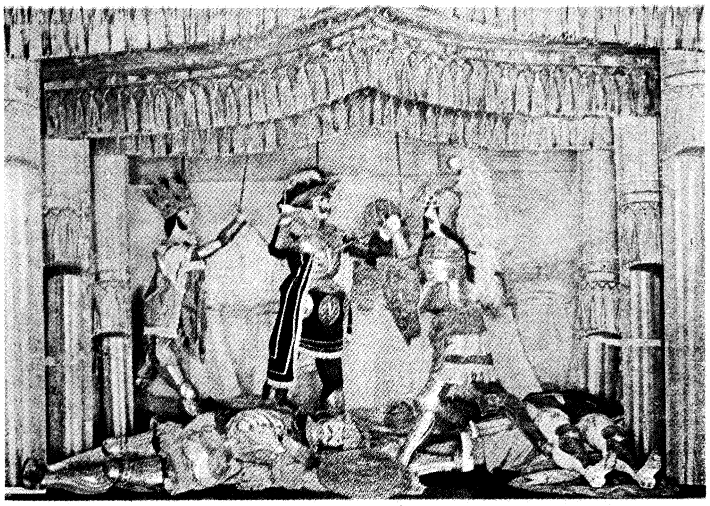
Chinese children playing with marionattes, detail from "The Hundred Children," a handscroll of the 17th century. In the British Museum.

By courtesy of the trustees of the British Museum

Marionettes or string puppets. These are full length figures controlled from above. Normally they are moved by strings or more often threads, leading from the limbs to a control or crutch held by the manipulator. Movement is imparted to a large extent by tilting or rocking the ontr[illegible] but individual strings are plucked when a decided movement is [illegible] d A simple marionette may have nine t[illegible] g[illegible] t each [illegible]g, one t[illegible] each [illegible] d, one t each shoulder, one to each ear [illegible] r [illegible] movements), and one to the base f the spine (for bowing); but special [illegible] t will q[illegible] sp[illegible] l strings th[illegible]t may double or tre[illegible] l[illegible] th[illegible] number. The manipulation of a many-stringed marionette i[illegible] a highly skilled [illegible] i[illegible] [illegible] l[illegible] are f two main [illegible] h[illegible] t l (or aeroplane) and vertical —and th[illegible] choice is largely a matter of personal preference.

The string marionette does [illegible]t seem t[illegible] have been fully developed until the nid 9th century, when the English marionettist Thomas Holden created a sensation with his ingenious figures and was followed by many [illegible] it t[illegible] Before [illegible]t time, the control of [illegible] ri[illegible]ett[illegible] [illegible] to h[illegible] b[illegible] b. a stout wire to the crown of the head, with subsidiary strings t[illegible] tl[illegible] hands and feet; even more primiti[illegible] methods of cc[illegible] may still be ob[illegible] in certain traditional folk theatres. In [illegible] ily [illegible] is an iron rod to the head, [illegible] t[illegible] rod to the sword arm, a[illegible]d a string t[illegible] the other arm; the legs hang free and a distinctive walking gait is imparted to the figures by a twisting and swinging of the main rod; in Antwerp, there are just rods to the head and to one arm; in Liège, there are no hand rods at

Primitive and developed forms



A scene from a 19th-century Sicilian puppet theatre enacting the battle of Roncesvalles. The Sicilian puppets were moved from above by both strings and rods. In the Puppentheatersammlung, Munich.

By courtesy of the Puppentheatersammlung, Munich

all, merely one rod to the head. Distinctive forms of marionette control are found in India: in Rajasthani a single string passes from one shoulder over the manipulator's hand and down to the other shoulder; in southern India there are marionettes whose weight is supported by strings attached to a ring on the manipulator's head, rods controlling the hands.

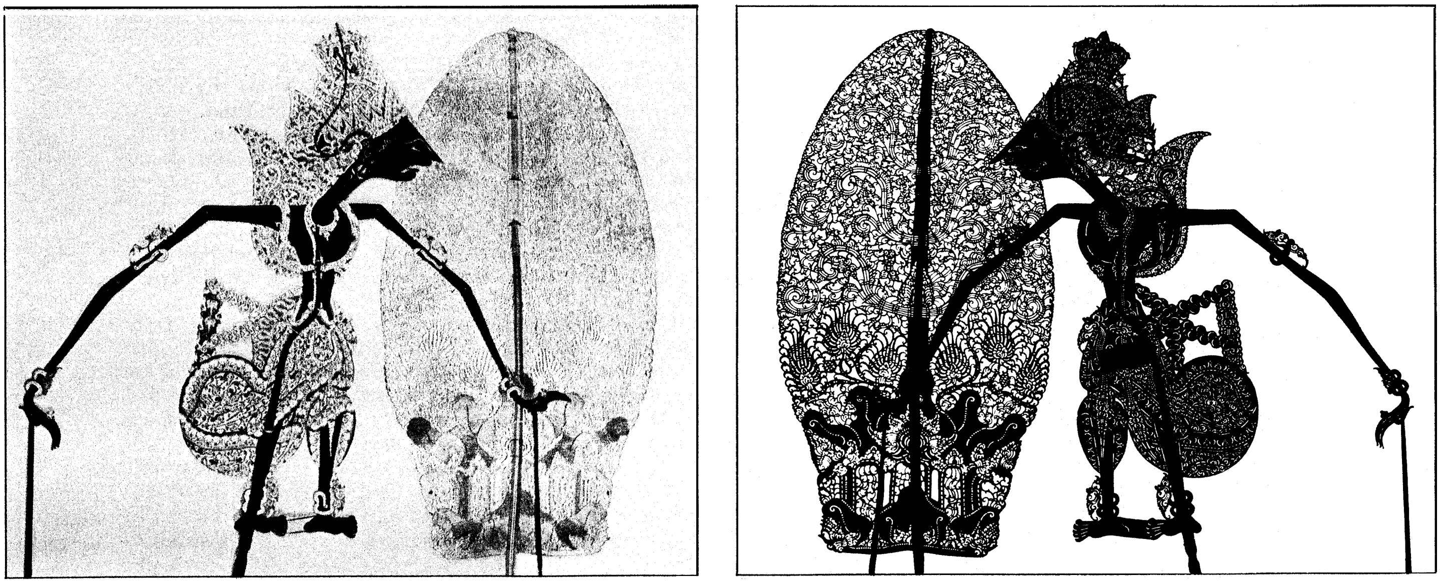
In European history the marionette represents the most advanced type of puppet; it is capable of imitating almost every human or animal gesture. By the early 20th century, however, there was a danger that it had achieved a sterile naturalism that allowed no further artistic development; some puppeteers found that the control of the figure through strings was too indirect and uncertain to give the firm dramatic effects that they required, and they turned to the rod puppet. But, in sensitive hands, the marionette remains the most delicate, if the most difficult, medium for the puppeteer's art.

Flat figures. Hitherto, all the types of puppets that have been considered have been three-dimensional rounded figures. But there is a whole family of two-dimensional flat figures. Flat figures, worked from above like marionettes, with hinged flaps that could be raised or lowered, were sometimes used for trick transformations; flat jointed figures, operated by piston-type arms attached to revolving wheels below, were used in displays that featured processions. But the greatest use of flat figures was in toy theatres. These seem to have originated in England by a printseller in about 1811 as a kind of theatrical souvenir; one bought engraved sheets of characters and scenery for popular plays of the time, mounted them and cut them out, and performed the play at home. The sheets were sold, in a phrase that has entered the language, for "a penny plain or twopence coloured," the colouring by hand in rapid, vivid strokes of the brush. During a period of about 50 years some 300 plays—all originally performed in the London theatres—were adapted and published for toy-theatre performance in what came to be called the "Juvenile Drama," and a hundred small printsellers were engaged in publishing the plays and the theatrical portraits for tinselling that often went with them. It was always a home activity, never a professional entertainment, and provided one of the most popular and creative fireside activities for Regency and Victorian families. Although few new plays were issued after the middle of the 19th century, a handful of publishers kept the old stock in print until the 20th century. After World War II this peculiarly English toy was revived. Toy theatres also flourished in other countries during the 19th century: Germany published many plays; Austria published some extremely impressive model-theatre scenery; in France toy-theatre sheets were issued; in Denmark a line of plays remains in print. The interest of these toy-theatre plays is largely social, as a form of domestic amusement, and theatrical, as a record of scenery, costume, and even dramatic gesture in a particular period of stage history. As genuine performances by flat figures, they represent one aspect of the puppet theatre.

Toy theatres

Shadow figures. These are a special type of flat figure, in which the shadow is seen through a translucent screen. They may be cut from leather or some other opaque material, as in the traditional theatres of Java, Bali, and-Thailand, in the so-called *ombres chinoises* (literally "Chinese shadows") of 18th-century Europe, and in the art theatres of 19th-century Paris; or they may be cut from coloured fish skins or some other translucent material, as in the traditional theatres of China, India, Turkey, and Greece, and in the recent work of several European theatres. They may be operated by rods from below, as in the Javanese theatres; by rods held at right angles to the screen, as in the Chinese and Greek theatres; or by threads concealed behind the figures, as in the *ombres chinoises* and in its successor that came to be known as the English galanty show. Shadow theatre need not be

By Courtesy Of the Puppentheatersammlung, Munich



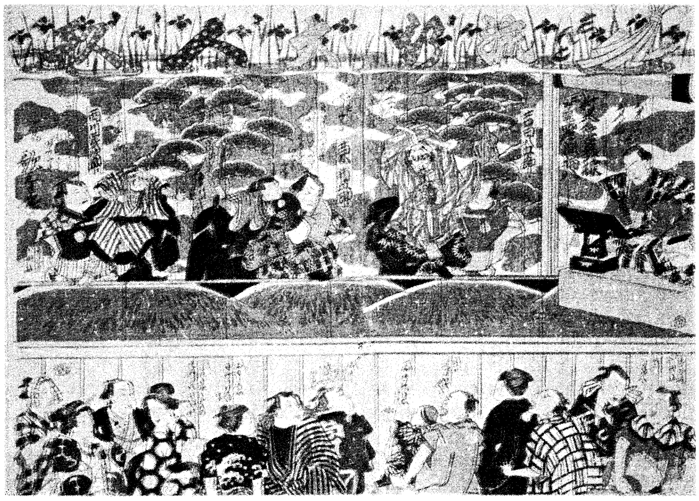
Indonesian wayang shadow puppet and decoration. (Left) Spectators may sit on the same side of the screen as the performer, watching the performance as presented by colourful rod puppets. (Right) Connoisseurs of the wayang art usually prefer to sit on the other side of the screen, viewing the performance as a shadow play.

limited to two-dimensional figures; rounded figures may be used effectively and the blurring of the sharp edge of a shadow that comes when the figure is not pressed sharply against the screen can provide an additional artistic element in the presentation. A particular type of shadow show that was conceived in terms of film is the silhouette films first made by the German film maker Lotte Reiniger in the 1920s; for these films, the screen was placed horizontally, like a table top, a light was placed beneath it, the camera was above it, looking downward, and the figures were moved by hand on the screen, being photographed by the stop-action technique. Her early films were in black and white, with figures made from thin sheet lead and card, but she made a few colour films in the 1950s. The shadow theatre is a medium of great delicacy, and the insubstantial character of shadow puppets exemplifies all the truest features of puppetry as an art form.

Bunraku puppets

Other types. These five types by no means exhaust every kind of figure or every method of manipulation. There are, for instance, the puppets carried by their manipulators in full view of the audience. The most interesting of these are the Japanese *bunraku* puppets, which are named for a Japanese puppet master, Uemura Bunrakuken, of the 18th century. These figures, which are about two-thirds life size, may be operated by as many as three manipulators: the chief manipulator controls with one hand head movements by means of strings inside the body, which may raise the eyebrows or swivel the eyes, while with his other hand he moves the right arm of the puppet; the second manipulator moves the left arm of the puppet; and the third moves the legs; the coordination of movement between these three artists requires long and devoted training. The magnificent costumes and stylized carving of the *bunraku* puppets establish them as among the most striking figures of their kind in the world.

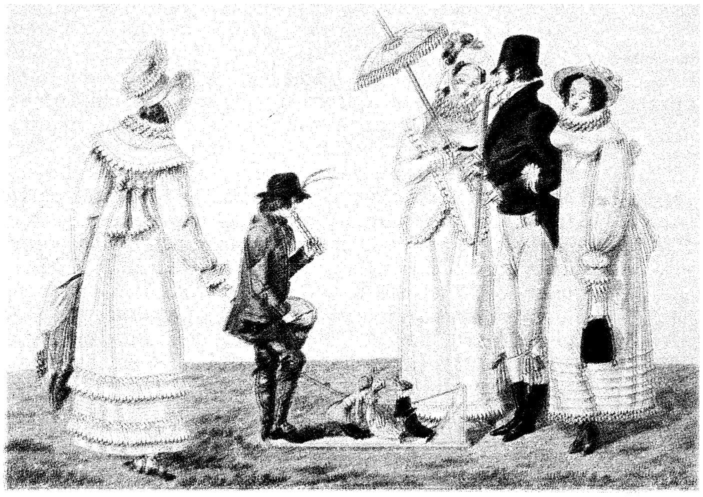
By courtesy of the Puppentheatersammlung, Munich



Japanese bunraku theatre; woodblock print by Utashige, 19th century. The puppeteers appear on stage with their puppets; the narrator is shown at the right.

Somewhat similar figures, though artistically altogether inferior, are the dummies used by ventriloquists; ventriloquism, as such, has no relation to puppetry, but the ventriloquists' figures, with their ingenious facial movements, are true puppets. The technique of the human actor carrying the puppet actor on to the stage and sometimes speaking for him is one that has been developed a great deal in some experimental puppet theatres in recent years. Sometimes the actor is invisible through the lighting technique of "black theatre," but sometimes he is fully visible. This represents a total rejection of much traditional thinking about the nature of puppetry, and it remains to be seen how widely the practice will become accepted.

Another minor form of puppet representation is provided by the jigging puppets, or *marionettes à la planchette,* that were, during the 18th and 19th centuries, frequently performed at street corners throughout Europe. These small figures were made to dance, more or less accidentally, by the slight variations in the tension of a thread passing through their chests horizontally from the performer's knee to an upright post. Similar were puppets held by short rods projecting from the figures' backs, which were made to dance by bouncing them on a springy board on the end of which the performer sat. The unrehearsed movements of figures like these, when loosely jointed, have a spontaneous vitality that more sophisticated puppets often miss. Another interesting, if elemental, type of puppet, the "scarecrow puppets," or *lileki,* of Slovenia, is constructed from two crossed sticks draped with old clothes; two of these figures are held up on either side of a bench draped with a cloth, under which the manipulator lies. The puppets talk with each other and with a human musician who always joins in the proceedings. The playlets usually end with a fight between the two puppets.



Marionnettes **a la** *planchette, or jigging puppets, being operated by a young puppeteer who provides his own accompaniment on his drum and whistle. "Les Petites Marionnettes," engraving from Le Bon Genre, published in France in* 1820.

Lalance—Ziolo

Still another minor puppet form is the finger puppet, in which the manipulator's two fingers constitute the limbs of a puppet, whose body is attached over the manipulator's hand.

The giant figures that process through the streets of

By courtesy of the Puppentheatersammlung, Munich

Amusement with a simple finger puppet; lithograph by an unknown artist, c. 1850.

Humans as puppets

some European towns in traditional festivities are puppets of a kind, though they do not normally enact any plays. The same applies to the dragons that are a feature of street processions in China and are to be found in some places in Europe — as, for example, at Tarascon, France. Indeed, when a man hides himself within any external frame or mask, the result may be called a puppet. Many of the puppet theatres in Poland today also present plays acted by actors in masks; the Bread and Puppet Theatre in the United States is another example of the same tendency. The divisions between human actors and puppet actors are becoming increasingly blurred; if, in the past, many puppets tried to look and act like humans, today many human actors are trying to look and act like puppets. Clearly, puppetry is being recognized not merely as a particular form of dramatic craft but as one manifestation of total theatre.

STYLES OF PUPPET THEATRE

Puppet theatre has been presented in many diverse styles and for many different kinds of audience. Throughout history, the chief of these has been the performance of folk or traditional plays to popular audiences. The most familiar examples are the puppet shows that have grown up around a number of national or regional comic heroes who appear in a whole repertory of little plays. Pulcinella, for example, was a human character in the Italian commedia dell'arte who began to appear on the puppet stages early in the 17th century; he was carried around Europe by Italian puppet showmen and everywhere became adopted as a new character, hunchbacked and hook-nosed, in the native puppet plays. In France he became Polichinelle, in England Punch, in Russia Petrushka, and so on. In England alone did this wide repertory of plays based on popular legend become limited to the one basic pattern of the Punch and Judy show. At about the time of the French Revolution, at the end of the 18th century, a great many local puppet heroes displaced the descendants of Pulcinella throughout Europe: in France it was Guignol, in Germany Kasperl, in the Netherlands Jan Klaassen, in Spain Christovita, and so on. All of these characters are glove puppets; most speak through a squeaker in the mouth of the performer which gives a piercing and unhuman timbre to their voices; and all indulge in the fights and other business typical of glove-puppet shows. It is a mistake, however, to regard them all as the same character; they are distinct national types. In Greece, the comic puppet hero is Karankiózis, a shadow puppet, who originally came from Turkey, where he is known as Karagoz.

Dramatic styles

The dramatic material in which these popular puppets play is sometimes biblical, sometimes based on folk tales, and sometimes from heroic sagas. A play on the Passion of Christ, for instance, is still presented by the Théâtre Toone in Brussels; the Faust legend has provided the classic theme for the German puppet theatre; the Temptation of St. Anthony for the French; and the poems of the Italian Renaissance poet Ariosto, handed on through many popular sources, provide the themes of crusading chivalry for the puppet theatres of Sicily and Liège. More specifically dramatic or literary sources were used by the travelling marionette theatres of England and the United States in the 19th century, when popular plays such as *East Lynne* and *Uncle Tom's Cabin* were played to village audiences almost everywhere.

In Asia, the same tradition of partly religious and partly legendary sources provides the repertory for the puppet theatres. Perhaps the most interesting of these is the *wayang purwa* of Java, from the Indian epics *Rāmāyaṇa* and *Mahābhārata,* in which a deposed deity (Semar) and his grotesque sons (Gareng and Petruk) provide the comic relief.

In distinction to these essentially popular shows, the puppet theatre has, at certain periods of history, provided a highly fashionable entertainment. In England, for instance, Punch's Theatre at Covent Garden, London, directed by Martin Powell from 1711 to 1713, was a popular attraction for high society and received many mentions in the letters and journalism of the day. In the 1770s to the 1790s, several Italian companies attracted fashionable audiences and the commendation of Dr. Samuel Johnson. In Italy, a magnificent puppet theatre was established in the Palazzo della Cancelleria in Rome in 1708, for which Alessandro Scarlatti, with other eminent composers, composed operas. In Austria-Hungary, Josef Haydn was the resident composer of operas for a puppet theatre erected by Prince Esterházy about 1770. In France, the *ombres chinoises* of François-Dominique Seraphin had been established at the Palais-Royal, in the heart of fashionable Paris, by 1781. The Italian scene designer Antonio Bibiena painted the scenery for a marionette theatre belonging to a young Bolognese prince, which performed in London in 1780. Exquisite Venetian marionette theatres preserved in the Bethnal Green Museum in London and the Cooper-Hewitt Museum in New York indicate the elegance of these fashionable puppet theatres of the 18th century.

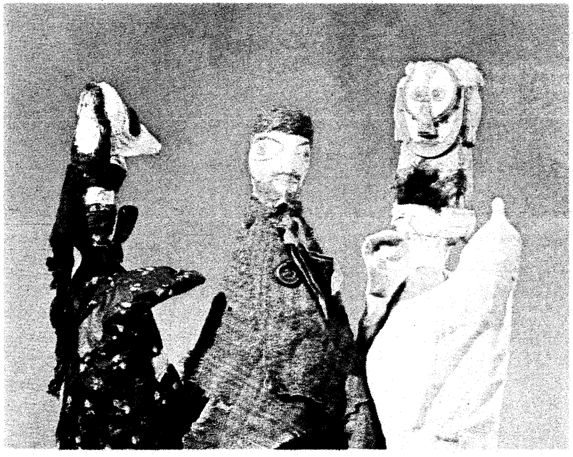
During the 18th century English writers began to turn to the puppet theatre as a medium, chiefly for satire. The novelist Henry Fielding presented a satirical puppet show, under the pseudonym of Madame de la Nash, in 1748. The caustic playwright and actor Samuel Foote used puppets to burlesque heroic tragedy in 1758 and sentimental comedy in 1773. In a similar vein, the dramatist Charles Dibdin presented a satirical puppet revue in 1775, and a group of Irish wits ran the Patagonian Theatre in London from 1776 to 1781 with a program of ballad operas and literary burlesques. In France there was a great vogue for the puppet theatre among literary men during the second half of the 19th century. This seems to have begun with the theatre created in 1847 at Nohant by George Sand and her son Maurice, who wrote the plays; well over a hundred plays were produced during a period of 30 years. These productions were purely for guests at the house; they are witty, graceful, and whimsical. Some years later another artistic dilettante conceived the idea of presenting a literary puppet show, but this time for the public; Louis Duranty opened his theatre in the Tuileries Gardens in Paris in 1861, but it lacked popular appeal and did not survive in its original form for very long. The next year Duranty's experiment inspired a group of literary and artistic friends to found the Theatron Erotikon, a tiny private puppet theatre, which only ran for two years, presenting seven plays to invited audiences. The moving spirit, however, was Lemercier de Neuville, who went on to create a personal puppet theatre that played in drawing rooms all over France until nearly the end of the century.

National preferences

All these literary puppet theatres in France had made use of hand puppets, while the English literary puppeteers of the previous century had used marionettes. In 1887 a French artist, Henri Rivière, created a shadow theatre that enjoyed considerable success for a decade at the Chat Noir café in Paris; Rivikre was joined by Caran d'Ache and other artists, and the delicacy of the silhouettes was matched by especially composed music and a spoken commentary. Another type of puppet was introduced to Paris in 1888 when Henri Signoret founded the Petit Théâtre; this theatre used rod puppets mounted on a base that ran on rails below the stage, the movement of the limbs being controlled by strings attached to pedals. The plays presented were pieces by classic authors—Cervantes, Aristophanes, Shakespeare — and new plays by French poets. The Petit Théâtre, like all the 19th-century French literary puppet theatres, performed infrequently to small audiences in a bohemian milieu; as a movement, this literary enthusiasm for the puppet theatre had little popular influence, but it served as a witness to the potential qualities of puppet theatre.

The puppet theatre in Japan entered literature with the plays of Chikamatsu (1653–1724). This writer, known as the Shakespeare of Japan, took the form of the existing crude Japanese puppet dramas and developed it into a great art form with over a hundred pieces, many of which remain in the repertory of the *bunraku* theatre today. In this form of theatre the text is chanted by a *jōruri* who is accompanied by a musician on a three-stringed instrument called a samisen.

By courtesy of Felix Klee, © Cosmopress, Geneva, and permission of S.P.A.D.E.M. 1971, by French Reproduction Rights. Inc.; photograph. Bil Baird Coliection



Hand puppets made by Paul Kiee (1879-1940); the centre puppet is a self-portrait. In the coilectlon of Felix Klee.

20th-century styles

In Europe, the art-puppet movement was continued into the 20th century by writers and artists associated with the Bauhaus, the highly influential German school of design, which advocated a "total" or "organic" theatre. One of its most illustrious teachers, the Swiss painter Paul Klee, created figures of great interest for a home puppet theatre, and others designed marionettes that reflected the ideas of Cubism. The eminent English man of the theatre Gordon Craig campaigned vigorously for the puppet as a medium for the thoughts of the artist. Between World Wars I and II and through the 1950s and 1960s, a number of artists endeavoured, in difficult economic conditions, to demonstrate that puppets could present entertainment of high artistic quality for adult audiences. The marionettes of the Art Puppet Theatre in Munich, for instance, were striking exemplars of the German tradition in deeply cut wood carving. A marionette theatre in Brunswick introduced a somewhat lighter and more satirical element. Both these theatres presented impressive productions of the traditional German puppet play of *Doktor Faustus.* In Austria, the Salzburg Marionette Theatre specializes in Mozart operas and has achieved a high degree of naturalism and technical expertise. In Czechoslovakia—a country with a fine puppet tradition—a marionette theatre presented musical turns interspersed with witty satirical sketches introducing the two characters who gave their names to the theatre: Hurvínek, a precocious boy, and Spejbl, his slow-witted father. In France, the prominent artists who designed for Les Comédiens de Bois included the painter Fernand Léger. Géza Blattner, a Hungarian living in Paris, created a theatre with ingenious, sophisticated rod puppets controlled by pedals and keys from below. Yves Joly stripped the art of the puppet to its bare essentials by performing hand puppet acts with his bare hands, without any puppets. The same effect was achieved by the Russian puppeteer Sergey Obraztsov, whose performance has a charm and a wit that are quite different from those of the great rod-puppet theatre that he directs. In England, the fine craftsman Waldo Lanchester played an important part in the marionette revival; his productions included the early madrigal opera *L'Amfiparnaso.* Jan Bussell, with the Hogarth Puppets, achieved an international reputation with his marionette ballets and light operas and also with his shadow-theatre productions designed by Lotte Reiniger.

In the United States, the artistic puppet revival was largely inspired by Ellen Van Volkenburg at the Chicago Little Theatre with productions that included *A Midsummer Night's Dream* in 1916. She later directed plays for Tony Sarg, who became the most important influence in American puppetry, with such large-scale marionette plays as *Rip Van Winkle, The Rose and the Ring,* and *Alice in Wonderland.* The Tatterman Marionettes, founded by William Duncan and Edward Mabley, followed the Sarg tradition with a wide repertory of plays, mostly commercially sponsored, and their touring production of *Peer Gynt* in 1937 represented an important artistic achievement. An important step in the development of puppetry in the United States was taken in 1967 with the opening of a permanent marionette theatre, the Bil Baird Theatre, in Greenwich Village, New York City; Bil Baird's marionettes were already well known from tours and television programs.

Meanwhile, the puppet theatre was continuing on a less exalted plane to demonstrate that it could still provide enjoyable entertainment for popular audiences. From the 1870s a number of English marionette companies had developed the technique of their art to an extraordinarily high level, and their influence was widely spread through Europe, Asia, and America by a series of world tours. Their performances made a great feature of trick effects: there was the dissecting skeleton, whose limbs came apart and then came together again; the Grand Turk, whose arms and legs dropped off to turn into a brood of children while his body turned into their mother; the crinolined lady, who turned into a balloon; the Scaramouch, with three heads; and a host of jugglers and acrobats. The last of the great touring marionette theatres in this tradition was the Teatro dei Piccoli of Vittorio Podrecca, which introduced the marionette pianist and the soprano with heaving bosom that have been widely copied ever since; this theatre closed down shortly after the founder's death in 1959.

During the 20th century there has been an increasing tendency to regard the puppet theatre as an entertainment for children. One of the first people to encourage this development was Count Franz Pocci, a Bavarian court official of the mid-19th century, who wrote a large number of children's plays for the traditional marionette theatre of Papa Schmid in Munich. Important also was Max Jacob, who developed the traditional folk repertory of the German Kasperltheater, between the 1920s and 1950s, into something more suited to modern ideas of what is suitable for children's entertainment. Almost all contemporary puppeteers have created programs for audiences of children.

The "interpreter

In this survey of the various styles of puppet theatre in different countries and in different cultures, there are certain features that are common to many otherwise differing forms. In many forms of puppet theatre, for instance, the dialogue is not conducted as if through the mouths of the puppets, but instead the story is recited or explained by a person who stands outside the theatre to serve as a link with the audience. This technique was certainly in use in England in Elizabethan times, when the "interpreter" of the puppets is frequently referred to; this character is well illustrated in Ben Jonson's *Bartholomew Fair,* in which one of the puppets leans out of the booth (they were hand puppets) and hits him on the head because it does not like the way he is telling the story. The same technique of the reciter outside the theatre is found in the Japanese *bunraku* theatre, in which the chanter contributes enormously to the full effect and is, indeed, regarded as the star of the company. The same technique is found in the French shadow theatre at the Chat Noir, and its imitators and successors, which depended to a great extent upon the chansonnier. Many recent puppet productions utilize the same technique. In traditional puppet theatres of Java, Greece, Sicily, and elsewhere, all the speaking is done by the manipulator. The plays consist of a mixture of narration and dialogue, and, though the performer will certainly vary his voice for the different characters, the whole inevitably acquires a certain unity that is one of the most precious attributes of the puppet theatre.

Music and lighting

Musical accompaniment is an important feature of many puppet shows. The gamelan gong and cymbal orchestra that accompanies a Javanese *wayang* performance is an essential part of the show; it establishes the mood, provides the cadence of the puppets' movements, and gives respite between major actions. Similarly, the Japanese samisen supports and complements the chanter. In the operatic puppet theatre of 18th-century Rome, the refined musical scores of Scarlatti and the stilted conventions and long-held gestures of the opera of that time must have been admirably matched by the slow, contrived but strangely impressive movements of the rod puppets. When, in 1662, Samuel Pepys visited the first theatre to present Punch in England, he noted in his famous diary that "here among the fiddlers I first saw a dulcimer played on with sticks knocking of the strings, and it is very pretty." Even an old-fashioned Punch and Judy show had a drum and panpipes as an overture. Puppets without music can seem rather bald. At one time the gramophone was used extensively by puppeteers, and more recently the tape recorder has provided a more adaptable means of accompanying a puppet performance with music and other sound effects.

Lighting effects can also play an important part in a puppet production. The flickering oil lamp of the Javanese *wayang* enhances the shadows of the figures on the screen; as long ago as 1781, the scene painter Philip James de Loutherbourg used a large model theatre called the Eidophusikon to demonstrate the range of lighting effects that could be achieved with lamps. Modern methods of dimming electric lamps have enabled directors of puppet productions to achieve astonishing and spectacular effects.

PUPPETRY IN THE CONTEMPORARY WORLD

Opportunities and problems

The puppet theatre in the contemporary world faces great difficulties and great opportunities. The audiences for the traditional folk theatres have almost disappeared. Punch and Judy on the English beaches and Guignol in the parks of Paris still draw a crowd, but the indoor theatres that once attracted humble audiences survive with difficulty, usually with the aid of a sympathetic town council or in the shelter of a local museum. Puppets are increasingly regarded as an entertainment only for children. They certainly do provide a kind of theatre to which children respond with great enthusiasm, and, in the general development of children's theatre, the puppet theatre has a part to play. Some puppeteers are happy to play only for children. But others are eager, today as in the past, to play also on an adult level; and for these, audiences are few. No professional puppet theatre can exist, in the West, on a purely adult repertory. Even those theatres that do play for children face great economic difficulties from the small size of audience to which puppets can play and from the modest admission charges that can be charged to children. If a few companies do continue to present performances of quality, this is a tribute to their dedication to their art.

There are some possible means of performance beyond the children's theatre. There are cabarets or night clubs, which provide an opportunity for slickly presented short turns but obviously no scope for serious drama. And there is television. At first sight, television would seem an ideal medium for puppetry, and many puppet shows have in fact appeared on it, but the great possibilities that it once seemed to offer have not been fully realized. Some puppets, indeed, have become national idols. In England, for instance, Muffin the Mule and his animal friends, manipulated by Ann Hogarth, appeared from 1946 on the top of a piano at which Annette Mills played and sang. Sooty is an endearing little animal whose adventures are guided by his friend and manipulator, Harry Corbett. In the United States, the Kuklapolitans were created by Burr Tillstrom in 1947; Kukla, a small boy, had a host of friends, including Ollie the Dragon, Beulah Witch, Fletcher Rabbit, and Madame Ooglepuss, who exchanged repartee with Fran, a human actress standing outside the booth. The show had enormous popularity for some ten years and presented a social message based upon friendliness and cooperation. The same spirit lies behind the puppets introduced into the more specifically educational "Sesame Street" show; some of the monsters, for instance, do all the wrong things, but, because they are such monsters, the moral is clearly that these things should *not* be done. It must be noted, however, that all these phenomenally successful television puppet characters appeared in association with human actors, whose personalities were an important constituent in the program. Pure puppet programs have not, in general, made quite the same impact. The truth would seem to be that a puppet theatre, more than any other kind of theatre, needs audience response to bring it alive, and this the television screen cannot provide.

Puppetry for television.
(Top) Fran Allison with Kukla and Ollie, two puppets created by Burr Tillstrom for the series Kukla, Fran and Ollie. (Bottom) Actors Loretta Long and Matt Robinson with two of Jim Henson's Muppets in a scene from the children's series Sesame Street.
By courtesy of (top) WTTW-TV, Chicago—Public Broadcasting Service. (bottom) Children's Television Workshop

State subsidies

The economic difficulties facing puppet companies in western Europe and the United States have been lifted in eastern Europe and China, where the state provides generous subsidies for puppet theatres. Whereas in the West a puppet theatre is lucky if it can afford to pay a company of five or six performers, it is not unusual for a puppet theatre in the East to employ 50 or 60 performers, artists, and technicians. There has been an extraordinary surge of interest in puppet theatre in eastern Europe since World War II, and while the state supports these theatres, there is very little sign of any direct political propaganda in their programs. The results of all this aid have often been impressive in the sheer weight of numbers and scenic effects, and the productions have often been experimental and imaginative. Mere size, however, does not necessarily guarantee artistic success, and some of the best of these theatres would seem to feel a lack of confidence in their medium by their restless searching for new methods of presentation through "black theatre," mask theatre, and other techniques.

Puppetry in schools

A great feature of education during the 20th century has been the introduction of puppet making into schools as a craft activity. The difficulties facing professional puppet theatre are entirely absent here, and a puppet performance can synthesize many of the arts and skills of a group of children in making, costuming, and manipulating puppets, in writing plays for them, and in acting them. When this activity was first introduced, undue importance was often placed upon the mere construction of figures according to certain set methods and upon the painstaking preparation of a showing, so that the creative release of the performance was long delayed and sometimes never reached. Today the tendency is to create puppets quickly from scrap materials or from natural objects and to perform them impromptu, without rehearsal, as a form of dramatic self-expression. It is from such activities that the therapeutic potentialities of puppets have been utilized by psychiatrists working with disturbed children.

The future of the puppet theatre will certainly be greatly influenced by the cross-fertilization between different traditions in puppetry that will result from puppeteers meeting each other and seeing each other's performances at international festivals of the puppet theatre, These festivals now take place almost every year and are usually sponsored by Unima, the Union Internationale des Marionnettes, an international society of puppeteers. This body was originally founded in 1929 and was reconstituted in 19.57; it has members in more than 50 countries and provides a common meeting ground for professional and amateur performers, critics, and enthusiasts. In the past the differing local traditions of puppet theatre have provided a rich variety for this minor but fascinating art. In the future the local traditions may need some form of self-conscious protection to preserve them, but an even richer variety will surely spring from the interplay of artistic experiment and achievement on an international level.

BIBLIOGRAPHY. A.R. PHILPOTT, *Dictionary of Puppetry* (1969), a brief but comprehensive guide to every aspect of the subject; *The Rosalynde Stearn Puppet Collection,* McGill University Library Special Collections IV (1961), contains a useful bibliography; CHARLES MAGNIN, *Histoire des marionnettes en Europe depuis l'antiquité jusqu'à nos jours*, rev. ed. (1862), the classic history, not yet superseded; ERNEST MAINDRON, *Marionnettes et guignols* (1900), valuable for chapters on East Asia and for carrying the French history up to the end of the 19th century; BIL BAIRD, *The Art of the Puppet* (1965), a magnificently illustrated general survey; MARGARETA NICULESCU (ed.), *The Puppet Theatre of the Modern World* (1967), an international presentation sponsored by the Union Internationale des Marionnettes; GEORGE SPEAIGHT, *The History of the English Puppet Theatre* (1955), includes European puppets up to the 17th century *(Punch and Judy: A History,* 1970, is largely abstracted from this work, but contains some additional material); and *The History of the English Toy Theatre* (1969; a revised edition of *Juvenile Drama,* 1946); PAUL MCPHARLIN, *The Puppet Theatre in America: A History,* ***1524–1948,*** rev. ed. (1969), with a supplement covering developmenrs since 1948 by MARJORIE BATCHELDER MCPHARLIN, including a select bibliography; OLIVE BLACKHAM, *Shadow Puppets* (1960), a description of these figures all over the world.

(G.St.)

Purcell, Henry

The most important English composer of his time, Henry Purcell composed music covering a wide field: the church, the stage, the court, and private entertainment. In all these branches of composition he showed an obvious admiration for the past combined with a willingness to learn from the present, particularly from his contemporaries in Italy. With alertness of mind went an individual inventiveness that marked him as the most original English composer of his time as well as one of the most original in Europe.

Birth and early life

Not very much is known of Purcell's life. He was probably born in the summer or autumn of 1659. His father was a gentleman of the Chapel Royal, in which musicians for the royal service were trained, and the son received his earliest education there as a chorister. When his voice broke in 1673, he was appointed assistant to John Hingston, keeper of the king's instruments, whom he succeeded in 1683. From 1674 to 1678 he tuned the organ at Westminster Abbey and was employed there in 1675–76 to copy organ parts of anthems. In 1677 he succeeded Mathew Locke as the composer for Charles II's string orchestra and in 1679 was appointed organist of Westminster Abbey in succession to the composer John Blow. A further appointment as one of the three organists of the Chapel Royal followed in 1682. He retained all his official posts through the reigns of James II and William III and Mary. He married in 1680 or 1681 and had at least six children, three of whom died in infancy. His son Edward was also a musician, as was Edward's son Edward Henry (died 1765). Purcell seems to have spent all his life in Westminster; he died there on November 21, 1695. His fatal illness prevented him from finishing the music for the operatic version of John Dryden and Sir Robert Howard's verse tragedy *The Indian Queen* (1664), which was completed by his brother Daniel (died 1717). Daniel Purcell had also been brought up as a chorister in the Chapel Royal and was organist of Magdalen College, Oxford, from 1688 to 1695. Before his brother's death, he was little known as a composer, but from 1695 to 1707 he was in considerable demand for music for stage productions in London until the advent of Italian opera brought his activities to an end.

Purcell. portrait by John Closterman (1656–1713). In the National Portrait Gallery, London.
By courtesy of the National Portrait Gallery. London

To later ages Purcell was best known as a songwriter because so many of his songs were printed in his lifetime and were reprinted again and again after his death. The first evidence of his mastery as a composer, however, is an instrumental work—a series of fantasias (or "fancies") for viols in three, four, five, six, and seven parts. The nine four-part fantasias all bear dates in the summer of 1680, and the others can hardly be later. Purcell was here reviving a form of music that was already out of date and doing it with the skill of a veteran. Probably about the same time he started to work on a more fashionable type of instrumental music—a series of sonatas for two violins, bass viol, and organ (or harpsichord). Twelve of these were published in 1683, with a dedication to Charles II, and a further nine, together with a chaconne for the same combination, were issued by his widow in 1697. The foreword to the 1683 set claimed that the composer had "faithfully endeavour'd a just imitation of the most fam'd Italian Masters"; but side by side with the Italianate manner there was a good deal that derived from the English chamber music tradition.

The instrumental movements are the most striking part of the earliest of Purcell's *Welcome Songs* for Charles II —a series of ceremonial odes that began to appear in 1680. Possibly he lacked experience in writing for voices, at any rate on the scale required for works of this kind; or else he had not yet achieved the art of cloaking insipid words in significant music. By 1683 he had acquired a surer touch, and from that time until 1694, when he wrote the last of his birthday odes for Queen Mary, he produced a series of compositions for the court in which the vitality of the music makes it easy to ignore the poverty of the words. The same qualities are apparent in the last of his odes for St. Cecilia's Day, written in 1692.

Compositions for the theatre

Purcell's genius as a composer for the stage was hampered by there being no public opera in London during his lifetime. Most of his theatre music consists simply of instrumental music and songs interpolated into spoken drama, though occasionally there were opportunities for more extended musical scenes. His contribution to the stage was in fact very modest until 1689, when he wrote *Dido and Aeneas* (libretto by Nahum Tate) for performance at a girls' school in Chelsea; this miniature opera achieves a high degree of dramatic intensity within a narrow framework. From that time until his death, he was constantly employed in writing music for the public theatres. These productions included some that gave scope for more than merely incidental music—notably music for *Dioclesian* (1690), adapted by Thomas Betterton from the tragedy *The Prophetess,* by John Fletcher and Philip Massinger; for *King Arthur* (1691), by John Dryden, designed from the first as an entertainment with music; and for *The Fairy Queen* (1692), an anonymous adaptation of Shakespeare's *Midsummer Night's Dream,* in which the texts set to music are all interpolations. In these works Purcell showed not only a lively sense of comedy but also a gift of passionate musical expression that is often more exalted than the words. The tendency to identify himself still more closely with the Italian style is very noticeable in the later dramatic works, which often demand considerable agility from the soloists.

Composer of anthems

Purcell's four-part fantasias, his first court ode, and his first music for the theatre, *Theodosius,* a play by Nathaniel Lee, all date from 1680. Some of his church music may be earlier than that, but it is not possible to assign definite dates. As far as is known, most of his anthems, whether for the full choir (full anthems) or with sections for soloists (verse anthems), were written between 1680 and 1685, the year of Charles II's death. The decline of the Chapel Royal during the reigns of James II and of William and Mary may have been responsible for the comparatively few works he produced during that period; or, alternatively, he may have been so busy with stage music and odes that he had little time or inclination for church music. The style of his full anthems, like that of the fantasias, shows a great respect for older traditions. His verse anthems, on the other hand, were obviously influenced, in the first instance, by his master at the Chapel Royal, Pelham Humfrey, who had acquired a knowledge of Continental styles when he was sent abroad to study in the mid-1660s. The most notable feature of these latter works is the use of expressive vocal declamation that is pathetic without being mawkish. The same characteristics appear in the sacred songs he wrote for private performance. Since composers for the Chapel Royal in Charles II's reign had the string orchestra at their disposal, Purcell took the opportunity to include overtures and ritornellos that are both dignified and lively. The most elaborate of all his compositions for the church are the anthem My *heart is inditing,* performed in Westminster Abbey at the coronation of James II in 1685, and the festal *Te Deum and Jubilate,* written for St. Cecilia's Day in 1694. Of these the anthem is the more impressive; the *Te Deum and Jubilate* suffers on the whole from a forced brilliance that seems to have faded with the passage of time.

Though the main period of Purcell's creative activity lasted for little more than the last 15 years of his life, he managed to crowd into it a very large number of compositions, including more than 100 secular songs and about 40 duets, apart from those that he contributed to plays. Many of the songs are quite substantial pieces, incorporating recitative and arias on the lines of the Italian solo cantata. A favourite device used widely by Purcell in his secular music, though rarely in his anthems, was the ground bass (a short melodic phrase repeated over and over again as a bass line, with varying music for the upper parts). This device can have an invigorating effect in lively pieces, while in laments, such as Dido's farewell, it can intensify the expression of grief. The chaconne in the second set of sonatas uses the same technique with impressive results. Works of this kind represent the composer at the height of his capacity. The numerous catches (rounds for three or more unaccompanied voices written as one melody with each singer taking up a part in turn), on the other hand, though accomplished enough are little more than an experienced musician's contribution to social merrymaking. Purcell seems to have abandoned instrumental chamber music after his early years. His keyboard music forms an even smaller part of his work: it consists of suites and shorter pieces for harpsichord and a handful of pieces for organ.

Works published during his lifetime

Apart from a large number of songs that appeared in vocal collections, very little of Purcell's music was published in his lifetime. The principal works were the *Sonatas of III Parts* (1683); *Welcome to all the pleasures,* an ode for St. Cecilia's Day, written in 1683 (published in 1684); and *Dioclesian,* composed in 1690 (1691). After his death his widow published a collection of his harpsichord pieces (1696), instrumental music for the theatre (1697), and the *Te Deum and Jubilate* (1697); and the publisher Henry Playford issued a two-volume collection of songs entitled *Orpheus Britannicus* (1698 and 1702), which went through three editions, last appearing at mid-18th century.

A few of Purcell's dramatic works, odes, and anthems were printed in the late 18th and early 19th centuries; but not until 1876, when the Purcell Society was founded, was a serious attempt made to issue all of Purcell's works. The first volume was published in 1878, the second in 1882. From 1889 to 1928 volumes appeared at intervals. Then the scheme was in abeyance until 1957, when a volume of miscellaneous odes and cantatas was published. It was finally completed in 32 volumes in 1965. Revision of earlier volumes proceeded simultaneously with the issue of later ones, beginning with a revised edition of *Dioclesian* in 1961.

MAJOR WORKS

STAGE WORKS: One opera, *Dido and Aeneas* (first performed 1689); semi-operas, *Dioclesiarz* (1690), *King Arthur* (1691), *The Fairy Queen* (adaptation of *A Midsummer Night's Dream,* 1692), *The Indian Queen* (1695), *The Tempest* (composed ? 1695); incidental music to 43 plays.

VOCAL MUSIC: 15 odes, including six for Queen Mary's birthday and four for St. Cecilia's Day; anthems, including the coronation anthem, *My heart is inditing* (composed 1685); *Te Deum and Jubilate* (1694); sacred songs, catches, duets, and secular songs.

INSTRUMENTAL MUSIC: *Chacony in G Minor;* 13 *Fantasias* (1680); two *In Nomines;* five *Pavans;* 12 *Sonatas of III Parts* (1683); 10 *Sonatas of IV Parts* (published 1697); violin sonata; three overtures.

KEYBOARD WORKS: Eight harpsichord suites (1696); *Musick's Handmaid,* part 2 only (published 1689); miscellaneous pieces and transcriptions.

BIBLIOGRAPHY. THURSTON DART, "Purcell's Chamber Music," *Proceedings of the Royal Musical Association,* 85: 81–93 (1958–59), a study of the chronology and style of these works; EDWARD J. DENT, *Foundations of English Opera* (1928), a standard work on 17th-century stage music in England and its antecedents; IMOGEN HOLST (ed.), *Henry Purcell: Essays on His Music* (1959), includes new information on *Dido and Aeneas;* R.E. MOORE, *Henry Purcell and the Restoration Theatre* (1961), a general study of Purcell's dramatic music and the conditions in which it was performed; J.A. WESTRUP, *Purcell,* 7th ed. (1973), a biography and detailed study of the works; and "Purcell's Parentage," *Music Review,* 25:100–103 (1964), discusses, with evidence, the question of whether Purcell was the son of the elder Henry Purcell or of Thomas Purcell; FRANKLIN B. ZIMMERMAN, *Henry Purcell, 1659–1695: An Arzalytical Catalogue of His Music* (1963), an indispensable work of reference, with a thematic *incipit* for each piece; *Henry Purcell, 1659–1695: His Life and Times* (1967), a detailed biography, with an account of Purcell's environment; and "Purcell's Family Circle Revisited and Revised," *Journal of the American Musicological Society,* 16:373–381 (1963), further details about Purcell's family, with particular reference to his parentage.

(J.A.W.)

Purification Rites and Customs

Purification rites and customs, based on concepts of purity and pollution, are found in all known cultures and religions, both ancient and modern, preliterate and sophisticated. Assuming a wide variety of types and forms, these rites and customs attempt to re-establish lost purity or to create a higher degree of purity in relation to the Sacred or Holy (the transcendent realm) or the sociocultural realm.

General concepts of purity and pollution. Every culture has an idea, in one form or another, that the inner essence of man can be either pure or defiled. This idea presupposes a general view of man in which his active or vitalizing forces, the energies that stimulate and regulate his optimum individual and social functioning, are distinguished from his body, on the one hand, and his mental or spiritual faculties, on the other. These energies are believed to be disturbed or "polluted" by certain contacts or experiences that have consequences for a person's entire system, including both the physical and the mental aspects. Furthermore, the natural elements, animals and plants, the supernatural, and even certain aspects of technology may be viewed as operating on similar energies of their own: they too may therefore be subject to the disturbing effects of pollution. Because lost purity can be re-established only by ritual and also because purity is often a precondition for the performance of rituals of many kinds, anthropologists refer to this general field of cultural phenomena as "ritual purity" and "ritual pollution."

The rituals for re-establishing lost purity, or for creating a higher degree of purity, take many different forms in the various contemporary and historical cultures for which information is available. Some purification rituals involve one or two simple gestures, such as washing the hands or body, changing the clothes, fumigating the person oı object with incense, reciting a prayer or an incantation, anointing the person or object with some ritually pure substance. Some involve ordeals, including blood-letting, vomiting, and beating, which have a purgative effect. Some work on the scapegoat principle, in which the impurities are ritually transferred onto an animal, or even in some cases (as among the ancient Greeks) onto another human being; the animal or human scapegoat is then run out of town and/or killed, or at least killed symbolically. Many purification rites are very complex and incorporate several different types of purifying actions (see also RITUAL).

Ritual purity and pollution are matters of general social concern because pollution, it is believed, may spread from one individual or object to other members of society. Each culture defines what is pure and impure—and the consequences of purity and pollution—differently from every other culture, although there is considerable cross-cultural overlapping on certain beliefs. Cultures also vary greatly in the extent to which purity and pollution are pervasive concerns: Hinduism, Judaism, and certain tribal groups such as the Lovedu of the African Transvaal or the Yurok of northern California in the United States seem highly pollution-conscious, whereas among other peoples pollution concerns are relatively isolated and occasional. Even within the so-called pollution-conscious cultures, attitudes toward the cultural regulations may vary considerably: the Yurok, on the one hand, are said to consider their purification rituals to be rather a nuisance, albeit necessary for the success of their economic endeavours; but Hindus, on the other hand, seem to incorporate and embrace more fully the many regulations and rituals concerning purity prescribed in their belief and social systems.

The transmission and symptoms of pollution

Pollution is most commonly transmitted by physical contact or proximity, although it may also spread by means of kinship ties or co-residence in an area in which pollution has occurred. Because purity and pollution are inner states (though there usually are outer or observable symptoms of pollution). the defiled man—or artifact, temple, or natural phenomenon—may at first show no outward features of his inner corruption. Eventually, however, the effects of pollution will make themselves known; the appearance of a symptom or disaster that is culturally defined as a consequence of pollution, for example, may be the first indication that a defiling contact has occurred. Common cross-cultural, human symptoms of pollution include: skin disease, physical deformity, insanity and feeblemindedness, sterility, and barrenness. Nature also may become barren as a result of pollution; but, on the other hand, the natural elements and magical or supernatural forces may run amok as a result of pollution.

In general, the vital energies of man, nature, or the supernatural, as a consequence of pollution, may become either hypoactive or hyperactive. The vital energies may tend to operate in a manner that leads toward decline, loss of potencies and fertility, and death; they may also. however, tend to operate in an opposite manner that leads toward excess, increase and perversion of potencies, and chaos. Both of these tendencies presumably contrast with the tendencies of a state of purity, although the properties, symptoms, or consequences of purity rarely are explicitly defined in cultural ideologies, in contrast to the wealth of detail elaborated on the consequences of pollution.

On the whole, purity seems to be equated with whatever a culture considers to be the most advantageous mode of being and functioning for achieving the paramount ideals of that culture. Thus, throughout most Asian religions (*e.g.*, Hinduism, Buddhism, Jainism, and Taoism), purity is equated with calmness (physical, mental, and emotional equilibrium) in keeping with the ideal goal—at least for religious adepts—of achieving spiritual transcendence or liberation. In contrast to such Asian religions, groups whose dominant cultural orientation is pragmatic and this-worldly, such as the Yurok, often equate the state of purity with vigour and quickness of mind and body.

Purity and pollution in relation to religious concepts. Concepts of purity and pollution may tend to merge with several concepts of religion: the sacred, sin, and the forces of evil.

Consequences of contact with the sacred or the polluted

Pollution and the sacred. The consequences of contact with both the sacred (the transcendent realm and objects infused with transcendent qualities) and the polluted may be identical, although the reasons for the consequences in the two cases are quite different. The dangers of contact with the sacred may also arise from the belief that the gods are offended by pollution; they will punish a person who defiles a sacred precinct or object (for example, a menstruating woman who enters a temple or shrine in Buddhism and many other religions). The gods may even punish an entire village or tribe for such an offense. To come into contact with the sacred is also viewed as dangerous because the sacred is highly powerful or "charged" with energy; thus, one must be properly strengthened (usually by purification) for the encounter. If one is not thus strengthened, he will be overwhelmed. Although contact with the sacred may have negative consequences for a person, this is not because the sacred is polluting. On the other hand, the dangers of encounter with a polluted person (*e.g.*, an "untouchable" in India) or object (*e.g.*, feces, in most cultures) arise directly from the pollution that passes from that person or object to oneself.

Pollution and sin. Purity and pollution beliefs may become incorporated into a religious morality system in which pollution becomes a type of sin and an offense against God or the moral order, and purity becomes a moral or spiritual virtue. Thus, for example, in the Old Testament, the pollution of birth must not only be cleansed by symbolic or ritual gestures; it must also be atoned for as a cultic sin that offends the sacred precincts of the Lord. In general, the more universalistic religions—Christianity, Buddhism, and Islām—seem to de-emphasize true pollution concerns, and to subsume them within their frameworks of moral and religious beliefs. Both the Qur'ān of Islam and the New Testament of Christianity show a sharp decrease in rules of specific pollution avoidances (*e.g.*, fewer food prohibitions) compared to the Old Testament. Similarly, the sacred texts of Buddhism stress the unimportance of specific avoidances and rituals (in implicit contrast to the multiple and detailed purity regulations of Hinduism) and the necessity for cultivating one's spiritual and moral development instead.

Pollution and the forces of evil. Ideas of pollution are often closely associated with beliefs in demons, sorcerers, and witches. All of the latter may be viewed, in part, as personifications of the powers of pollution. People in polluted states are believed to be dangerous not only to others because they may spread their pollution, but they themselves are often thought to be in danger of attack by demons, who are attracted by the defiled person's impurities (see also ANGELS AND DEMONS; MAGIC; WITCHCRAFT).

CATEGORIES AND THEORIES OF POLLUTION AND IMPURITY

Categories of pollution and impurity. Four major categories of what various religions and societies have regarded as polluting or inherently impure phenomena may be distinguished. Virtually any type of impure person, object, or state (as defined in various cultures) may be assigned to one of these four categories, or may be shown to have symbolic associations with one (or sometimes with several) of these four sets.

Universal belief that bodily functions involve pollution

Physiological processes. The functions of the human body are, for the most part, universally considered polluting, although all functions are not considered polluting in all cultures. The intensity with which the various processes are abhorred also varies from culture to culture. The list of polluting organic processes and things includes menstruation, sexual intercourse, birth, illness, death, and all bodily excretions and exuviae (urine, feces, saliva, sweat, vomit, blood, menstrual blood, semen, nasal and oral mucous, and hair and nail cuttings). Associated with this category symbolically may be various persons, animals, natural objects, sense-related objects, and professions: women in general (because they menstruate), pregnant women, prostitutes, and widows (the latter because of their additional association with death); pigs, dogs, and other scavengers because they eat or associate with excrement and garbage; carrion-eating animals because of their association with death; leftover food, because it has come in contact with saliva via the fingers or utensils that have touched the mouth, or because it may visually resemble vomit or the undigested contents of the stomach; pungent vegetables or spices (such as garlic, onions, and leeks) and strong-smelling meats or fish because they cause foul breath odours; food in general because of its ultimate state as excrement; certain professions because their members are required to handle corpses or bodily exuviae; and things associated with lowness—the entire body below the navel, the feet, the hem of the garment, the floor or ground—because most bodily excretions derive from the lower part of the body.

Violence and associated processes. A second major category of polluting phenomena involves violence and all associated aspects. This entire category may be reduced to beliefs in the polluting nature of blood and death, but the extensive development of various ideas connected with violence pollution merit its being classified as a separate category. Violence pollution involves a wide variety of activities: murder, hunting, warfare, physical fights, quarrelling, cursing or speech that is considered foul, aggressive language, lying, and various aggressive human passions (*e.g.,* greed, anger, and hatred). Various phenomena considered polluting in one culture or another may be placed in this category because of their symbolic associations with violence: Satan, demons, witches, predatory ghosts, and the practice of black magic; alcohol because it stimulates aggressive impulses; carnivorous, predatory, and aggressive animals; meat because of the act of slaughtering the animal; certain professions because their members manufacture weapons or kill or fight for a living.

Anomalies. The third major category includes strange, unusual, or unclassifiable phenomena: (1) certain events of nature (*e.g.,* comets or lunar or solar eclipses); (2) unusual deaths (*e.g.,* death by lightning); (3) unusual births (*e.g.,* twins or other multiple births, breech deliveries, miscarriages, or stillbirths); (4) physical deformities, especially sexual deformities (*e.g.,* monorchids [men having one testicle], hermaphrodites, or eunuchs); (5) speech defects and voices appropriate to the opposite sex; (6) unusual developmental sequences (*e.g.,* children who cut their upper teeth before their lower); (7) anomalous animals or types of plants that have features of several species; (8) anomalous states of matter, such as viscosity, which is neither solid nor liquid; (9) persons in liminal (threshold or transitional) categories or states (*e.g.,* persons undergoing initiation rites, strangers, or captives); (10) persons not considered fully in control of their faculties (*e.g.,* children, drunken persons, the insane, or the mentally or physically handicapped, such as cretins); and (11) perversions of social relationships, especially sexual, that a culture generally considers to be normal (*e.g.,* adultery, homosexuality, bestiality, incest, births of children to unwed parents or as a result of adulterous relationships, or the breaking of vows of celibacy by monks or nuns). That pollution results from a confusion of classification rules may explain beliefs that certain objects must not be mixed lest pollution result. The Old Testament prohibition (also found in certain African groups) that meat and milk should not be mixed with one another or the prohibition in the Vedas (ancient Hindu scriptures) against carrying water and fire at the same time are examples of attempts to maintain classificatory purification rules (see also ANIMALS AND PLANTS IN MYTH AND LEGEND; DIETARY LAWS AND CUSTOMS).

Pollution in relation to caste professions and ethnic structures

Social classifications: classes and castes. The belief that the lower castes pollute the upper castes has been explicit in India, where a true caste system has existed. These lower castes, to some extent, are considered polluting because they engage in professions that have been or are associated with the physiological processes or with violence. Many lower caste occupations (*e.g.,* pottery making or basket weaving), however, do not have such associations, and thus the categorization of pollution attached to all lower castes cannot be so explained. Outside true caste systems, there are de facto systems of racial or ethnic hierarchy, in which certain races or ethnic groups are considered to be inherently lower than others. In most such systems, the notion that the lower groups pollute the higher is not stated explicitly in terms of pollution; the language of racial or ethnic prejudices in such systems, however, is often strongly reminiscent of pollution concepts—*e.g.,* that the lower groups are "dirty," have peculiar bodily odours, engage in sexual promiscuity or perversions, are "animals," or are violent and dangerous. Relations between the dominant race or ethnic group and the subordinate one often resemble the relations between upper and lower castes in India. In such social systems, eating together and intermarriage generally are not condoned, and segregated neighbourhoods and public facilities to maintain minimal physical contact are encouraged by law or custom (see also CASTE SYSTEMS).

Theories of pollution and impurity. Though these four major categories indicate the great diversity of phenomena considered polluting cross-culturally, no one culture considers every item noted in these categories as polluting. Furthermore, within a single culture, not every item considered polluting is necessarily polluting to every member of the society, because the connotation of pollution often is dependent upon the occasion and on the status of a person. The pollution of death, for example, may be confined to those who have actual contact with the corpse, the immediate family of the deceased, certain categories of kinsmen, or all members of the village in which the death has occurred.

The rules dictating avoidance of certain groups or individuals because of the threat of pollution may be seen as means that a society has at its disposal for emphasizing its important social categories. Thus, in the case of death, if relatives on the father's side but not on the mother's side are considered polluted by the death, it may be theorized that this is one of the society's ways of emphasizing the greater social significance of the patrilateral relatives in the kinship system. Sociologists and anthropologists, on the one hand, tend to stress such social implications of pollution rules. On the other hand, some psychologists, philosophers, and theologians are more interested in explaining what there is about polluting events and processes (*e.g.,* death and menstruation) in themselves that would result in their being considered polluting in so many cultures.

Psychoanalytic and social-classification theories

Two general theories have been proposed in relation to these emphases or questions. The first theory derives primarily from psychoanalytic theories developed by Sigmund Freud in which the quest for sexual, excretory, and aggressive pleasures are viewed as instinctual drives in man that are repressed or greatly limited in the socialization of the individual. Hence, because many of the phenomena viewed as polluting cross-culturally are related to these concerns, pollution fears are interpreted as projections or symbolizations of these repressed instincts. The second theory that attempts to explain the specific content of pollution-belief systems (as opposed to the social effects of those beliefs) maintains that, in a very broad sense, things are considered polluting by virtue of their relationship to cultural classification. This theory holds that everything considered polluting in any culture either is anomalous in relation to basic cultural categories or is positioned at the extremities—*i.e.*, the margins—of major conditions or situations of individual or social existence. Birth and death, for example, are at the margins of an individual's life, and the lower castes are at the margin of society.

Both of these theories, however, contain certain problems that may be resolved by subsuming them under a more general theory. The theory derived from psychological considerations is regarded by many scholars as being too narrow in scope because it ignores many types of pollution data; the theory based upon cultural classification, because it is capable of such broad interpretation, loses its coherence as a theory. A more general view incorporates these two theories within a single more fundamental one based on denial. Thus, pollution fears might be interpreted as symbolizations of any material that is denied full expression—psychologically, culturally, or socially. The Freudian theory, emphasizing the psychoanalytic notion of repression of instinctual drives, thus becomes significant in interpreting the first two categories—physiological processes and aggression (*i.e.*, violent emotional processes). The classification theory, which emphasizes cultural attempts to ignore or suppress phenomena that do not fit its cognitive-classification schemes, then becomes significant in interpreting the third category of polluting things—anomalies, unusual occurrences or types of persons, and "mixings." To account for the fourth category, involving the fear of lower castes, classes, and ethnic groups as polluting, the sociopolitical notion of oppression may thus be introduced. All these concepts—repression, suppression, and oppression—are related to the notion of something or someone being forcibly prevented from expression; that is, of being under some sort of pressure. This idea suggests why polluting things are viewed as threatening and not simply as interesting peculiarities of the world, because things under pressure are volatile, liable to escape, or capable of erupting at any moment.

TYPES OF PURIFICATION RITES

Occasions and symbolism of purification rites. Purification rites are required whenever there has been some kind of polluting contact. In addition, cultures may institutionalize regular, periodic purification rituals on the general principle that pollution occurs all the time. Important changes of status or quests for special or sacred status may be viewed as progressions from lesser to greater states of purity, and such changes or quests thus entail rites that promote the anticipated progressions. Purification is invariably required before any contact with the sacred. Purification also is generally considered necessary after any kind of traffic with the demonic forces and black magic, because these contacts with the nether realm are viewed as polluting experiences. Purification rites also may be required before undertaking a major endeavour in order to ensure the participant's success and a right relationship with the special powers involved in the project.

Rites before or after contact with the sacred

Though every culture has rituals to rectify unavoidable pollution, prescriptions of avoidance, abstention, separation, and seclusion are utilized to minimize contacts with polluting persons, objects, or places. Seclusion devices, which confine the very pure or the very polluted within an enclosed area away from other members of society, include menstrual huts, nuptial huts, and birth huts. Initiates are generally confined to special houses or isolated from the community by living for certain required periods of time in the bush or forest. Priests often withdraw to the inner rooms of temples to prepare for or to participate in contacts with the sacred; monks and nuns confine themselves or are confined to monasteries in order to remain undefiled by the world, among other reasons. Seclusion or containment may also be symbolically effected by the use of veils or by the drawing of circles or other enclosures around the object in question. Under the general heading of segregation, groups of different grades of purity may retire to their respective parts of a town when their periods of contact with other members of their community are completed for the day. Men may have special houses for their esoteric activities from which women are excluded. Impure persons may be required to cook over a separate fire; persons of different grades of purity often are not permitted to eat together, to sleep under the same roof or in the same room, and, almost universally, to marry or have sexual relations with one another. Finally, complete abstention, for a fixed period of time, from such polluting activities as sex, eating, and other sensuous indulgences is a significant aspect of purification processes in many societies around the world.

Classification of purification rites. Various kinds of avoidances and abstentions represent the passive aspect of purification. The active aspect consists of the purification rites themselves. Such rites may be classified according to the principle on which they operate.

Rites of cleansing

The removal of pollution. Based on the analogy of cleansing outer dirt or stains by means of bathing or washing in everyday life, purification of man's inner state of being is almost universally believed to be effected by rituals involving various forms of washing. The polluted individual might be required to swim or bathe in the sea, a river, a pond, or special tank. Bathing in swift-flowing streams is often considered especially effective because the rapidly flowing water not only removes the impurities but carries them away. A polluted person might wash his entire body with water or only certain parts of the body that represent the body or person as a whole—rinsing or cleaning the mouth by other means is common. Water may be poured, sprinkled, thrown, or blown upon a polluted person or object. Simply touching water is a purifying gesture in the Vedas; gazing at it is considered purificatory in Sri Lanka (Ceylon). In the absence of water various kinds of moist substances may be used—clay, mud, wet herbs, or plants. The Qur'ān (the Islāmic sacred scriptures) directs desert dwellers and travellers to rub themselves with high clean soil because of the scarcity of water. In cultures in which saliva is not considered polluting, expectorating or breathing on something may be viewed as purificatory gestures.

Other modes of purification based on the analogy of cleansing outer dirt include: the use of wind or aeration to blow or carry away the impurities; sweeping a house or certain area of the ground or brushing the polluted person or object, often with a brush made of fibres from a symbolically pure source; scraping the surface of a polluted object or utensil; shaving and cutting the hair and nails; removing clothing and washing it or destroying it; and putting on clean or new clothes.

The expulsion of pollution. Based on the analogy of expelling internal physical poisoning or corruption, a second category of purification rites involves the actions of expelling, ejecting, purging, or drawing out the pollution from the defiled person or object. The use of purgatives in purification rites to induce vomiting is not uncommon. Sweat baths and steam baths are believed to bring the impurities out of the person as symbolized by the emerging sweat. Some purification rites involve bloodletting in order to drain out impurities. The use of salt in some rites may be based on the fact that salt has drawing or draining properties. In corporate acts of expelling pollution, an entire community may purge itself of a polluted individual in its midst by excommunicating him and forcing him to leave the religious group, caste, tribe, or area.

Ceremonies involving scapegoats

The *transfer* of pollution. Closely connected with the practice of drawing pollution from the defiled person or object is the notion that pollution may be transferred from a person or community to another object that is either immune to pollution itself or that can be discarded or destroyed. The most dramatic rites embodying this principle are scapegoat ceremonies in which pollution is transferred to an animal or person by either touching, bathing with, or simply pronouncing the pollution transferred to the scapegoat. The scapegoat is then run out of town or killed, actually or symbolically. The victim may further be made into an offering or sacrifice to the gods on the general ritual principle of keeping the gods satisfied. In the classic scapegoat ceremony of the Old Testament, as noted in Leviticus, chapter 16, the animal—called Azazel (a desert demon)—was simply released to wander the wilderness; in Bali (in Indonesia) birds act as scapegoats and are then released to fly away.

Less dramatically, pollution may be transferred to a relatively worthless talisman (charm). Some talismans are regarded as convenient because they are disposable and of little value; after they have served their purposes in specific situations they are thrown away. In Bali a three-month-old child is purified by transferring his impurities to a chicken; this chicken may then become his pet and continue to absorb the pollutions to which the child is exposed. It may never be killed or eaten, and when it dies it is buried with respect.

Forms and objects of pollution-destroying rites

The destruction of pollution. Pollution is also believed to be eliminated by destroying the polluted object. The killing of the scapegoat belongs to this general category; more dramatically, a severely polluted person may himself be killed rather than being allowed the opportunity to transfer his impurities onto a more dispensable animal or object. The execution of a polluted person or a scapegoat animal often takes the form of drowning, choking, suffocating, or clubbing so that the pollution might not escape with a flow of blood. Polluted metal objects may be melted down; polluted fires are extinguished; polluted clothing, utensils, and other items are torn, broken, and often buried.

The most common means of destroying pollution is by burning the polluted objects. Fire is a most efficient destroyer; when the flame no longer exists there is virtually nothing left of the objects. Fire is generally conceived, however, as having more positive purifying properties, not only destroying pollution but creating purity.

The *transformation* of pollution into purity. Fire is perhaps one of the most symbolically complex phenomena in the history of human culture. It renders raw meats and vegetables into cooked and edible food, base minerals into useful and durable metals, and porous dirt and clay into watertight pottery. It destroys the forests and brushlands, but its ashes make the earth fertile and productive. Fire is thus viewed as a powerful transformer of the negative to the positive. Because of such properties, fire is commonly found in purification rites throughout the world. Polluted persons may be required to walk around, jump over, or jump through fire. Polluted items may be singed, fumigated, or smoked. The widespread use of incense smoke in purification rites is based on the transforming powers of fire, as well as on the additional purificatory powers of sweet smells. Polluted persons or things may be rubbed with ashes or soot, and polluted objects may be boiled, subject to the double purifactory powers of fire and water. Exposure to sun and to intense heat are also regarded as practices falling into this same general category. The extinguishing of old fires in temples and villages and the kindling of new ones are common practices after a death or as part of annual renewal and purification ceremonies. Alchemic experiments, which attempt to purify mineral substances and turn them into gold, involve boiling or melting down the solution or elements over pure and intense heat and then recrystallizing them in newer and higher forms (see also ALCHEMY; TAOISM).

Phenomena viewed as intrinsically pure

The introduction of purity. In addition to the cleansing, purging, destruction, and transformation of pollution, most purification rites involve the positive introduction of purity. Many phenomena are considered inherently pure; ingestion of, or contact with, or simply exposure to such phenomena is believed to bring purity to the object of the ritual.

Objects, activities, or persons commonly considered to have intrinsic purity cross-culturally include: fire; water; sweet smells created by flowers, fragrant plants and herbs, perfumes, fragrant oils, or incense; milk, ghee, and other dairy products; white objects; earth in its natural form; sacred objects (*e.g.*, relics) and sacred personages (*e.g.*, priests); the recitation of spells, incantations, and names of gods; magical amulets and stones; gold and, in one culture or another, silver, bronze, jade, and crystal; virgins; the right as opposed to the left side of things in many cultures (*e.g.*, the Abaluyia of Kenya); morning, sunshine, and daylight as opposed to darkness; whole or perfect objects, including circles and wheels and perfect numbers—*e.g.*, the number nine (because the digits of any of its square products always add up to nine) or four (because quaternity is viewed as perfection); and physically perfect specimens of their species. In addition, cultures idiosyncratically define certain things as pure because of special cultural associations: cow dung and cow urine are pure in Hinduism because of the sacredness of the cow; dogs are considered to be pure in Zoroastrianism (a religion founded in the 6th century BC by the Iranian prophet Zoroaster) because as scavengers they purify the world for everyone else (most cultures view dogs as impure because of their scavenging habits); and all cool things are considered pure among the Lovedu of the Transvaal because pollution is associated with heat.

Use of pollution to achieve purity

Other purification rites. Purification practices in which pollution is introduced qua pollution in order to achieve purity are also found in various religions and cultures. These rather paradoxical practices work on several different principles. The use of garlic, sulfur, or an amulet made of impure materials apparently operates on the principle of like attracting like; the impure amulet draws the impurity encountered in some situation toward itself, thus preventing it from polluting the wearer of the charm. Another set of practices apparently works on an inoculation principle—a baby, a magical implement, or a special work area may be briefly exposed to menstrual blood, for example, to protect it against future pollution from the same kind of item. A third group of such paradoxical practices, found primarily in Asian religions, involves immersing oneself in what is viewed as utter pollution, either by meditating on foul things or by actually keeping oneself permanently unclean, in order to achieve transcendence over pollution. Ordeals, mutilations, and blood rituals in general may also be regarded as fitting the transcendence pattern.

In highly developed and elaborated systems of thought, purity and pollution meet and merge. Buddhist monks are considered to be extremely pure, yet they are directed to make their robes from cemetery cloths, and beds or litters used in funerals may be donated to their monasteries. Buddhist relics with great purifying power are often composed of bits of hair, nails, and bones (albeit of the Buddha or other great saints); in Sri Lanka the word (*dhātu*) for such relics is the same as the word for semen. Monks and nuns of Jainism (an Indian religion founded by Mahavira in the 6th century BC) are ordered not to bathe and under no circumstances to clean their teeth. In Hinduism, if a Brahmin (a member of the highest caste) enters a street of the untouchables (outcastes), he is polluted, but the whole street also falls prey to disease, famine, and sterility. In a Burmese folktale, an alchemist became discouraged with his experiments and threw his alchemic stone into a latrine pit; on contact with the excrement, the stone achieved purity—thus indicating that contacts with pollution may bring about purity.

Ritual renewal and sacrificial rites

Many rituals considered to effect purification do not utilize any of the specific purifying techniques outlined above. They simply make use of techniques believed to have generalized ritual efficacy, no matter what the disorder. Thus, some purification rites involve reversals, especially reversals of roles between men and women, on the general principle that they represent a return to chaos and then a change back to order. Another widely practiced ritual principle involving the symbolism of reversal is that of death and rebirth; man and the world, with all their disorders, are symbolically put to death and then symbolically renewed in a purer and better state. Because blood is associated with both life and death, the use of blood in purification rites is often central to the symbolic renewal process. Nearly all rituals involve the reading or reciting of spells, texts, or prayers that have a generalized efficacy over negative forces, and in many cases purification may be accomplished by these means without any further symbolization of cleansing or a re-creation of purity. When pollution becomes one of many possible offenses against the gods, purification may be accomplished simply by making sacrifices or offerings to the gods. Pollution often becomes identified with immoral or sinful behaviour and in such instances purification may be effected by punishment of the offender, by the offender's spiritual atonement, or by acts of penance and virtue, such as giving alms. Purity also may become identified with the struggle against the demonic forces, and in this transcendent dimension purification is effected in rites of exorcism or in rites that placate the demons. The use of weapons in purification rites is often based on a symbolic battle with the forces of evil; the use of firecrackers in some purification rites is viewed as a means of frightening away the demons; the use of curses, abuse, ridicule, and ribaldry in purification rites among the ancient Greeks, for example, was regarded as forms of protection against the demons. Some purification rites involving blood are structured in terms of giving demons what they want in order to turn away their polluting presences (see also SACRIFICE).

EXAMPLES OF COMPLETE PURIFICATION RITES

Most full-scale purification rites combine several of the principles outlined above. A few of the immense number of complex purification rites in the religions and cultures of the world follow.

Rite for purifying a cured leper in ancient Judaism. In the Old Testament purification rites for a person who has been cured of leprosy, as described in Leviticus, the leper and the priest meet outside the camp, and the priest examines the man to ascertain that he is cured. The priest then calls for two live, clean birds, cedar wood, a scarlet item, and hyssop (an aromatic herb). One of the birds is killed in an earthen vessel over running water. The live bird and the other ingredients are then dipped in the blood of the dead bird and used to sprinkle blood seven times upon the leper while the priest pronounces him clean. The live bird is then allowed to fly away. The leper washes his clothes, shaves off all his hair, and washes himself, after which he is allowed to enter the camp, although he must remain outdoors for seven days. On the seventh day he once again shaves off his hair, including his eyebrows, and washes his clothes and body. On the eighth day he goes to the temple to make various offerings to the Lord. The priest then takes some of the blood of one of the offerings and places it on the man's right ear, thumb, and large toe of the right foot, after which he does the same with some oil that is being offered, also pouring some oil on the man's head. The sacrifices are then offered to the Lord upon the altar, thus completing the required ritual: "the priest shall make atonement for him, and he shall be clean."

The Navajo sweat-emetic rite. The Navajo sweat-emetic rite is part of most major Navajo ceremonials for curing illness or rectifying other ritual disturbances. It is specifically viewed as a rite of purification.

A ritual hut is prepared with sand paintings, and a fire is then built. A procession of patients, led by the chanter, enters the hut and circumambulates the fire, pausing at each of the four directions to sing an appropriate chant. In some cases there is fire jumping; the men are required to jump over the fire, and the women to walk as close to it as possible. The audience then enters, with men and women sitting in segregated groups. The chanter heats wooden pokers in the fire and applies them to himself, mainly on the legs, and then to all the patients. Basins in front of each patient are filled with the emetic formula, the fire procession is repeated, and the emetic is then drunk. Everyone is expected to vomit; if they do not, it is regarded as inauspicious. Vomiting is done into receptacles containing sand, and the contents of these receptacles may then be sprinkled with ashes from the pokers. A bull-roarer (a heavy stone on a string that produces a deep roaring sound when whirled) is sounded outside six times and then brought in and applied to the patients. The audience leaves the hogan (hut) in procession, this time led by assistants who carry out the basins with their contents. The contents of the basins are deposited neatly in a row outside the hogan and allowed to be dispersed by the natural elements. The patients, however, remain inside the hogan, perspiring in the heat. Later, the audience re-enters; the fire is broken up and extinguished, and all remnants of it are removed to a place near the basin area. The chanter sprinkles all present with a medicinal lotion and then fumigates everyone with incense. All then leave in procession and dress outside.

The Zoroastrian "Great Purification" rite. The "Great Purification" rite (*baresnum*) of Zoroastrianism originally was intended for purification from serious polluting contacts, especially for corpse bearers after contact with death. The rite was later pre-empted for initiation into the priesthood, or for attaining higher statuses within it.

In preparation for this rite a priest seeks a piece of ground regarded as clean (*i.e.*, dry and unfrequented by men or animals). He then cuts down any trees located on the area selected. Nine pits are dug in a certain arrangement; furrows are drawn around three, then around six. and finally around all nine pits. Thereafter, the whole area is covered with sand. After these activities have been completed, the priest stands outside the outer furrow, and the subject requiring purification advances to the first pit and is told to recite praises to the "Purity of Thought." The priest, holding a stick with nine knots and with a spoon fastened to the end, uses the spoon to pour consecrated cow's urine (gomez) upon the hands of the subject, who washes his hands with the urine three times. He then washes his entire body with gomez, progressing from the head down to the feet. The pollution is said to leave the toes in the form of a foul-smelling fly. After the one seeking purification has washed himself with *gomez,* the priest recites purifying formulas. This process is iepeated at each of the first six pits; at a prescribed distance from the seventh pit, the subject. sits down and rubs himself 15 times with sand, making sure that he is completely dry. At the seventh pit he washes his body once, from head to toe, with water; at the eighth pit he does this twice and at the ninth pit three times. His body is then fumigated with the smoke of fragrant wood, after which he dresses in clean clothes. In certain versions of the ceremony, a dog is presented to the candidate, who, after each washing at each pit, must touch the left ear of the dog with his left hand. At the end of the ceremony the candidate is required to recite the following formula: “The Evil Spirit of pollution is put down. The head and the body have become purified. The soul has been purified. The dog is holy, the priest is holy."

The candidate then retires to a house and is required to have no contact with fire, water, cultivated land, trees, cattle, men, or women. On the fourth, seventh, and tenth days he again bathes with gomez and then with water. After the final bath he is considered "perfectly purified."

POLLUTION BELIEFS IN MODERN SOCIETY

Pollution beliefs and fears occur in modern society as well as in any other, although they are not systematized and usually not understood as such. Racism and other forms of prejudice apparently play upon pollution fears. Of less serious consequence are such notions that warts result from masturbation (traditionally considered a polluting or impure practice in conventional Western societies), that there is something dangerous or polluting in intercourse with menstruating women, and that (as in a New York state law) men and women should not have their hair cut or beauty services performed in the same room. Physiological processes (*e.g.*, urination and other forms of elimination) are often viewed with disgust, and as a result many modern notions of sanitation are based on not entirely rational principles. The highly developed mortuary profession (especially in Western countries) protects persons in contact with death not only from grief but probably from pollution fears as well. On the whole, however, there are fewer pollution beliefs in modern society than in traditional societies. This trend may be attributed in part to the assimilation of these beliefs into moral and religious concepts.

BIBLIOGRAPHY. Few works deal directly with the subject of purification rites. MARY DOUGLAS, *Purity and Danger* (1966), is a major recent work dealing with the problems of purity and impurity. HUTTON WEBSTER, *Taboo: A Sociological Study* (1942); and FRANZ STEINER, *Taboo* (1956), deal with pollution taboos as part of the general field of ritual prohibitions. The *Encyclopaedia of Religion and Ethics,* vol. 10 (1919), though very dated, has a long article on "Purification" with many examples. Religious texts are among the best available sources: The *Old Testament,* the Egyptian *Book of the Dead,* and the *Sacred Books of the East.* The latter includes texts of Hinduism, Buddhism, Islām, Zoroastrianism, and Taoism. For a good summary of Zoroastrian purification rites (which are extensive and elaborate), see J.J. MODI, *The Religious Ceremonies and Customs of the Parsees,* 2nd ed. (1937). For ancient Greece, see LOUIS MOULINIER, *Le Pur et l'impur dans la pensée des grecs, d'Homère à Aristote* (1952); and JANE HARRISON, *Prolegomena to the Study of Greek Religion,* 3rd ed. (1922). The material on purity and pollution in primitive and other traditional societies is found in many sources. For excellent synopses of African thought systems, see DARYLL FORDE (ed.), *African Worlds* (1963); and for a North American tribe, see GLADYS REICHARD, *Navaho Religion,* 2nd ed. (1963). MARGARET MEAD, *Sex and Temperament in Three Primitive Societies* (1935 and 1963), brings together material on three Pacific Islands societies.

(S.B.O.)

Puritanism

Puritanism, a reform movement in the Church of England during the late 16th and 17th centuries, sought to carry the Reformation beyond the stage it reached at the beginning of the reign of Queen Elizabeth I (1588–1603). The name Puritan apparently was first used in the 1560s against those who thought it was necessary to "purify" the Church of England from remnants of Roman Catholic "popery."

NATURE AND SIGNIFICANCE

Puritans became noted for a spirit of moral and religious earnestness that determined their whole way of life. Their ideal life style was prefigured in Geoffrey Chaucer's description of the parson in *The Canterbury Tales,* almost two centuries earlier. What made the Puritans of the 16th and 17th centuries different from Chaucer's parson, however, was their heroic attempt through church reform to make their own life style the pattern for a nation.

Goals of the Puritan reform

The Puritan program for reform aimed to develop the Church of England by placing in every parish a pastor who would proclaim the Word truly (in a Protestant sense), administer the sacraments rightly (with no semblance of "popish idolatry"), and maintain discipline seriously (admonishing and correcting members of the parish in order to transform profane lives into holy lives).

In their efforts to transform the nation, Puritans were responsible both for revolutionary activities that led to civil war in England and for the founding of colonies in America as working models through which the people could be instructed. The religious, social, economic, political, literary, artistic, and intellectual institutions of the modern English-speaking world have been influenced by the Puritan spirit. The nature of that influence continues to be a matter of lively debate.

HISTORY

Origins of Puritanism (from the Reformation to 1603). The ideas that gave Puritanism its energy were rooted in the thought of England's earliest Protestants, of whom the most important was William Tyndale (c. 1494–1536). Tyndale viewed England as a covenant nation, like ancient Israel, existing under the judgment and mercy of God. He spoke of the Protestant Reformation as a means through which God in his grace was giving England an opportunity to renew that covenant. If England repented and purged the land of "popish idolatry" and sought to keep the laws of God, Tyndale contended, God would make good his promises. Otherwise, England would suffer the wrath of God in the form of war, famine, or plague. Temporal prosperity was possible, he said, in a reformed England.

Reformation origins. When the Church of England under Henry VIII was separated from Rome by the Act of Supremacy in 1534, much Roman Catholic form and substance remained. English Protestants worked for further reformation but had to proceed with extreme caution for fear of being burned at the stake as heretics.

On his ascension to the throne in 1547, young Edward VI was hailed by Archbishop Thomas Cranmer and other Protestants as England's Josiah, the young 7th-century-BC king of Judah who enforced the Deuteronomic reform. Edward, it was held, would rid the land of idolatry so that England might be blessed. Protestantism advanced rapidly during his reign through the systematic reformation of doctrine, worship and discipline—the three external marks of the true church. A reformed confession of faith and a prayer book were adopted, but the reformation of the ecclesiastical laws that would have defined the basis of discipline was blocked in Parliament by the most powerful of the English nobility.

The external marks of the church

The death of Edward and England's return to Roman Catholicism in 1553 under Queen Mary was interpreted by Protestants as a judgment by God upon a nation that had not taken the Reformation seriously enough. Many, including Cranmer, died as martyrs to the Protestant cause. Others fled to the European continent. Those in exile experimented with more radical forms of worship and discipline. Leading clergymen published material justifying rebellion against an idolatrous ruler. Many saw in Geneva, which was a haven for English exiles, a working model of a disciplined church. Exiles produced two large volumes of incalculable consequence for English religious thought. John Foxe's *Actes and Monuments,* popularly known as *The Book of Martyrs,* and the Geneva Bible were the most popular books in England for many years after they were published. They provided a view of England as an elect nation chosen by God to bring the power of the Antichrist (understood to be the pope) to an end. An England obedient to God would receive his favour. Otherwise, it would experience his plagues.

The Book of Martyrs and the Geneva Bible

Elizabeth, beginning her rule in 1558, was hailed as the glorious Deborah (12th-century-BC Israelite leader), the "restorer of Israel." She did not restore it far enough for English Protestants, however. Two statutes promulgated in her first year—the Act of Supremacy, stating that the queen was "supreme governor" of the Church of England, and the Act of Uniformity, ensuring that English worship should follow *The Book of Common Prayer*—defined the nature of the English religious establishment. In 1563 the primary church legislative body, the Convocations of Canterbury and York, defined standard doctrine in the Thirty-nine Articles, but attempts in the Convocation to reform the prayerbook further and to produce a reformed discipline failed. Defeated there, the reformers came to rely more on Parliament where they could always depend on strong support.

Emergence of Puritanism. Puritanism surfaced during a controversy over vestments when the Queen demanded that a reluctant archbishop, Matthew Parker, enforce uniformity in the dress of the clergy. Parker did so with the publication of his *Advertisements* in 1566. Those who regarded the prescribed garb as remnants of popery were called "Precisians" or "Puritans" at this time. Later opponents of Puritans, such as John Whitgift (professor and vice chancellor of Cambridge University and later archbishop of Canterbury), were sympathetic on this point, and they urged the higher authorities not to make such an issue of dress. The issue was more serious than dress, however, for it involved the maintenance of the Queen's supremacy.

Thomas Cartwright's contributions

In 1570 Thomas Cartwright, in lectures delivered at Cambridge University, proposed that the presbyterian form of church government was the type ordained by God in the Bible. He was dismissed from his position as Lady Margaret professor of divinity by Whitgift and fled to Geneva. Two years later an anonymously published *Admonition to the Parliament* called for the abolition of the entire hierarchy of the Church of England in order to establish the right scriptural government there and thereby save England from plagues to be sent by God. It advocated a form of government in which there would be equality of all clergy and a body of elders and pastors in each congregation who would correct and punish "all willful persons and contemners" of the Law of God. Whitgift, in reply, maintained that the government of the church should be suited to the government of the state and that Episcopal government best suited monarchy. Cartwright took up the pamphlet war and insisted that the government of the civil state should be determined by what best suited the divinely ordained government of the church.

Later, informal groups of clergy and laymen from several churches met together to expound and to discuss the Scriptures. These groups were viewed by the Queen as a political threat. Edmund Grindal, who succeeded Matthew Parker as archbishop of Canterbury, favoured these meetings (called prophesyings) for their educational value. He refused to carry out the Queen's orders to have them suppressed and wrote to her, "Remember, Madam, that you are a mortal creature . . . and although ye are a mighty prince, yet remember that He which dwelleth in heaven is mightier." The Queen could find no precedent to deprive Grindal of his archbishopric, but she did suspend him from the exercise of his office.

In the 1580s there was a concerted effort in England to set up area presbyteries, or classes, that could help the congregations under them develop their programs of discipline. The *Second Book of Discipline* (1578) recommended supplanting bishops with supervisory assemblies of pastors and elders on area, provincial, national, and world levels. This Presbyterian Puritan program was defeated through the joint efforts of John Whitgift, then archbishop of Canterbury, and Richard Bancroft, a member of the High Commission and later archbishop of Canterbury, who set up an intelligence system to keep abreast of Puritan activities. Extended ecclesiastical Court of High Commission and Star Chamber proceedings effectively brought this Presbyterian movement to a halt.

The Separatists

Some Puritans, concerned with the long delay in reform, decided upon a "reformation without tarrying for any." As Separatists they rejected compromise and repudiated the state church. Around the Separatist movement arose voluntary congregations that covenanted with God and with themselves, chose ministers by common consent, and put into practice the Puritan marks of the true church. The leaders of the Presbyterian movement, such as Thomas Cartwright, repudiated such activity and sought to disassociate themselves from the Separatists. The state repressed Separatists more savagely than it did the Presbyterians, and two laymen were hanged in 1583 for selling Separatist tracts. Three Separatist clerical leaders, John Greenwood, Henry Barrow, and John Penry, were hanged in 1593; others went to Holland as exiles.

In the last decade of Elizabeth's reign, Puritan reformers had to put aside their grand plan for the nation and concentrate on the cultivation of individuals and parishes. There were still many pastors through whom this work could be accomplished and bishops who were friendly to them. Also, there were noble patrons who placed Puritans in parish positions and members of Parliament who made common cause with them. Puritans controlled colleges and professorships at Oxford and Cambridge, and people of means contributed to lectureships assuring large audiences for Puritan preaching. Thus the Puritan spirit continued to spread.

Puritanism under the Stuarts (from 1603 to 1649). *Events under James I.* Puritan hopes were raised when James VI of Scotland succeeded Elizabeth as James I of England in 1603. James was known to be Calvinist in theology, and he had once signed the Negative Confession of 1581 favouring the Puritan position. In 1603 the Millenary Petition (with a claimed thousand signatures) presented Puritan grievances to the King, and in 1604 the

The Hampton Court Conference

Hampton Court Conference was held to deal with them. The petitioners were sadly in error in their estimate of the King, who had learned by personal experience to resent Presbyterian clericalism. At Hampton Court he coined the phrase, "no bishop, no king." Outmanoeuvred in the conference, the Puritans were made to appear petty in their requests.

As a seal upon the Hampton Court Conference, James appointed Richard Bancroft to be Whitgift's successor as archbishop of Canterbury and encouraged the Convocation of 1604 to draw up the *Constitutions and Canons* against Nonconformists. Conformity in ecclesiastical matters became a pattern in areas where forms of nonconformity had survived under Elizabeth. Though a number of the clergy were deprived of their positions, others took evasive action and got by with minimal conformity. Members of Parliament supported them in their position by arguing that since the canons had not been ratified by Parliament they did not have the force of law.

Puritans remained under pressure, but men of Puritan sympathies still came close to the seat of power in James's reign. The enforced reading from pulpits of James's *Book of Sports,* dealing with recreations permissible on Sundays, in 1618, however, was a further affront to those who espoused strict observance of the sabbath, making compromise more difficult.

Increasing numbers of Separatist groups could not accept compromise, and in 1607 a congregation from Scrooby, England, fled to Holland and then migrated on the "Mayflower" to establish the Plymouth Colony on the shore of Cape Cod Bay in 1620.

Events under Charles I. William Laud emerged as an effective opponent of Puritans, and, in 1628, he was appointed bishop of London by Charles I. London was regarded as the centre of Puritanism, and a policy of "thorough" (so called from Laud's expressed determination to carry his schemes "thorough" [*i.e.*, "through"] any obstacle) was begun there in order to destroy Puritan power. Men who were not Separatists found their positions increasingly difficult to maintain; hence, plans were laid for the founding of the Massachusetts Bay Colony.

Laud became archbishop of Canterbury in 1633 and was an effective leader under Charles in dealing with ecclesiastical matters in England. He misjudged Scotland's loyalty to the episcopal system, however, when he attempted to introduce into the Church of Scotland a liturgy comparable to the Anglican *Book of Common Prayer.* When "Laud's Liturgy" was introduced at the Church of St. Giles, at Edinburgh, a riot broke out that led to a popular uprising that restored Presbyterianism in Scotland.

Charles sought to put down the Scots, but his armies were no match for the Scottish forces. In 1640 he was faced with an army of occupation in northern England demanding money as a part of their settlement. Short of funds, Charles was forced to call Parliament, without which he had been trying to rule since 1629.

The Long Parliament

Suddenly Puritans were in a situation where they not only could preach freely but were being looked to for their leadership by people in power. The first act of the Long Parliament, as it came to be called (1640–53), was to set aside Nov. 17, 1640, as a day of fasting and humiliation. Cornelius Burges and Stephen Marshall were appointed to preach that day to members of Parliament. It was a time of great opportunity for the Puritans. Their sermons urged the nation to renew its covenant with God in order to bring about true religion through the maintenance of "an able, godly, faithful, zealous, profitable, preaching ministry in every parish church and chapel throughout England and Wales" and through the establishment of a civil magistracy that would be "ever at hand to back such a ministry."

Hundreds of similar sermons were preached on monthly fast days and on other occasions before Parliament during the next few years, urging the people to adopt true doctrine, pure worship, and the maintenance of discipline as a means to claim God's blessing so that England might become "our Jerusalem, a praise in the midst of the earth."

The Westminster Assembly

Civil war. An acceptable compromise between King and Parliament could not be found, and civil war broke out between those loyal to the King and those committed to the Parliament. Members of Parliament called together a committee of over a hundred clergymen from all over England to advise them on "the good government of the

Church." This body, the Westminster Assembly of Divines, convened on July 1, 1643, and continued daily meetings for more than five years.

A majority of the Puritan clergy of England probably would have opted for a modified episcopal church government. Parliament, however, needed Scotland's military help. It adopted the Solemn League and Covenant, which committed the Westminster Assembly to develop a church polity close to Scotland's presbyterian form. A small, determined Assembly group of "Dissenting Brethren" held out for the freedom of the congregation, or "Independency," as opposed to the power of presbytery. Others, called Erastians, wanted to limit the offenses under the power of church discipline. Because both groups had support in Parliament, the reform of church government and discipline was frustrated.

The New Model Army

Dissent within the assembly was negligible compared with dissent outside it. Pamphlets by John Milton, Roger Williams, and others schooled in Puritanism pleaded for greater freedom of the press and of religion. Such dissent was supported in the New Model Army, a Parliamentarian army of 22,000 men organized and disciplined under Sir Thomas Fairfax (1612–71) as commander in chief and Oliver Cromwell (1599–1658), and the real power in England was passing to the military leaders who had defeated all royalist forces. Late in 1648 the victors feared that the Westminster Assembly and Parliament would reach a compromise with the defeated King Charles that would destroy their gains for Puritanism. In December 1648 Parliament was purged of members unsatisfactory to the Army, and in January 1649 King Charles was tried and executed.

The age of Cromwell (1649–60). Both Parliament and the assembly continued to sit on a "rump" basis (containing only a remnant after the purges), and Oliver Cromwell emerged as England's Lord Protector. Cromwell was a typical Puritan in his seeing the judgment and mercy of God in events. Military successes to him were definite signs of the blessing of God upon his work.

The Independent clergyman John Owen guided the religious settlement under Cromwell. He maintained that the "reformation of England shall be more glorious than of any Nation in the world, being carried on, neither by might nor power, but only by the spirit of the Lord of Hosts." Error was a problem for both Cromwell and Owen, but, as Owen expressed it, it was better for 500 errors to be scattered among individuals than for one error to have power and jurisdiction over all others.

Such was the basis for a pluralistic religious settlement in England under the Commonwealth in which parish churches were led by men of Presbyterian, Independent, Baptist, or other opinions. Though Jews were permitted to live in England, those publicly holding religious views, such as Roman Catholics or Unitarians, were unacceptable. Cromwell was personally willing to tolerate *The Book of Common Prayer,* but his Parliament was not. Voluntary associations of churches were formed, such as the Worcestershire Association to keep up a semblance of church order among churches and pastors of differing persuasions. The latter association was organized by Richard Baxter (1615–91), whose persuasive and successful use of a mild discipline in Kidderminster served as an example for other pastors.

Radical reform groups

Radical groups appeared, each seeking to implement its particular vision of the New Jerusalem. The Levellers (a republican and democratic political party) in the New Model Army, in 1647 and 1648, interpreted the liberty that comes from the free grace of God offered to all men in Christ as having direct implications for political democracy. The Diggers (agrarian communists) in 1649 planted crops on common land, first at St. George's Hill near Kingston and later at Cobham Manor, also near Kingston, to encourage God to bring soon the day when all men would live in an unstructured community of love with a communal economy. The Fifth Monarchy Men (an extreme Puritan millennialist sect) in 1649 presented their message of no compromise with the old political structures and advocated a new structure, composed of saints joined together in congregations with ascending representative assemblies, to bring all men under the kingship of Jesus Christ. All these radical groups that accepted variations on Puritan themes eventually died out. A group of lasting significance, however, was the Society of Friends, or Quakers, under George Fox (1624–91). With their program of no minister, no sacraments, and no liturgy, they pushed the Puritan logic of no remnants of popery to its ultimate limit.

Though Cromwell believed that it would be better to permit Islām to be proclaimed in the land than to have "one of God's children persecuted" and though he had respect for George Fox personally, Quakers were persecuted for their refusal to pay tithes. One Quaker, James Nayler, was punished cruelly for blasphemy—his tongue was bored with a hot iron, he was whipped on two occasions, and he was imprisoned for two years at hard labour.

The Restoration (1660–85). After the death of Cromwell (1658), conservative Puritans supported the restoration of King Charles II. They hoped for a modified episcopal government, such as had been suggested in 1641 by the late archbishop of Armagh, James Ussher (1581–1656). Such a proposal was satisfactory to many of Episcopal, Presbyterian, or Independent persuasion. When some veterans of the Westminster Assembly went to Holland in 1660 to meet with Charles before he returned, the King made it clear that there would be modifications to satisfy "tender consciences."

These Puritans were outmanoeuvred in their attempt to obtain a comprehensive church, however, by those who favoured the strict episcopal pattern of Laud. A new Act of Uniformity was passed on May 19, 1662, by the Cavalier Parliament. The act required reordination of many pastors, gave unconditional consent to *The Book of Common Prayer,* advocated the taking of the oath of canonical obedience, and renounced the Solemn League and Covenant. Between 1660 and when the act was enforced on August 24, 1662, almost 2,000 Puritan ministers were ejected from their positions.

Persecution of Puritans

As a result of the Act of Uniformity, English Puritanism entered the period of the Great Persecution. One effect of the Civil War on English politics was that Parliament had become more secure in its power. After 1660 the Puritans were faced with a Parliament that did not favour them as a result of the acceptance of the return of the royalist and episcopal positions. Shortly before the Act of Uniformity (1662), the Corporation Act was enacted; it eliminated many Puritan magistrates. The Conventicle Act of 1664 punished any person over 16 years of age for attending a religious meeting not conducted according to *The Book of Common Prayer.* The Five Mile Act of 1665 prohibited any ejected minister from living within five miles of a corporate town or any place where he had formerly served.

Conservative Puritans, although forced into separatism, did not give up the idea of comprehension (inclusiveness of various persuasions). There were conferences with sympathetic bishops and brief periods of indulgence in Puritan preaching that were initiated by the King. Fines and jailings, however, set the tone.

The Glorious Revolution

During the short reign of Charles's Roman Catholic brother, James II (1685–88), fear of Roman Catholic tyranny united politically both establishment and Nonconformist Protestants. This new unity brought about the "Glorious Revolution" (1688), establishing William and Mary on the throne. The last attempt at comprehension failed to receive approval by either Parliament or the Convocation under the new rulers. In 1689 England's religious solution was defined by an Act of Toleration that continued the established church as episcopal but also made it possible for dissenting groups to have licensed chapels. The Puritan goal to further reform the nation as a whole was transmuted into the more individualistic spiritual concerns of Pietism or else the more secular concerns of the Age of Reason.

Puritanism in America. *Virginia.* A decade before the landing of the "Mayflower" (1620) in Massachusetts a strong Puritan influence was planted in Virginia. Leaders of the Virginia Company who settled Jamestown in 1607 saw themselves in a covenant relation to God and they carefully read the message of their successes and failures. A typical Puritan vision was held by the Virginia settler Sir Thomas Dale. His strict application of severe laws disciplining the Jamestown community in 1611 probably saved the colony from extinction, but he also earned a reputation as a tyrant. Dale thought of himself as a labourer in the vineyard of the Lord, as a member of Israel building up a "heavenly New Jerusalem." Like Oliver Cromwell later, whom he resembled, Dale interpreted his military success as a direct sign of God's lending "a helping hand."

Puritan clergymen saw excellent opportunity for their cause in Virginia. The Rev. Alexander Whitaker, the "apostle of Virginia," wrote to his London Puritan cousin in 1614, "But I much more muse, that so few of our English ministers, that were so hot against the surplice and subscription, come hither where neither is spoken of."

The church in Virginia, however, became more directly aligned with the English establishment when the settlements were made into a royal colony in 1624.

The Puritan experiment in America

Massachusetts Bay. In New England, however, the Puritans had their greatest opportunity. Between 1628 and 1640 the Massachusetts Bay Colony was developed as a covenant community. Gov. John Winthrop stated the case concisely in his lay sermon on board the "Arbella" before the colonists landed,

> Thus stands the cause between God and us; we are entered into covenant with Him for this work; we have taken out a commission; the Lord hath given us leave to draw our own articles . . . Now if the Lord shall be pleased to hear us and bring us in peace to the place we desire, then hath He ratified this covenant and sealed our Commission, [and] will expect a strict performance of the articles contained in it.

Lack of performance of the articles, in this view, would bring down the wrath of God.

The pattern for church organization in the colony was determined by John Cotton, who pursued "that very Middle-way" between English Separatism and the presbyterian form of government. Unlike the Separatists, he held the Church of England to be a true church, though blemished; and, unlike the Presbyterians, he held that there should be no ecclesiastical authority between the congregation and the Lordship of Christ. Cotton proposed that the church maintain its purity by permitting only those who could make a "declaration of their experience of a work of grace" to be members. Cotton's plan ensured that church government should be in the hands of the elect, the chosen of God.

Taking their cue from Thomas Cartwright, the Puritans of the Bay Colony fashioned the civil commonwealth according to the framework of the church. Only the elect could vote and rule in the commonwealth. The church was not itself to govern, but it was the means through which were prepared "instruments both to rule and to choose rulers." Biblical law was the primary law for the ordering of both church and state.

The colony prospered; thus it seemed evident that God was blessing Puritan performance. As a result, the leadership could not take kindly to those who were publicly critical of their basic program. Hence, Roger Williams in 1635 and Anne Hutchinson in 1638 were banished from the colony in spite of their ability to declare experience of the work of grace.

More troublesome than these dissenters were persons such as Mary Dyer, who followed Anne Hutchinson across the border and who later became a Quaker and then returned to Massachusetts Bay. She and other Quakers who returned again and again after being punished and banished were finally hanged. It was difficult for the state to keep the church pure.

The Cambridge Platform and the Half-Way Covenant

In order to head off a possible new form of church government dictated from England at the time of the Westminster Assembly, churches from the four Puritan colonies of Massachusetts Bay, Plymouth, Connecticut, and New Haven met in a voluntary synod in 1648. They adopted the Cambridge Platform, in which the congregational form of church government was worked out in detail. The standard for church membership came under question when it was found that numbers of second generation residents could not testify to the experience of grace in their lives. This resulted in the Half-Way Covenant of 1657 and 1662 that permitted baptized, moral, and orthodox persons to share in the privileges of church membership except for partaking of communion.

Late in the 17th century it was apparent to all that the ideal commonwealth was not being maintained. Ministers pointed to wars with the Indians and other problems as signs of God's judgment. Visitation by demonic powers in the form of witches was believable to people expecting the wrath of God. The Salem witchcraft trials and hangings took place in 1692 at a period of declining confidence in the old ideal.

Other colonies. Massachusetts Bay, Plymouth, Connecticut, and New Haven were variations on the main theme of realizing the Holy Commonwealth in America. Roger Williams and the other founders of Rhode Island must also be regarded as Puritans with the "one principle, that every one should have liberty to worship God according to the light of their consciences."

William Penn's "holy experiment" in Pennsylvania represented another Puritan variation, only this time under Quaker norms. When Penn came into the ownership of this vast tract of land he saw it as a mandate from God to form an ideal commonwealth. In New Jersey, Puritans from the New Haven colony who were dissatisfied with the Half-Way Convenant sought to re-establish the pristine Puritan community at Newark. Maryland, which had been established under Roman Catholic auspices, soon had a strong Puritan majority among its settlers.

There was no colony in which the Puritan influence was not strong in one form or another. One estimate is that 85 percent of the churches in the original 13 colonies were Puritan in spirit.

BELIEF, WORSHIP, AND PRACTICES

Belief. Puritan belief developed from the manner in which William Tyndale, Thomas Cranmer, and other early English Protestants sought to be faithful to the Bible. England's first Protestants were humanists, in the tradition of Desiderius Erasmus, before they became acquainted with the work of Martin Luther. They fitted the humanist concern for the "discipline of Christ" into their adaptation of Protestant theology. Huldrych Zwingli, Martin Bucer, Heinrich Bullinger, and Pietro Martire Vermigli had great influence on Puritan thought. In the last decade of the 16th century John Calvin's influence reached its peak among Puritans. They systematized Calvin's ideas and put them into the context of their English theological tradition to produce a "covenant theology." As John Robinson, the pastor to the Pilgrim Fathers, stated, Puritans did not "stick fast where they were left by that great man of God [Calvin], who yet saw not all things," for "the Lord hath more truth yet to break forth out of his holy Word."

Influence of continental reformers

Worship. Puritans sought to purify worship according to a biblical model, eliminating anything that suggested idolatry. Central to their worship was the preaching of the Word, prayers, and the singing of Psalms. Plain preaching that involved explicating the meaning of a biblical text and then applying it to the congregation was emphasized. Sermons lasting from one to two hours were often given both in the morning and the afternoon on Sundays and also during the week, especially on market days.

Extemporaneous prayers were preferred, and these tended to be much longer than those in *The Book of Common Prayer.* To safeguard the biblical character of worship, the Psalms became the basic hymnbook for Puritans. Many paraphrases of the Psalms were published, including the *Bay Psalm Book,* the first book printed in the English colonies of America.

The *Bay Psalm Book*

The sacrament of the Lord's Supper was celebrated frequently in some churches and more rarely in others. There was a great concern that no unworthy person be admitted. They celebrated the sacrament around a table rather than before an altar.

Daily family worship was practiced, with the pastor's household providing the example for his people.

Ecclesiastical polity. Puritans experimented with many types of episcopal, presbyterian and congregational forms of church organization. They went from a more functional approach in the early days of Elizabeth, where the primary concern was the workability of the government, to a concern for finding and reduplicating the biblical pattern. Yet, even at the end of the Puritan period in 1690, one finds Functionalists, such as Richard Baxter, who saw truth in each form of church government but whose greatest concern was to find a working formula. In almost all cases, Puritan church government on the local level consisted of a pastor governing along with lay leaders elected by the congregation. Any church structure above the congregational level was maintained for the welfare of the local congregations.

The maintenance of discipline. The purpose of church government was to ensure that true doctrine was proclaimed, proper worship was performed, and that the discipline was maintained. The key to the Puritan understanding of discipline is Matt. 18:15–17, which deals with the admonition to confront fellow believers with their faults and if they are unresponsive to reject them. Separatists sought to maintain this discipline only among members of their congregations and to use persuasion upon the rest of mankind. The more conservative Puritans believed that it was their duty to maintain the discipline in the whole community.

Puritans sought to bring order out of the disorder that the "Fall of Adam" had brought to all mankind. Their, degree of achievement was indicated by events interpreted as signs of God's blessing or judgment. Individual Puritans kept diaries by which they could measure their personal "pilgrim's progress," and Puritan preachers in their sermons assessed the state of the community.

Ministers. Since the interpretation and proclamation of the Word of God was central to the laws, the life, and the very sense of meaning in the community, the ministers of the Word were of paramount importance. Through them the congregation and the community came to know who they were, where they were, and where they were going. It was desirable that a minister be highly educated in his vocation by becoming "conversant with Hebrew and Greek and in such arts and sciences as are handmaids unto divinity."

Education of ministers

The schools and universities of England were shaped to produce such men first by the humanists and then by their Protestant successors. The English models were brought to America by the Massachusetts Bay colonists who founded the Boston Latin School, Harvard College, and the rudiments of the American public school system within ten years of their landing.

"Blue laws." The attempt to enforce discipline upon the whole community resulted in the production of "blue laws," so named from some laws bound in blue paper in the New Haven colony. The Puritans were not so different in their moral ideals from Roman Catholics, Lutherans, Jews, and the adherents of other faiths for whom the biblical injunctions were normative. The Puritans, however, tried harder to enforce the Ten Commandments and other precepts.

Many laws had to do with observing the sabbath commandment. Sabbatarianism was not peculiar to Puritans, for the Anglican Lancelot Andrewes recommended the same. Some Puritan sermons, however, seemed to take the position that if the sabbath were observed rightly just once, the Kingdom of God would come.

Missionary activities. One reason for establishing colonies in America was the conversion of Indians. John Eliot pioneered in this endeavour in Massachusetts by translating the Bible into an Indian language. His translation was the first Bible printed in North America. Though he had considerable success in making converts, his work was frustrated by the demand for land by settlers who pushed the Indians back until a last desperate attempt to forestall the inevitable took place in King Philip's War of 1675. That war scattered Eliot's Indian congregations and communities. Great interest in his work had been evidenced among Puritans in England who organized a missionary society to raise money for his cause.

INFLUENCES OF PURITANISM

The original Puritans had committed themselves to the authority of the Word of God without any intervening human agency. They had expected to produce religious uniformity, but they actually contributed to greater diversity. The logic of their position eventually resulted first in an attempt to accommodate variety, then in religious toleration, and finally in religious freedom.

Contributions to religious freedom and the development of democracy

The contribution of Puritanism to the development of democracy followed a similar course. Their theology of election seemed to indicate an aristocracy of the chosen of God, rather than democracy. Those who were the elect of God were all on the same level, however, and the rule of the saints in New England was democratic as far as the saints themselves were concerned. When men became less sure of their ability to distinguish the true saints from others, they extended the franchise until eventually the democracy of the saints became the democracy of the community.

Fear of the tendency to sin that was in every man encouraged the view that all power on earth be limited. Puritan influence can thus be seen in the checks and balances that were written into the Constitution of the United States. To counteract sin, Puritan institutions of home, family, and church were all schools for character that produced individuals with a strong enough sense of personal morality and civic responsibility to make constitutional democracy workable.

Traces of the Puritan vision live on in the basic commitments of the United States. Thomas Jefferson's Latin motto on the reverse side of the one-dollar bill reminds Americans that their nation is the foundation of that "new order of the ages" for which men have hoped. The second inaugural address of Abraham Lincoln resounds with the Puritan theme of God's judgment and grace to be seen in human events. When Dwight D. Eisenhower reminded Americans that their nation was based upon a deeply felt religious faith and that he personally did not care whether one was Protestant, Catholic, or Jew, Oliver Cromwell would have understood.

BIBLIOGRAPHY

English Puritanism: W. HALLER, *The Rise of Puritanism* (1938, pa. 1957), and its sequel *Liberty and Reformation in the Puritan Revolution* (1955, pa. 1963), on English Puritanism from 1570 to 1648; P. COLLINSON, *The Elizabethan Puritan Movement* (1967), definitive treatment of the Puritans during the reign of Queen Elizabeth; C. HILL, *Society and Puritanism in Pre-Revolutionary England* (1964), a discussion of the nontheological factors encouraging the rise of Puritanism in England; M.L. WALZER, *The Revolution of the Saints: A Study in the Origins of Radical Politics* (1965), an analysis of Puritanism as a response of anxious men to change and disorder.

American Puritanism: Seminal scholarly studies by P.G.E. MILLER are *Orthodoxy in Massachusetts, 1630–1650: A Genetic Study* (1933), *The New England Mind:* vol. 1, *The Seventeenth Century*, vol. 2, *From Colony to Province* (1961), and *Errand into the Wilderness* (1956, pa. 1964). Two additional works on this subject are E.S. MORGAN, *Visible Saints: The History of a Puritatt Idea* (1963), a description of Puritan attempts to realize their ideals; and A. SIMPSON, *Puritanism in Old and New England* (1955, pa. 1961), a history of the struggle of Puritans to preserve their authority in society.

(J.C.S.)

Pushkin, Aleksandr

The greatest Russian poet, Aleksandr Pushkin (1799–1837) was also the founder of his country's modern literature. His use of language, astonishing in its simplicity and profundity, formed the basis of the style of novelists Ivan Turgenev, Ivan Goncharov, and Leo Tolstoy. His novel in verse, *Yevgeny Onegin*, the first Russian work to take contemporary society as its subject, pointed the way to the Russian realistic novel of the 19th century.

LIFE

Pushkin was born in Moscow on June 6 (May 26, old style), 1799. His father came of an old boyar family; his mother was a granddaughter of Abram Hannibal, who, according to family tradition, was an Abyssinian princeling bought as a slave at Constantinople and adopted by

Peter the Great, whose comrade in arms he became. Pushkin immortalized him in an unfinished historical novel, Arap Petra Velikogo (published 1837; The Negro of Peter the Great).

By courtesy of the State Tretyakov Gallery, Moscow

Pushkin, oil painting by Orest Kiprensky. 1827. In the State Tretyakov Gallery, Moscow.

Early years. Like many aristocratic families in early 19th-century Russia, Pushkin's parents adopted French culture, and he and his brother and sister learned to talk and to read in French. They were left much to the care of their maternal grandmother, who told Aleksandr, especially, stories of his ancestors in Russian. From Arina Rodionovna, his old nurse, a freed serf (immortalized as Tatyana's nurse in Yevgeny *Onegin*), he heard Russian folktales. During summers at his grandmother's estate near Moscow, he talked to the peasants and spent hours alone, living in the dream world of a precocious, imaginative child. He read widely in his father's library and gained stimulus from the literary guests who came to the house—Vasili Pushkin, his father's brother, a minor poet; I.I. Dmitriev; N.M. Karamzin; V.A. Zhukovsky; and K.N. Batyushkov—young writers opposed to the prevailing French Classicism.

First published work

In 1811 Pushkin entered the newly founded Imperial Lyceum at Tsarskoye Selo (later renamed Pushkin) and while there began his literary career with the publication (1814, in Vestnik Evropy, "The Messenger of Europe") of his verse epistle "To My Friend, the Poet." In his early verse, he followed the style of his older contemporaries, the Romantic poets Batyushkov and Zhukovsky, and of the French 17th- and 18th-century poets, especially the Vicomte de Parny, and their tradition of light verse.

While at the Lyceum he also began his first completed major work, the romantic poem Ruslan and Lyudmila (published 1820), written in the style of the narrative poems of Ariosto and Voltaire but with an old Russian setting and making use of Russian folklore. Ruslan, modelled on the traditional Russian epic hero, encounters various adventures before rescuing his bride, Lyudmila, daughter of Vladimir, grand prince of Kiev, who, on her wedding night, has been kidnapped by the evil magician Chernomor. The poem flouted accepted rules and genres and was violently attacked by both of the established literary schools of the day, Classicism and Sentimentalism. It brought Pushkin fame, however, and Zhukovsky presented his portrait to the poet with the inscription "To the victorious pupil from the defeated master."

St. Petersburg. In 1817 Pushkin left the Lyceum and, accepting a post in the foreign office at St. Petersburg, plunged into social life. He was elected to the exclusive Arzamás, a literary society founded by his uncle's friends, and also became an active member of the liberating movement that had begun among the progressive aristocracy as a result of the upsurge of patriotism after the Napoleonic invasion of 1812.

Pushkin joined the Green Lamp association, which, though founded (in 1818) for discussion of literature and history, became a clandestine branch of a secret society, the Union of Welfare. In his political verses and epigrams, widely circulated in manuscript; in the ode to liberty "Volnost" (1817); and in "Derevnya" (1819: "The Village"), written under the influence of the 18th-century radical thinker and writer A.N. Radishchev, he made himself the spokesman for the ideas and aspirations of those who were to take part in the Decembrist rising of 1825, the unsuccessful culmination of a Russian revolutionary movement in its earliest stage.

Exile in the south. For these political poems, in May 1820 Pushkin was banished from St. Petersburg to a remote southern province. Sent first to Yekaterinoslav (now Dnepropetrovsk), he was there taken ill and, while convalescing, travelled in the northern Caucasus and later to the Crimea with General Rayevski, a hero of 1812, and his family. The impressions he gained provided material for his "southern cycle" of romantic narrative poems.

Influence of Byron

The "*southern* cycle." Like many western European writers, Pushkin had fallen under the spell of Byron's poetry, and, as he himself said, his southern poems "smack of Byron." They introduced Byronic Romanticism to Russia. But even in the first of them, *Kavkazski* plennik (1822; *The* Prisoner of the Caucasus), realist tendencies are discernible. Pushkin here creates from his own experience a psychologically exact portrait of a typical representative of the rising generation in Russia, who, disappointed in love and friendship and dissatisfied with social life in the capital, seeks freedom in the primitive beauty of the Caucasus and in the simple life of its inhabitants, untainted by "civilization." Taken prisoner by the Circassians and finally liberated by a Circassian girl who loves him, he is unable to respond to the passion of this "maid of the mountains" because at heart he is cold and prematurely aged; and she, in despair, throws herself into a mountain torrent and is drowned.

In the second poem of the cycle, *Bratya razboiniki* (1821–22, published 1827; The Robber Brothers), a work permeated with passionate ardour for freedom and based on an event that took place during his stay at Yekaterinoslav, he describes the courageous escape from prison of two brothers, who, although chained and fettered, had swum across the Dnieper and got away.

Bakhchisaraisky *fontan* (1821–23, published 1824; *The* Fountain of Bakhchisarai) is based on a legend. Girei, a warlike nomad, khan of the Crimea, conceives a deep and pure love for his captive, the Polish princess Maria Potocka; but one of his wives murders her out of jealousy. In her memory, Girei erects a fountain (which Pushkin had seen in the khan's palace) surrounded by marble basins into which tearlike drops of water fall melodiously.

Yevgeny *Onegin.* Although this "southern cycle" of poems confirmed the reputation of the author of Ruslan and Lyudmila and Pushkin was hailed as the leading Russian poet of the day and as the leader of the romantic, liberty-loving generation of the 1820s, he himself was not satisfied with it. In May 1823 he started work on his central masterpiece, the novel in verse, Yevgeny *Onegin* (published 1833), on which he continued to work intermittently until 1831. In it he returned to the idea (which had first found artistic expression in *The* Prisoner of the Caucasus) of presenting a typical figure of his own age but in a wider setting and by means of new artistic methods and techniques.

Yevgeny *Onegin* unfolds a panoramic picture of Russian life. The characters it depicts and immortalizes—Onegin, the disenchanted skeptic; Lensky, the romantic, freedom-loving poet; and Tatyana, the heroine, a profoundly affectionate study of Russian womanhood: a "precious ideal," in the poet's own words—are typically Russian and are shown in relationship to the social and environmental forces by which they are molded. Although formally the work resembles Don Juan, Pushkin rejects Byron's subjective, romanticized treatment in favour of objective description and shows his hero not in exotic surroundings but at the heart of a Russian way of life. Thus, the action begins at St. Petersburg, continues on a provincial estate, then switches to Moscow, and finally returns to St. Petersburg.

At Kishinyov and Odessa. Pushkin had meanwhile been transferred first to Kishinyov (1820–23) and then to Odessa (1823–24). His bitterness at continued exile is expressed in letters to his friends—the first of a collection of correspondence that became an outstanding and enduring monument of Russian prose. At Kishinyov, a remote outpost in Moldavia, he devoted much time to writing, though also he plunged into the life of a society engaged in amorous intrigue, hard drinking, gaming, and violence. At Odessa he made passionate love to the wife of his superior, Count Vorontsov, governor general of the province. He fought several duels, and eventually the count asked for his discharge. Pushkin, in a letter to a friend intercepted by the police, had stated that he was now taking "lessons in pure atheism." This finally led to his being again exiled to his mother's estate of Mikhaylovskoye, near Pskov, at the other end of Russia.

Russian features of his poetry

At Mikhaylovskoye. Although the two years at Mikhaylovskoye were unhappy for Pushkin, they were to prove one of his most productive periods. Alone and isolated, he embarked on a close study of Russian history; he came to know the peasants on the estate and interested himself in noting folktales and songs. During this period the specifically Russian features of his poetry became steadily more marked. His ballad "Zhenikh" (1825; "The Bridegroom"), for instance, is based on motifs from Russian folklore; and its simple, swift-moving style, very different from the brilliant extravagance of *Ruslan and Lyudmila* or the romantic, melodious music of the "southern" poems, emphasizes its stark tragedy.

In 1824 he completed *Tsygane* (published 1827; *The Gypsies),* begun earlier as part of the "southern cycle." In this, the most mature of his verse romances, he puts into the mouth of an old gypsy, a representative of the people, a condemnation of the individual romantic hero who wants freedom "only for himself." The style is dramatically harsh and spare, and the descriptive passages and dialogue are vigorous and realistic. At Mikhaylovskoye, too, he wrote the provincial chapters of *Yevgeny Onegin;* the poem *Count Nulin* (1825, published 1827), based on the life of the iural gentry; and, finally, one of his major works, the historical tragedy *Borir Godunov* (1824–25, published 1831).

His debt to Shakespeare

Boris Godunov. This tragedy marks a break with the Classicism of the French theatre and is constructed on the "folk-principles" of Shakespeare's plays, especially the histories and tragedies, plays wiitten "for the people" in the widest sense and thus universal in their appeal. Written just before the Decembrist rising, it treats the burning question of the relations between the ruling classes, headed by the tsar, and the masses; it is the moral and political significance of the latter, "the judgment of the people," that Pushkin emphasizes. Set in Russia in a period of political and social chaos on the brink of the 17th century, its theme is the tragic guilt and inexorable fate of a great hero—Boris Godunov, son-in-law of Malyuta Skuratov, a favourite of Ivan the Terrible, and here presented as the murderer of Ivan's little son, Dmitri. The development of the action on two planes, one political and historical, the other psychological, is masterly and is set against a background of turbulent events and ruthless ambitions. The play owes much to Pushkin's reading of early Russian annals and chronicles, as well as to Shakespeare, who, as Pushkin said, was his master in bold, free treatment of character, simplicity, and truth to nature. Although lacking the heightened, poetic passion of Shakespeare's tragedies, *Boris* excels in the "convincingness of situation and naturalness of dialogue" at which Pushkin aimed, sometimes using conversational prose, sometimes a five-foot iambic line of great flexibility. The character of the pretender, the false Dmitri, is subtly and sympathetically drawn; and the power of the people, who eventually bring him to the throne, is so greatly emphasized that the play's publication was delayed by censorship. Pushkin's ability to create psychological and dramatic unity, despite the episodic construction, and to heighten the dramatic tension by economy of language, detail, and characterization make this play an outstanding achievement and a revolutionary event in the history of Russian drama.

Return from exile. After the suppression of the rising of Dec. 14 (O.S.), 1825, the new tsar Nicholas I, aware of Pushkin's immense popularity and knowing that he had taken no part in the Decembrist "conspiracy," allowed him to return to Moscow in the autumn of 1826. During a long conversation between them, the Tsar met the poet's complaints about censorship with a promise that in the future he himself would be Pushkin's censor and told him of his plans to introduce several pressing reforms from above and, in particular, to prepare the way for liberation of the serfs. The collapse of the rising had been a grievous experience for Pushkin, whose heart was wholly with the "guilty" Decembrists, five of whom had been executed, while others were exiled to forced labour in Siberia. Among them had been his closest friends at the Lyceum, I.I. Pushchin and the poet V.K. Kyukelbeker (see the poem "Arion" and the epistle to Siberia, "Vo glubine sibirskikh rud . . ." ["In the depths of Siberia's mines . . ."], both written in 1827). Thus, when the Tsar asked him outright how he would have acted had he been in St. Petersburg on December 14, he answered that he would have joined the ranks of the insurgents.

Pushkin's interest in reforms

Pushkin saw, however, that without the support of the people, the struggle against autocracy was doomed. He considered that the only possible way of achieving essential reforms was from above, "on the tsar's initiative," as he had written in "Derevnya." This is the reason for his persistent interest in the age of reforms at the beginning of the 18th century and in the figure of Peter the Great, the "tsar-educator," whose example he held up to the present tsar in the poem "Stansy" (1826; "Stanzas"), in *The Negro of Peter the Great,* in the historical poem *Poltava* (1828, published 1829), and in the poem *Medny vsadnik* (1833, published 1837; *The Bronze Horseman).* In this last, Pushkin poses the problem of the "little man" whose happiness is destroyed by the great leader in pursuit of ambition, by telling a "story of St. Petersburg" set against the background of the flood of 1824, when the river took its revenge against Peter I's achievement in building the city. The poem describes how the "little hero," Yevgeny, driven mad by the drowning of his sweetheart, wanders through the streets. Seeing the bronze statue of Peter I seated on a rearing horse and realizing that the Tsar, seen triumphing over the waves, is the cause of his grief, Yevgeny threatens him and, in a climax of growing horror, is pursued through the streets by the "Bronze Horseman." Its descriptive and emotional powers give the poem an unforgettable impact and makes it one of the greatest in Russian literature.

After returning from exile, Pushkin found himself in an awkward and invidious position. The Tsar's censorship proved even more exacting than that of the official censors, and his personal freedom was curtailed. Not only was he put under secret observation from the police but he was openly supervised by its chief, Count Benckendorf. When in 1829, during the Russo-Turkish War, after his applications to go to Transcaucasia had been refused, he managed to visit the front lines without permission, for which he was severely reprimanded by Count Benckendorf. This visit, on which he met several of the exiled Decembrists, is described in *Puteshestviye v Arzrum* (published 1836; *A Journey to Erzuıum).*

Reaction of critics and friends

Moreover, his works of this period met with little comprehension from the critics, and even some of his friends accused him of apostasy, forcing him to justify his political position in the poem "Druzyam" (1828; "To My Friends"). The anguish of his spiritual isolation at this time is reflected in a cycle of poems about the poet and the mob (1827–30) and in the unfinished *Egipetskiye nochi* (published 1837; *Egyptian Nights).*

Yet it was during this period that Pushkin's genius came to its fullest flowering. His art acquired new dimensions, and almost every one of the works written between 1829 and 1836 opened a new chapter in the history of Russian literature. He spent the autumn of 1830 at his family's Nizhny Novgorod (now Gorky) estate Boldino, and these months are the most remarkable in the whole of his artistic career. During them he wrote the four *malenkiye tragedii* ("little tragedies"): *Skupoy rytsar* (1836; *The Covetous*

Knight); Mozart i Salieri (1831; Mozart and Salieri); Kamenny *gost* (1839; The Stone Guest); Pir vo vremya chumy (1832; Feast in Time of the Plague); and the five short prose tales collected as Povesti pokoynogo Z.P. Belkina (1831; Tales of the Late *I.P.* Belkin); the comic poem of everyday lower class life Domik v Kolomne (1833; A Small House in Kolomna); and many lyrics in widely differing styles; as well as several critical and polemical articles, rough drafts, sketches, and so on.

One of Pushkin's most characteristic features was his wide knowledge of world literature, shown in particular by his interest in English literature from Shakespeare and Byron to Sir Walter Scott and the Lake poets, in Dostoyevsky's phrase, his "universal sensibility," his ability to re-create the spirit of different races at different historical epochs without ever losing his own individuality. This is particularly marked in the "little tragedies," which are concerned with an analysis of the "evil passions" and, like Pikovaya *Dama* (1834; The Queen of Spades), exerted a direct influence on the subject matter and techniques of the novels of Dostoyevsky.

Last years. In 1831 Pushkin, after a turbulent courtship and objections from her mother, married Natalya Nikolayevna Goncharova and settled in St. Petersburg. Once more he took up government service and was commissioned to write a history of Peter the Great. Three years later he received the rank of Kammerjunker (gentleman of the emperor's bedchamber), partly because the Tsar wished Natalya to have the entree to court functions. The social life at court, which he was now obliged to lead and which his wife enjoyed, was ill suited to creative work, but he stubbornly continued to write. Without abandoning poetry altogether (the fairy tales in verse, skazki, belong to this period, as well as Medny vsadnik, or The Bronze Horseman), he turned increasingly to prose. Alongside the theme of Peter the Great, the motif of a popular peasant rising acquired growing importance in his work, as is shown by the unfinished satirical Istoriya sela Goryukhina (1830, published 1837; History of the Village of Goryukhino); the novel Dubrovsky (1832–33, published 1841); Stseny *iz* rytsarskikh vremen (1835, published 1837; dramatic Scenes from the Age of Chivalry); and finally, the most important of his prose works, the historical novel of the Pugachov Rebellion, *Kapitan*skaya dochka (1833–36, published 1836; The Captain's Daughter), which had been preceded by a historical study of the rebellion, Istoriya Pugachova (1833, published 1834).

Prose writings

Meanwhile, both in his domestic affairs and in his official duties, his life was becoming more intolerable. In court circles he was regarded with mounting suspicion and resentment, and his repeated petitions to be allowed to resign his post, retire to the country, and devote himself entirely to literature were all rejected. Finally, on February 8, 1837, in a duel forced on him by influential enemies, defending his wife's honour, Pushkin fell, mortally wounded. His adversary, Georges d'Anthès, an officer in the guards and the adopted son of the Dutch ambassador Baron Heeckeren, had emigrated from France after the revolution of 1830 and was the husband of Natalya's sister. Pushkin died of his wound at St. Petersburg on February 10 (January 29, O.S.), 1837.

ASSESSMENT

Even during his lifetime Pushkin's importance as a great national poet had been recognized by Nikolay Vasilyevich Gogol, his successor and pupil; and it was his younger contemporary, the great Russian critic, democrat, and revolutionary Vissarion Grigoryevich Belinsky, who produced the fullest and deepest critical study of Pushkin's work, which still retains much of its relevance. To the later classical writers of the 19th century, Pushkin, the creator of the Russian literary language, the author of the standard works of Russian literature, the "poet of reality" (as he said of himself), stands as the cornerstone of Russian literature, in Maksim Gorky's words "the beginning of beginnings."

Pushkin has become an inseparable part of the cultural and spiritual world of the Russian people and of the other peoples of the former tsarist empire to whom he bequeathed his work (in the poem "Ya pamyatnik sebye vozdvig" [1836; "Exegi monumenturn"]). He has exerted, too, a profound influence on other aspects of Russian culture: most notably, in opera.

Pushkin's work—with its nobility of conception and its emphasis on civic responsibility (shown in his command to the poet-prophet to "fire the hearts of men with his words"); its life-affirming vigour; and its confidence in the triumph of reason over prejudice, of light over darkness, of human charity over slavery and oppression—has struck an echo all over the world. Translated into all the major languages, his works are regarded both as expressing most completely Russian national consciousness and as transcending national barriers.

MAJOR WORKS

Poetic works

NARRATIVE POEMS: *Ruslan i* Lyudmila (1820; *Ruslan* and Lyudmila), a mock-heroic folk epic. The "southern cycle": Kavkazsky plennik (1822; The Prisoner of the Caucasus); Bratya razboiniki (written 1821–22, published 1827; The Robber Brothers); Bakhchisaraysky *fontan* (1824; The Fountain of Bakhchisaray); Tsygany (1827; The Gypsies); begun as part of the "southern cycle" but completed at Mikhaylovskoye. (EPIC POEMS): *Poltava* (1829; Eng. trans.): Medny vsadnik (1837; The Bronze Horseman). (HUMOROUS EPICS): Graf Nulin (1827; Count Nulin); "Tazit," or "Galub" (published posthumously 1837); Domik v Kolomne (1833; A Small House in Kolomna). (NOVEL IN VERSE): Yevgeny *Onegin* (written 1823–31, published complete 1833; Eugene *Onegin*). (BALLADS): "Kazakh (1815; "The Cossack"); "Zhenikh (1825; "The Bridegroom"). (FAIRY TALES IN VERSE): "Skazka o pope i o rabotnike yego Balde" (1840; "The Tale of the Priest and His Helper Bald"); "Skazka o tsare Saltane" (1832; "The Tale of Tsar Saltan"); "Skazka o mertvoy tsarevne" (1834; "The Tale of the Dead Princess"); "Skazka o rybake i rybke" (1835; "The Tale of the Fisherman and the Fish); "Skazka o zolotom petushke" (1835; "The Tale of the Golden Cockerel").

POLITICAL POEMS: (circulated in manuscript, all posthumously published): "Volnost" (written 1817; "Ode to Freedom"); "Skazki Noel" (1818); "Derevnya" (written 1819; 'The Village"; "In the Country"); "Kinzhal" (1821); "Vo glubine Sibirskikh rud . . ." (written 1827; "In the depths of Siberia's mines . . ."; "Message to Siberia"); "Druzyam" (written 1828; "To My Friends").

Dramatic works

Boris Godunov (written 1824–25, published 1831; Eng. trans.), historical tragedy in blank verse and prose. The "little tragedies" (malenkiye tragedii), written 1830: Skupoy rytsar (1836; The Covetous Knight); Kamenny *gost* (1839; The Stone Guest); Pir vo vremya chumy (1832; Feast in Time of the Plague); and Mozart i Salieri (1831; Mozart and Salieri), short dramatic episodes in blank verse; Rusalka (published posthumously 1837); Stseny iz rytsarskikh vremen (published posthumously 1837; Scenes from the Age of Chivalry).

Prose works

SHORT STORIES: *Povesti pokoynogo* I.P. Belkina (1831; Tales of the Late I.P. Belkin); Pikovaya *Dama* (1834; The Queen of Spades).

NOVELS: Kapitanskaya dochka (1836; The Captain's Daughter); Arap Petra Velikogo (1837; The Negro of Peter the Great); Egipetskiye *nochi* (1837; Egyptian *Nights*); Dubrovsky (1841; Eng. trans.).

OTHER PROSE: *Puteshestviye* v *Arzrum* (1836; *A* Journey to *Erzurum*); Istoriya Pugachova (written 1833, published 1834).

English translations

The most comprehensive selection is in A. Yarmolinsky (ed.), The Works of Alexander *Pushkin* (1939, new ed. 1946). There are translations of selected poems, ballads, and fairy tales by Maurice Baring in Poems Translated from *Pushkin* (1931); by Oliver Elton in Verse from *Pushkin* and Others (1935); and by Walter Morison in Pushkin's Poems (1945; shorter poems only). Selected fairy tales are translated by B.L. Brasol in The Russian Wonderland (1936). The Little Tragedies (1946) have been translated by V. de S. Pinto and W.H. Marshall. Translations of Evgeni *Onegin* include those by O. Elton (1937; rev. ed. 1948); D. Paull Radin and G.Z. Patrick (1937); V. Nabokov (4 vol., 1964); and B. Deutsch (1965). Thomas Keane, The Prose Tales of *Pushkin* (1894), has been superseded by G. Aitken, The Complete Prose Tales of *Pushkin* (1966). There is a translation of the novels and some stories by N. Duddington in The Captain's Daughter and Other Tales (1933, rev. ed. 1961).

BIBLIOGRAPHY. The most detailed biographical and critical studies of Pushkin in English are: D.S. MIRSKY, *Pushkin* (1926, reprinted 1963); ERNEST J. SIMMONS, *Pushkin* (1937, reprinted 1964); and "Pushkin: The Poet As Novelist," in *An Introduction to Russian Realism* (1965); DAVID MAGARSHACK, *Pushkin: A Biography* (1967); WALTER N. VICKERY, *Pushkin: Death of a Poet* (1968); V.V. VERESAYEV, *Pushkin: A Biographical Sketch* (1937; orig. pub. in Russian, 1936); IRAKLY ANDRONIKOV, *The Last Days of Pushkin* (Eng. trans. 1957), an essay on Pushkin's duel and death, written after the discovery of new materials; JANKO LAVRIN, *Pushkin and Russian Literature* (1947, reprinted 1969), on Pushkin's place and significance in the development of Russian literature; and SAMUEL H. CROSS and ERNEST J. SIMMONS (eds.), *Centennial Essays for Pushkin* (1937, reprinted 1967), a collection of essays published on the centenary of Pushkin's death, including Simmons's "Biographical Study of Pushkin"; GEORGE V. VERNADSKY, "Pushkin and the Decembrists"; and ARTHUR P. COLEMAN, "Pushkin and Mickiewicz."

There are thousands of books and articles in Russian that cover all aspects of Pushkin's life and works. Б.П. ГОРОДЕЦКИЙ, Н.В. ИЗМАЙЛОВ, and Б.С. МЕЙЛАХ (eds.), *Пушкин: Итоги и проблемы изучения* (1966), is an excellent collection of articles on Pushkin by 14 Russian authorities. The fundamental biographies in Russian are НИКОЛАЙ ЛЕОНТЬЕВИЧ БРОДСКИЙ, *А.С. Пушкин, биография* (1937); and ЛЕОНИД ПЕТРОВИЧ ГРОССМАН, *Пушкин*, 3rd ed. (1960).

(D.D.B.)

Pym, John

The English statesman John Pym, the architect of Parliament's victory over Charles I in the Civil War (1642–46), laid the foundations of the parliamentary sovereignty that prevailed in England from the 17th century onward. Pym also was largely responsible for the system of taxation that survived in England until the 19th century and for the enduring close relations between the English government and the City of London.

By courtesy of the trustees of the British Museum

Pym, engraving by G. Glover, 1644, after a portrait by Edward Bower.

Pyrn was born in 1583 or 1584, the eldest son of Alexander Pym of Brymore, Somerset, who died when John was a child; his mother married Sir Anthony Rous, a client of the Russells, the earls of Bedford. Pyrn was educated at Oxford but took no degree, and at the Middle Temple, but was not called to the bar. Through Bedford influence he became a local official of the Exchequer. From 1621 to his death Pym sat in every Parliament, usually for the Russell borough of Tavistock. He soon made a name as an enemy of popery and Arminianism (high-church Anglicanism) in high places, as a sound financier, an expert on colonial affairs, and a good committeeman. He was no extremist, however, but a loyal subject anxious to maintain good relations between crown and Parliament.

From 1630 Pym was treasurer of the Providence Island Company, which sought to open trade with Spanish America—peacefully, if possible, by force, if not. In the 11 years from 1629 to 1640, during which the King chose to ıule without Parliament, this company brought together the men, mostly Puritans, who were to lead the Parliamentary party in the 1640s. Opposition to Charles's tax of Ship Money to support the Royal Navy (a tax without parliamentary approval) was organized by adventurers of the company; in August 1640 the petition of 12 peers demanding a Parliament was drafted by Pyrn and another adventurer.

Leader of the middle group in Parliament

When the Long Parliament met in November 1640, Pyrn headed what has been called the middle group, whose central position allowed it to dominate the House of Commons. His policy was that of his patron, the earl of Bedford: to force the King to accept a government in which Parliament, representing the wealth of the country, had confidence. Their main obstacle was Charles's toughest adviser, Thomas Wentworth, 1st earl of Strafford, who was executed as a traitor in May 1641. It was difficult to prove a man deep in the King's confidence a traitor, but Pym argued that "to endeavour the subversion of the laws of this kingdom was treason of the highest nature." Thus, by implication, even a king was capable of committing treason: here was the germ of the charge on which Charles was to be executed in 1649. There were great popular demonstrations in London calling for Strafford's execution, and Pyrn was accused of fomenting them.

Pym forced Charles to accept an act forbidding the dissolution of Parliament without its consent. This was followed by acts abolishing the whole apparatus of personal government and finance. On paper Charles had accepted that he must rule through Parliament, but he had no real intention of accepting this, and he had to be coerced. The main issue became control of the armed forces. When rebellion broke out in Ireland (October 1641), all agreed that it must be crushed; but Parliament rightly feared a military coup if the King were given command of the army. The House of Commons said it would act in Ireland without the King unless Charles changed his ministers. This virtual declaration of revolution was reinforced by the Grand Remonstrance, listing the grievances of the kingdom as Pym's group saw them and demanding ministers trusted by Parliament and an Assembly of Divines nominated by Parliament to reform the church. This Remonstrance, carried by 159 votes to 148, was printed and circulated to rally support outside Parliament; its opponents henceforth formed a Royalist party. Pym was one of five members of Parliament whom Charles tried to arrest in January 1642. They took refuge in the City, from which they returned in triumph when the King quitted London.

Pym's political philosophy

Before and during the Civil War, Pym's political philosophy was summed up in the phrase "I know how to add sovereign to [the King's] person, but not to his power." He believed that the King reigned but did not rule alone: power should be balanced between him and Parliament. "To have printed liberties," Pym once said, "and not to have liberties in truth and realities, is but to mock the kingdom." Pyrn never seems to have contemplated abolishing the monarchy, and he was certainly no democrat; but he used popular pressure to achieve his ends. When war started, he set about creating an army, the machinery to administer it, and the excise and assessment (later the land tax) to pay for it. His City connections helped him to raise loans. He created the network of committees at Westminster and in the counties that ran the country for the next 17 years. When military stalemate threatened, Pyrn called in Scottish assistance even at the price of concessions to Presbyterianism that went further than he wished. Ever pragmatic, when the House of Lords made difficulties, he told them that the Commons could run the country alone.

Pym was a sonorously eloquent speaker, arguing each particular issue from first principles without ever being doctrinaire. His skill as a parliamentary tactician was unrivalled. He balanced between radicals—some of whom were republicans—and the peace party, which was so frightened of social upheaval that it would have accepted almost any terms from the King.

Pym preserved the unity of Parliament for three years with no party organization, no discipline, no whips. This great achievement wore him out, and when he died on December 8, 1643, no one could replace him; Parliament's supporters split into wrangling groups. But by then Pym had laid the basis for victory and had sketched the shape of the England that was to survive the 17th century.

BIBLIOGRAPHY. There is no definitive biography. Those by JOHN FORSTER, *John Pyin* (1837); C.E. WADE, *John Pym* (1912); and S.R. BRETT, *John Pym, 1583–1643: The Statesman of the Puritan Revolution* (1940), are all inadequate. J.H. HEXTER, *The Reign of King Pym* (1941), is the most important modern book, dealing primarily with the period after 1640. A.P. NEWTON, *The Colonising Activities of the English Puritans* (1914), discusses the Providence Island Company.

(J.E.C.H.)

P'yongyang

P'ybngyang, capital and largest city of the Democratic People's Republic of (North) Korea, is the second largest city (after Seoul) of all Korea. P'ybngyang still appears on many maps as Heijo, which is the Japanese pronunciation of the Chinese characters for the name. The name itself means "large flat [*p'yŏng*] field [*yang*]," which describes the plain on which the city was established. Located in the heavily populated northwestern region of North Korea, P'yirngyang is built on both sides of the Taedong-gang (Taedong River) about 30 miles from the Yellow Sea. The area of the metropolitan city (1967) is 77 square miles (199 square kilometres), and its population in 1967 was 1,364,000.

Founding by the Chinese

History. P'ybngyang is reputed to be the oldest city in Korea. The ancient capital of the legendary Tangun dynasty (2333 BC) was located on the site where, according to legend, the present city of P'ybngyang was founded in 1122 BC. The city's recorded history begins in 108 BC with the founding of a Chinese trading colony at Lolang (Nangnang in Korean) near P'ybngyang. A walled fortress lying on the route of march for Chinese or Mongol armies invading Korea and Japanese armies invading China, P'yŏngyang figured prominently in the wars that signalled shifts in power among the nations of northeastern Asia.

In 427 P'ybngyang became the capital of Koguryb, one of the several rival kingdoms that had established themselves in Korea, and in the 7th century it was besieged on several occasions by Chinese invaders. The first of these invasions, in 612, ended in a disastrous defeat for the Chinese army when it was ambushed inside the city. Chinese armies returned in 644 and again in 661 but did not capture the city until 668. The fall of P'ybngyang marked the end of the Koguryb dynasty. Later, the kings of the Kbryo dynasty (918–1392) made P'ybngyang their secondary capital. Japanese invaders attacked and briefly occupied P'ybngyang in 1359. In 1592 the Japanese again attacked. This time the city fell. A year later Chinese armies, as allies of the Koreans, retook the city after an exceptionally bloody battle. Thirty-four years later an army of the Manchus captured P'ybngyang and burned it to the ground before they withdrew.

Attack on the "General Sherman"

The scars of successive invasions left their mark on the Koreans, who became suspicious of all foreigners. In 1866 an American ship, the USS "General Sherman," sailed up the Taedong-gang with the object of opening Korea to foreign trade. Undeterred by Korean demands that it turn back, the "General Sherman" sailed on until it ran aground just opposite P'ybngyang. The Koreans attacked the ship, set it on fire, and killed the survivors who swam ashore. Today the "General Sherman" is to North Koreans a symbol of United States aggression and of Korea's victory. Two five-inch guns from the ship still guard the entrances of two of P'ybngyang's historical museums.

When Korea finally opened its doors to foreigners, P'ybngyang soon became the base of an intensive campaign to bring Christianity to Korea. Over a hundred churches were built in the city, which in the 1880s was reputed to have more Protestant missionaries than any other city in Asia.

During the Sino-Japanese War (1894–95), P'ybngyang was the site of yet another major battle and again suffered great destruction. Plague followed war, and in 1895 P'ybngyang was left a virtually deserted and ruined city. During the Japanese occupation of Korea (1910–45), P'ybngyang was built up as an industrial centre, and its population grew rapidly. The city escaped destruction in World War II.

In the Korean War (1950–53), P'ydngyang was captured by United Nations forces in 1950 and recaptured by Chinese and Korean Communist forces shortly after. During the remainder of the war the city was subjected to intensive United States bombing and almost totally destroyed. Since 1953 the city has been rebuilt with Soviet and Chinese assistance.

The contemporary city. *The city site.* From its original site on the northern bank of the Taedong-gang, the city has expanded in all directions. New residential areas have been built on the southern bank of the river, which is crossed by several bridges; and the river itself is being straightened and made deeper, while trees have been planted on both of its sides. The city's recreational park and an educational centre are located atop a cluster of hills in the city.

Although remnants of the old stone walls and gates can still be found, as a result of the Korean War few buildings are more than 18 years old. Under North Korea's socialist economy, the present city did not simply grow; it was planned. The government has made it a show place with numerous parks and gardens. Row after row of prefabricated concrete apartment houses were erected along wide, tree-lined boulevards.

Climate. P'yŏngyang's proximity to the sea does not protect it from the harsh continental climate of North China and the Northeast (Manchuria). Winters are cold; the average temperature in January is about 18° F (−8° C). Summers are warm, with average temperatures in July and August in the upper 70s and wet. Over half of the total yearly rainfall, which averages 37 inches, falls during these two months.

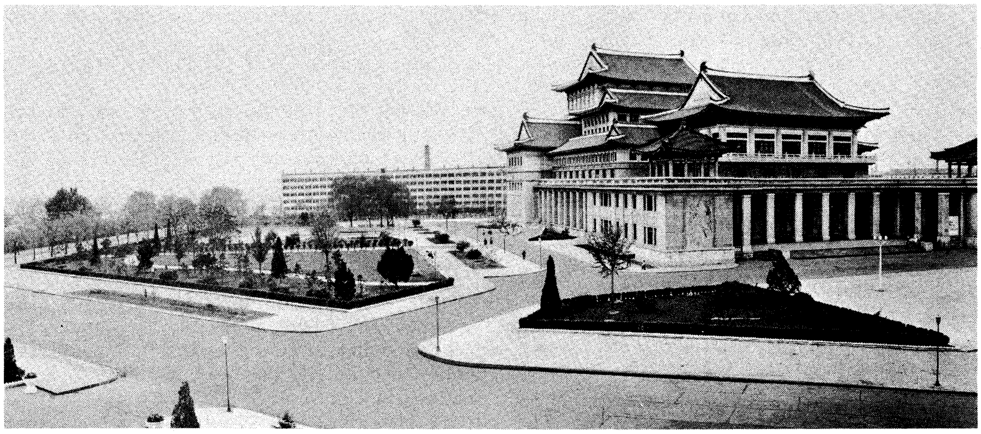
International air and rail connections

Transportation. Invariably on the itinerary for Communist dignitaries and delegations from various revolutionary movements touring Asia, P'ybngyang is normally closed to other visitors. From Sunan airport, ten miles north of the city, flights connect P'yŏngyang with Peking; there are also air connections with Irkustsk, Moscow, and Omsk; and regularly scheduled domestic air service links P'ybngyang with Ch'bngjin on the east coast of North Korea. P'yirngyang is the hub of North Korea's rail system, which is being electrified. There is direct international rail passenger car service with both Moscow and Peking. Ocean vessels call at Namp'o at the mouth of the Taedong-gang on the west coast, and smaller boats on the river provide regular passenger and freight service between that port and P'ybngyang.

Buses provide public transportation within P'ybngyang. Trucks also carry workers to the factories each morning. Only high government officials are allowed private autos. Most people live near the places where they work.

Demography. P'ybngyang's population at the start of the 1970s was estimated to be slightly over 1,000,000 with another 500,000 to 600,000 in the outside metropolitan area. The population density of the city is about 18,000 per square mile. During the later years of the Japanese occupation, thousands of Koreans left the farms to work in the new factories being built by the Japanese to support their war effort. Urbanization continued after the war when thousands more who had been conscripted to work in factories in Japan and Manchuria were repatriated. The returning workers preferred to remain in the cities, although the government tried to get them to accept agricultural jobs in the rural areas.

The growth of P'ybngyang was reversed briefly during the Korean War, as thousands were evacuated from the city to avoid the bombing. At the end of the war, the city, which had more than 340,000 people in 1944, had only



Bolshoi **Theatre,** P'ydngyang.
Tass—Sovfoto

80,000. Since then its population has increased more than tenfold, putting heavy pressure on government construction of urban housing.

Effect of the city on traditional ways of life

City life has had a profound impact on the social structure of the Korean people. Confucian concepts and clan ties have broken down. The typical apartment in the city is far too small to accommodate the extended families of old Korea. Family units have to be smaller, and, as the apartments are assigned by the government, neighbours can no longer claim to have known each other since childhood.

P'yŏngyang, unlike most other Asian cities, does not have a large Chinese minority. North Korea's population is remarkably homogeneous, with only 24,000 resident aliens, mainly Chinese, in the entire country. During the years of Japan's rule, there had been a large resident Japanese population in P'ybngyang, but hardly any remain and intermarriage between the Koreans and Japanese was rare.

Class distinctions are founded not on race but on a person's role in the economy and position in the Communist Party. Housing is assigned according to one's job title. Preference is shown to party members, especially in the purchase of scarce consumer goods.

Housing. The government constructs all urban housing and regulates rents according to the income of the tenants. Tenants are assigned to an apartment and cannot move unless they are promoted to a better apartment. There are four categories of apartments. Workers are entitled to group one housing: a single room with a half-size kitchen. Group four housing, which consists of two rooms, one room with a wood floor, storage room, and a bathroom with toilet, is reserved for such people as department or bureau chiefs in the central government, managers of state enterprises, and college professors. Only top-ranking officials of the government and the Communist Party may have individual houses. Rents are low; an average person reportedly spends no more than 3 percent of his salary on rent, including utilities.

Architecturalfeatures. A huge bronze statue of Ch'bllima, a winged horse of Korean legend, atop a pedestal 150 feet high dominates the skyline of P'ybngyang. According to legend, Ch'ŏllima was able to carry its riders a thousand *li* a day (a *li* is a Chinese unit of measurement that equals about one-third of a mile). To North Koreans, Ch'bllima symbolizes the economic progress made after the end of the Korean War. Sections of the Inner and Northern walls and Hyunmoo Gate are still standing, and several old temples and pavilions that date from the Koguryb kingdom have been reconstructed in the original architectural style. Modern landmarks include the Grand Theatre; the Okryoo Hall, which contains a large banquet hall for official functions and recreation facilities for the workers; and the Moranbong Stadium, which seats 70,000 people. Beneath Moran-bong (Moran Hill), the city's main recreational centre, is a huge underground theatre. A large television antenna marks the northern boundary of the city.

Mills, factories, and commerce

Economy. Located in the midst of North Korea's industrial heartland, P'ybngyang is a major textile and food-processing centre. Coal deposits near the city provide fuel for its factories. Two of the country's largest textile mills are in P'ybngyang: the P'ybngyang Cotton Works, which employs more than 11,000 people, and the P'yŏngyang Silk Works, which employs more than 5,000. While silkworms are raised in North Korea, ginned cotton, raw wool, and rayon yarn used by the mills must be imported. The city also contains sugar refineries, rubber factories, ceramics factories, railroad workshops, and an arsenal.

As the political capital of a thoroughly Socialist country, P'ybngyang is also the country's economic headquarters. An army of bureaucrats is employed in planning and managing the economy. The major commercial institutions are operated by the government. These include the Korean Central Bank, which supervises the flow of all capital; the Korean Industrial Bank, the only bank that grants loans; and the Foreign Trade Bank, which handles all financial settlements resulting from foreign trade.

The People's Assembly and the People's Committee

Government. Urban government in North Korea consists of a People's Assembly and People's Committee. Members of the People's Assembly are locally elected and serve a four-year term. They meet twice a year in regular session to review the work of the People's Committee. Members of the People's Committee are elected by the People's Assembly and are responsible to it. Composed of a chairman, vice chairman, secretary general, and committee members, the People's Committee draws up economic plans and directly supervises the work of the functional bureaus. An important function of the People's Committee is the regulation of prices in the state-owned retail stores. Being the capital, P'ybngyang is a "special city," in which the central government probably plays a larger role than in other directly ruled cities. Below the city government are the people's assemblies and people's committees of the *tong,* or wards, of the city, and the *kun* or counties located on the lightly populated outskirts of P'ybngyang. Below these are people's assemblies and committees for each street or block.

Electricity and water. The city's power supply comes from hydroelectric plants on the Yalu River to the north and a 500,000-kw thermal electric plant in the city. Little power is available to consumers. Few apartment buildings have elevators; electrical appliances are used only by high-ranking officials; the average tenant is limited to the use of two 40-watt light bulbs.

The city's water supply is adequate, but plumbing facilities are not. Even in modern apartment buildings, water may have to be carried up to the apartment by hand. Usually there is one toilet facility on each floor, and outhouses are still in evidence in the backyard of modern buildings.

Public health. Modern hospitals have been built in P'yŏngyang with assistance from the Soviet Union and Communist bloc countries. Medical care is provided by the state at little or no cost. Immunization teams regularly comb the city block by block, immunizing those who cannot show satisfactory shot records.

Education. Education begins in state-run nurseries where working parents deliver their children each morning. This is followed by primary school, middle school, then vocational training, higher technical schooling, or entrance into a university. At the top of the educational system is Kim 11-sung University (founded 1946), located on the outskirts of P'ybngyang overlooking the Taedong-gang. P'ybngyang also has a medical school and a Communist university for training party leaders. Adult education is encouraged through classes taught in the factories.

Cultural life. P'ybngyang is the centre of North Korea's cultural life. It is the home of the P'ybngyang Theatre, where performances are given by the P'ybngyang Song and Dance Troupe, the National People's Opera, and the National Opera; the National Dancing Theatre; and the P'ybngyang Young Students Palace, where children are trained in art, music, and drama. Cultural halls where workers put on simple skits rich in ideological content are located in each district. During the Korean War, dramatic performances were given at the huge underground theatre beneath Moran-bong. The city's museums include the Fatherland Liberation War Memorial Hall, the State Central Historical Museum, the State Central Fine Arts Museum, and the Korean National Folk Museum. The annual National Fine Arts Exhibition is held in P'ybngyang, and the State Central Library is located there.

Communications media

All of North Korea's ten principal newspapers are published in P'ybngyang. These include "Democratic Korea," official paper of the Presidium of the Supreme People's Assembly, and "Labour Daily," put out by the Central Committee of the Korean Workers Party. There are also a number of foreign language newspapers published in P'ybngyang, including, in English, *The Pyongyang Times* (weekly) and *The People's Korea* (daily), but these are intended primarily for distribution overseas. Numerous journals are published by various agencies of the government and mass organizations.

As in other Communist countries in Asia, there are few private wireless radio receivers in North Korea. Radio broadcasting is by direct wire to loudspeakers, which are set up in factories and in open spaces in the towns and cities, and to receivers in private homes. A television broadcasting station has been established in P'ybngyang with Soviet guidance, although in the early 1970s there were only a few thousand television sets in the city.

Leisure time is scarce. Only high-ranking officials have every Sunday off. Workers usually have only one day off every other week and can expect to work overtime to meet demands for higher production. An average day begins at 6:00 AM, with mandatory calisthenics. Fourteen-hour workdays are not uncommon.

The city has a zoo and numerous parks for recreation. Boating on the Taedong-gang is popular in the summer, as is soccer. Winter sports include ice skating, sledding, and skiing.

BIBLIOGRAPHY. Few Western visitors have been allowed to visit North Korea. As a result, there are few sources of information in English that are up to date. Current information must be distilled from North Korean publications and broadcasts or gathered second hand from the few journalists and Japanese businessmen who have visited the city. *Communist North Korea: Bibliographical Survey* (1971), published by the United States Department of the Army, is the most recent guide to other sources in English. Its appendixes contain useful information about government and military organization as well as various statistics. The *Area Handbook for North Korea*, published also by the United States Department of the Army (1969), is the most complete source of information on P'yongyang. Some additional information may be gleaned from the *North Korean Central Yearbook* published annually by the government of North Korea and translated by the Joint Publications Research Service of the United States Department of Commerce. RYU HUN, *Study of North Korea* (1966); and ROBERT A. SCALAPINO (ed.), *North Korea Today* (1963), are useful only for their description of the government. SHANNON MCCUNE, *Korea, Land of Broken Calm* (1966), contains good descriptive material but is already dated. *Again Korea* (1968), by WILFRED G. BURCHETT, an Australian Communist, tends to be more of a political diatribe but does contain some descriptive material.

(B.M.J.)

PyreneesRange

A mountain chain stretching from the shores of the Mediterranean Sea on the east to the Bay of Biscay on the Atlantic Ocean on the west, the Pyrenees (French, Pyrénées; Spanish, Pirineos) form a high wall between France and Spain that has played a significant role in the history of both countries and of Europe as a whole. The range is some 270 miles (430 kilometres) long; it is barely six miles wide at its eastern end, but at its centre it reaches 100 miles in width. At its western end it blends imperceptibly into the Cordillera Cantábrica (Cantabrian Mountains) along the northern coast of the Iberian Peninsula. Except in a few places, where Spanish territory juts northward or French southward, the crest of the chain marks the boundary between the two countries, though the tiny, autonomous principality of Andorra (*q.v.*) lies among its peaks. The highest point is the Pico de Aneto, at 11,169 feet (3,404 metres), in the Maladeta (Accursed) region of the Central Pyrenees.

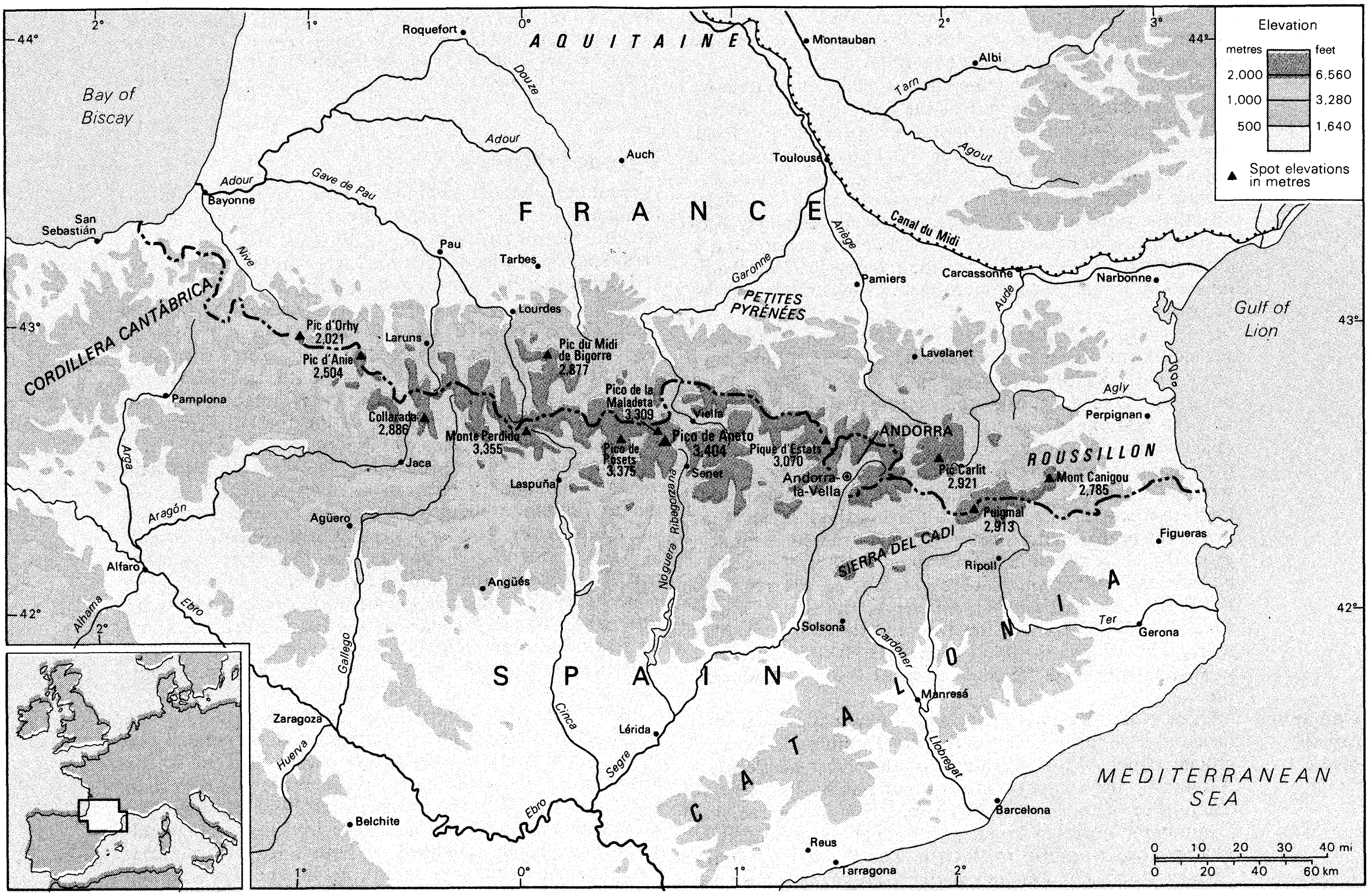
The Pyrenees long have been a formidable land barrier between Spain and Portugal on the Iberian Peninsula and the rest of Europe, helping in some way to tie these two countries more closely to Africa than the rest of Europe, and to the sea. From Pio Carlit (9,583 feet [2,921 metres]) near the eastern limit of the Pyrenees to the peaks of Orhy and Anie, a succession of mountains rise nearly 9,800 feet; at only a few places, all well to the west, can the chain be crossed through passes lower than 6,500 feet. In both the lower eastern and northwestern sectors, rivers dissect the landscape into numerous small basins. The range is flanked on both sides by broad depressions, the Aquitaine to the north, the Ebro to the south, both receiving waters from the major rivers flowing out of the mountains, the Garonne of France and the Ebro of Spain.

Scientificstudy. Until the 17th century, a general lack of knowledge about the Pyrenees permitted the repetition over centuries of errors and misconceptions about the mountains that had been propounded by such authors of antiquity as Diodorus Siculus of Sicily and the Greek geographer-historian Strabo (both, 1st century BC). In 1582 the first explorations were made, followed by botanical works from the academies at Montpellier-de-Médillan, France, and by other studies, including those of the 18th-century Swiss pioneer alpinist Horace Bénédict de Saussure. The earliest military map of the region dates from 1719, while early topographical studies were the bases of frontier treaties.

In the 19th century the first topographical and geological maps were made of the mountains, the latter beginning a series of geological interpretations and controversies among French and Spanish scientists. German studies added to the interpretive geology, but only in 1933 was the first study made that was based on modern research methods. In 1951 Luis Sol6 Sabaris published ***Los Pirineos,*** which dealt with many aspects of the range from various sources. More recent studies by universities, technical institutes, and national councils for research in France and Spain have produced an increasing amount of knowledge, and scholars of the subject have banded together in the International Union of Pyrenean Studies, which sponsors congresses to coordinate all phases of Pyrenean research.

International studies of the region

The physical environment. *The relief and its origins.* The Pyrenees represent the geological renewal of an old mountain chain rather than a more recent and vigorous mountain-building process that characterizes the Alps. Some 500,000,000 years ago the region now occupied by the Pyrenees was covered with the folded mountains created during the Paleozoic Era, a type known as Hercynian, which is represented by the Massif Central in France, the Ardennes, and the Meseta Central. Although these other massifs have had a comparatively quiet history of



The Pyrenees mountain range.

internal deformation, or tectonism, since their emergence, the Pyrenean block was submerged in a relatively unstable area of the Earth's crust, and its history was a troubled one from the beginning of the Mesozoic Era about *225,000,000* years ago.

The earliest formations, which were sediments severely folded over a granitic base, were submerged and covered by secondary sediments. They later were lifted once again into two parallel chains running to the north and south of the original Hercynian massif. These became the two zones of pre-Pyrenean ridges, of which the Spanish is the more fully developed and which are now great spurs of the main Pyrenean chain.

Under the forces of folding, the more recent and comparatively more plastic layers folded without breaking, but the original rigid base was split and dislocated. In the vicinity of the breaks, hot springs appeared and some metal-containing deposits formed. This upheaval affected chiefly the central and eastern regions. During this era, erosion was incessant, and, in the most exposed of the raised areas, weathering wore away the softer terrain and uncovered the old Hercynian sedimentary formations, occasionally reaching the deeper granitic bedrock.

Even today the old rocks, slates, schists, limestones transformed into marble (all of which come from old sediments transformed by great pressures and enormous heat), and granites of various kinds make up the spine, or axial zone, of the chain. The geological phases of this zone, which rises and widens from west to east and ends by sinking, with a steep drop of nearly 9,800 feet, into the depths of the Mediterranean, have determined the evolution of the massif as a whole.

North–south sequences of topography

The structure of the Pyrenees, therefore, is characterized by patterns of relief and of underlying structure running in a north–south sequence (like the base rock); these alternate with depressions, some of which are the result of internal deformations, others of erosion of less resistant overlying deposits. In a cross section directly through the central area, where the tectonic activity reached its fullest width and development, it is possible to distinguish, from north to south, two strips of the comparatively recent pre-Pyrenean fold, one Spanish and one French, in juxtaposition with the axial massifs. An outer strip to the north consists of folds constituting the Petites Pyrénées. Cut into channels, they permit the passage of rivers. Nearer the middle of the range rise the Inner Ridges, represented by the mighty cliffs of the Ariège, which contain the primary, or granitic, axial zones. On the Spanish side, the series is repeated in the opposite direction, but it is more highly developed and thicker. Thus the Interior Ridges—*e.g.*, Monte Perdido and the massif of Collarada—are sometimes higher than the neighbouring primary axial peaks. They are followed, going south, by a broad, pre-Pyrenean, middle depression, with a succession of marine and continental deposits of varying hardness that constitute the valleys of such tributaries of the Ebro as the Aragón. This depression continues across the rest of the pre-Pyrenean ridges, among which are new secondary outcrops that form the fringe of Exterior Ridges and the northern rim of the depression of the Ebro; they are not, however, as thick or as important as the Interior Ridges.

From the structure of their relief and from the climatic conditions (especially on the south) that derive from the geographical situation of the chain, the Pyrenees have been divided into three natural regions: the Eastern, or Mediterranean, Pyrenees (in Spanish, Pirineos Orientales, and Pyrénées Orientales in French); the Central Pyrenees; and the Western Pyrenees. The different vegetation, the linguistic divisions of the people, and—to a point—the racial characteristics and certain cultural distinctions would appear to confirm this classification.

Pyrenean climate and climatic effects. Major factors in the climate are the two abutting bodies of water and the extensive continental areas to the north and south. The oceanic influence penetrates southward across the

Marine and terrestrial influences on climate

low peaks of the Western Pyrenees, as far south as Pamplona, tempering somewhat the differences of climate between the northern and southern slopes. This is not the case in the rest of the chain, especially the Central Pyrenees. The contrast in humidity between the French and Spanish sides is remarkable. To the north, the oceanic influence, meeting no obstacles on the French plains of the Aquitaine, penetrates and goes a little beyond the north–south watershed of the French rivers flowing into the Mediterranean. To the east, the levanters, winds from the east and southeast, carry damp air from the Mediterranean, some of which falls as precipitation over the southeastern part of the eastern spurs. As a result, these regions are humid, while to the northeast the French depressions of the Roussillon acquire very Mediterranean characteristics.

South of the Central Pyrenees the valley of the Ebro—which runs in a general northwest–southeast direction and is blocked by the southwest–northeast-trending Catalonian ranges near the eastern coast of Spain—acts as a "little continent." Hence, its climate is one of great thermal contrasts that are exaggerated by the generally high altitude of the peninsula, but it is Mediterranean and unlike anything known in other European countries. Thus the variegated climatic pattern of the Pyrenees ranges from the limpid, sunny atmosphere of the continental zone to the mild mists of the northwest and includes all transition stages in between.

River systems. The hydrographic system consists basically of series of parallel valleys that descend from peaks of over 9,800 feet and from passes, most of which in the Central and Central-Eastern areas are above 6,500 feet. They are bordered by high, dividing ridges in a north–south direction, perpendicular to the axis of the chain. This type of valley produces short, torrential rivers that drop a long distance over short stretches; only seldom do these rivers flow, like the Aragón, through valleys that, as in the Alps, have both gentle slope and greater length. Their flow, extremely variable, especially on the southern side, is heavily influenced by the climate, as well as by the relief. Different maximum low waters occur in summer and winter; the spring, with maximum rain and melting snow, usually sees the greatest flows. In the Western Pyrenees and the northern zone, the rainfall pattern helps produce greater regularity; hence, flow is only slightly lower in summer. On the south, a few torrential rivers are fed principally by melting snows, a few largely by rain, but most from a combination of sources.

The river patterns and flow have been most important since antiquity in man's utilization of both the land and the rivers—from the floating of timber rafts downstream, which can only be done in the spring, to harnessing waterpower for industry and irrigation on the southern side by means of dams. The torrential flow of many of the rivers is the cause both of the crystalline purity of the Pyrenean waters and of their excellence and richness as fishing streams.

Glaciation. The present Pyrenean glaciers, perhaps more frequent on the northern than on the southern slopes, have been reduced to high basins—cirques or hanging valleys—at heights of over 9,800 feet. During the Pleistocene and Recent epochs (within the past 2,500,000 years), however, especially in the Central and much of the Eastern Pyrenees, glaciers left widespread erosion and various important sediments. The present-day lower lakes, or *ibdnes*, and idyllic meadows with their winding rivulets are among their marks. Glacier tongues were also the main causes of the deep valleys containing the river system.

Hotsprings. The fractured areas have many hot springs, both sulfurous and saline. The former are found throughout the axial massif, while the latter occur at the edges. These springs were exploited in Roman times, rediscovered in the 17th century, and reorganized and modernized toward the end of the 19th century. Over 20 famous examples occur on the French side; those in Spain are as numerous but are less fully exploited.

Bioecology and human uses. Forms of life in the Pyrenees have some remarkable characteristics that cannot be explained merely by the influences of climate and soil. The historical vicissitudes of the chain and its isolation at the southwestern limit of the main European peninsula, far from the centres of dispersion and variation of the various species (including man), have influenced the structure and character of its population.

Animal life. Some groups among the fauna, such as the cave-dwelling animals and frogs and toads, represent a migratory wave that come from ancient Tyrrhenia, associated with Corsica and Sardinia, and displaced certain native European species, relegating them to the Cordillera Cantábrica. The Pyrenean fauna is rich today, both in larger herbivores as well as in the variety and abundance of predators, though some species—*e.g.*, the wolf and the lynx—have disappeared. The southern Pyrenees represent one of the last important reserves for wild European fauna driven out of sectors more heavily populated by man. The present distribution and differentiation of large, warm-blooded animals is undoubtedly connected with the climate and the landscape, but the dominant feature is evidence of central European origin.

Distant origins of plant and animal life

Vegetative types. Similar comments may be made as to the origin of all cold-blooded animals, as well as of the vegetation. Basic differentiations exist among the latter. Pyrenean flora of tropical origin differentiated without any ancient European competition as the new chain replaced the old Hercynian; flora of Arctic origin, brought southward during relatively recent ice ages, are represented by two different branches of orophiles, or plants adapted to mountain life, from central Europe and from Siberia. Other orophiles have long been differentiated, but they are of Mediterranean origin and are dominant in the drier, sunnier parts of the southern slopes. An Atlantic group of flora predominates in the Western Pyrenees.

Climate and to a lesser extent soil are the conditioning elements in the vegetation of the Pyrenees, and there are two traditional methods of classifying vegetative types. One is classified by horizontally defined geographical regions, the other by vertically defined variations of climate, which varies with altitude.

In the northwest–southeast direction, the vegetation shows a marked and gradually decreasing oceanic influence; the contrary is the case with the Mediterranean influence from southeast to northwest. The exposure of the mountain surfaces and the conditions of local climate caused by mountain relief create special localized enclaves of all kinds. The most characteristic feature of the oceanic influence is the predominance of flat-leaf deciduous trees in the forests of the lower levels and the medium-height mountains, while the Mediterranean influence, represented by evergreen flat-leaved trees, is not only dominant in hot surroundings but also bears drought conditions better.

The variety of altitudinal vegetation shows itself in levels. From the medium-height mountain upwards, the flat-leaf woods at about 5,200 feet are replaced by needled conifers that require less water. This subalpine level, sometimes as low as 6,500 feet but usually above 7,800, gives way to the more sparsely covered pastures of the alpine level. This altitudinal scheme pertains in the vegetation east of the Orhy and Anie peaks. The oceanic influence, however, with its greater rainfall gives the west of the chain a different pattern. Flat-leaved deciduous beeches may be found as high as 5,850 feet, with some mix of the subalpine conifers, and there the high pastures are more resistant to damp and permanent snow. Overall the landscape is more like that of the high mountains of western Europe.

The Mediterranean influence expands through the entire valley of the Ebro, but it acquires marked signs of a more variable continental climate in the Central Pyrenees. Here, great quantities of mountain pines, which are more drought resistant, take the place of all kinds of flat-leaved deciduous trees in the higher, colder, and drier parts of the medium and higher levels of the southern slopes.

Patterns of human settlement. Similar factors influence man's utilization of the land, the kinds of crop, and the farming system of each district. In the Eastern Pyrenees and at the foot of the southern slope, summer droughts

Interactions of man and his environments

make irrigation and the development of a variety of horticultural crops imperative: plots of vegetables require borders of fruit trees. In this dry land, the woods have not been properly used, and the typical Mediterranean trilogy of olives, vines, and cereals predominates. In the Western Pyrenees, the abundant rainfall makes irrigation unnecessary for obtaining potatoes, sweet corn, and forage crops. The landscape there is profoundly marked by the traditional European checkering of tilled fields framed by more or less luxuriant hedgerows, shrubs, and climbing plants.

Few crops reach the mountain fringe of deciduous trees, and there the wood is used for firewood, furniture industries, and construction. At the subalpine level, meadows and pasture occupy nearly all the land; at 4,500 feet lies a zone balanced between woodland and pasture, while higher up the land is devoted almost entirely to stock breeding. This last activity must be intensive, for only about 90 days a year are suitable for grazing at the higher altitudes. The seasonal process of moving the flocks up and down the mountains maintains all the rich flavour of old traditions, but this way of life faces an increasing number of obstacles. The intensive farming of the plain lessens the acreage available for winter pasturing, and modern communications and economics eat away at the old ways.

As in all mountains, history has shown that the three basic resources are water, forests, and livestock. Agriculture, which has scant possibilities here, is always subordinate to stock breeding. Mining remains of little importance because it is so scattered. The hydroelectric potential is considerable. The ibdnes hanging at over 6,500 feet and the great rivers that can be dammed permit the establishment of industries without great transportation costs —*e.g.*, textiles and chemicals and above all those that utilize wood. Thus, there is a certain specialization or at least a tendency toward specialization in each of the three Pyrenean areas. The woodland and pastoral character is most apparent on the north side of the central area. The Western Pyrenees are markedly industrialized, while the Eastern Pyrenees are essentially agricultural, but industrial expansion is growing in their lower valleys. The transformation achieved goes hand in hand with the gradual development of communications, which is more noticeable at the two ends of the chain. The other regions remain in somewhat of an archaic substratum of economic life, and it will be difficult to stop the flight from the land without intelligent renovation of equipment and variation in the traditional modes of living. Finally, tourism, which began with the old Roman spas, is being revitalized with winter sports, hunting and fishing, and by a love of nature in general. This pathway offers the people of the Pyrenees a growing sense of forward movement.

BIBLIOGRAPHY. In English the reader will find the following of particular interest: NINA EPTON, *Navarre* (1957); HENRY R. FEDDEN, *The Enchanted Mountains* (1962); J.B. MORTON, *Pyrenean* (1938); and HENRY MYHILL, *The Spanish Pyrenees* (1966). In Spanish, LUIS SOLE SABARIS, *Los Pirineos* (1951), covers both sides of the Pyrenees with scholarly detail and illuminating generalizations; R VIOLANT Y SIMORRA, El *Pirineo espaiiol: Vida, usos, costumbres, creencias y tradiciones de una cultura inilenaria que desaparece* (1949), is a more popular volume with hundreds of photographs, giving greater emphasis to the human aspects of Spanish Pyrenean life. In French, the classic M. SORRE, *Les Pyrénées* (1922), is the first modern scientific study; P. ARQUE, *Géographie des Pyrénées Francaises* (1943), deals with the French Pyrenees.

Pyroxenes

Occurrence and use

The pyroxenes are an important group of rock-forming silicate minerals of variable composition, among which calcium-, magnesium-, and iron-rich varieties predominate. The name is derived from the Greek pyro, "fire," and *xenos*, "stranger," in allusion to their occurrence in volcanic rocks, which was initially thought to be accidental rather than indigenous. Pyroxenes are the principal minerals in many kinds of igneous (crystalline) and metamorphic (altered) rocks, however, and they occur in meteorites and lunar samples as well. Omphacite, a variety rich in sodium and aluminum, and aluminous enstatite are thought to be abundant in the Earth's mantle—the zone directly beneath the crust—and analysis of the ıadiation emitted by certain cool stars reveals the probable presence of pyroxenes throughout the known universe.

Certain pyroxenes are useful as economic mineral guides, or indicators: johannsenite is generally associated with copper, zinc, and lead ores; omphacite occurs with diamond in kimberlites (relatively rare igneous rocks that are rich in iron and magnesium); and spodumene and aegirine (lithium aluminosilicate and sodium ferrisilicate minerals, respectively) are often indicators of pegmatites—very coarse-grained crystalline rocks that commonly contain a rare-mineral assemblage. Aside from this usage as indicators or guide minerals, only jadeite, nephrite—ornamental jade—and spodumene, a commercial source of lithium, are of economic importance.

Form, habit, and colour

All pyroxene species exhibit a prismatic crystal form in which the angles between crystal faces are about 87° and 93°. The pyroxenes have good cleavages parallel to these directions—*i.e.*, they tend to split along planes that have these trends. As a consequence, the pyroxenes possess nearly square cross sections when viewed perpendicularly to their principal cleavage directions, and this characteristic is diagnostic. Most of the pyroxenes in rocks in the Earth's crust, however, occur as irregularly shaped grains; short, stubby, prismatic crystals are less common.

Aegirine commonly occurs as long, needle-like crystals, and spodumene as prismatic crystals that are as large as 12 metres (40 feet) long and two to six feet wide. These two minerals are associated with very coarse-grained pegmatites.

The colour of pyroxenes varies with slight changes of chemical composition: augite is brown or green; diopside is either colourless or white to green; and aegirine is green to greenish black. Some varieties of jadeite and spodumene may be violet or pink.

This article treats the crystal structure and chemical composition of the pyroxenes and their natural formation and synthesis. For further information on crystal structure in general, see CRYSTALLOGRAPHY; and for an overview of rock-forming minerals and their place within the hierarchy of all minerals, see SILICATE MINERALS; and MINERALS. Ornamental varieties of pyroxenes are further considered in the article GEMSTONES. See also GEOCHEMICAL EQUILIBRIA AT HIGH TEMPERATURES AND PRESSURES for relevant coverage of conditions of pyroxene formation.

CLASSIFICATION AND PROPERTIES

Variation of composition and physical properties

Mineral names and groups. The naming of a pyroxene depends upon its major element composition and crystal system, and the more refined aspects of minor element content and crystal structure often require prefixes. Diopside, for example, is primarily a calcium (Ca) magnesium (Mg) silicate ($CaMgSi_2O_6$) and crystallizes in the monoclinic system, which refers to three unequal crystallographic axes, two intersecting obliquely and the third perpendicular to these two. Diopside has the basic pyroxene crystal structure but differs in detail from another monoclinic pyroxene such as spodumene. A trace content of chromium also is sufficient in some cases to warrant the name chrome diopside. Enstatite is primarily magnesium silicate ($Mg_2Si_2O_6$) and usually crystallizes in the orthorhombic system, which refers to three unequal crystallographic axes that are mutually perpendicular, but polymorphism (the ability of a mineral species to occur in different forms) gives rise to other crystal structures with the same composition. These may be orthorhombic such as protoenstatite or monoclinic such as clinoenstatite. A trace content of aluminum may warrant the name aluminous enstatite. For diopside, a marked increase of iron in place of magnesium leads to hedenbergite; magnesium in place of calcium leads to clinoenstatite; and magnesium and iron in place of calcium yield augite and ultimately pigeonite. This subgroup of pyroxenes, which is by far the most common, can first be expressed chemically in terms of three end-members (pure compounds): calcium silicate ($Ca_2Si_2O_6$), magnesium silicate ($Mg_2Si_2O_6$), and iron (Fe) silicate ($Fe_2Si_2O_6$).

The pyroxene formulas listed in the Table are relevant only to these end-member compositions, and the sections of this article on crystal structure and chemical composition explain the relationships between these and the more common intermediate variants. The general composition of the pyroxenes, in fact, must be expressed as: XYZ_2O_6, in which X is zero or one, Y is one or two, and X + Y = 2. The compositional range can be indicated by the array of possible substituting elements—*i.e.*, X = calcium, sodium, or, less commonly, lithium; Y = magnesium, iron, aluminum, or, less commonly, manganese, nickel, chromium, and titanium; and Z = silicon, but aluminum or ferric iron also may occur.

Certain optical properties serve to differentiate the pyroxenes, but other physical properties tend to be rather uniform. The hardness of these minerals is 6 to 6½ on Mohs' scale (ranging from talc = 1 to diamond = 10), the lustre is vitreous or glassy, and the form and habit also tend to be rather similar because of the good cleavage at 87" and 93" as previously noted. Density variations (see the Table) are diagnostic of iron enrichment.

The Pyroxene Minerals

mineral	chemical formula	unit-cell dimensions*			density (g/cm³)
		a	*b*	*c*	
Diopside	$CaMgSi_2O_6$	9.73	8.91	5.25	3.25
Hedenberaite	$CaFe^{2+}Si_2O_6$	9.85	9.02	5.26	3.54
Enstatite	$Mg_2Si_2O_6$	18.23	8.81	5.19	3.21
Ferrosilite (Ortho)	$Fe_2^{2+}Si_2O_6$	18.43	9.06	5.26	3.96
Aegirine (or Acmite)	$NaFe^{3+}Si_2O_6$	9.65	8.79	5.29	3.56
Jadeite	$NaAlSi_2O_6$	9.50	8.61	5.24	3.25
Spodumene	$LiAlSi_2O_6$	9.50	8.30	5.24	3.16
Johannsenite	$CaMnSi_2O_6$	9.83	9.04	5.27	3.45

*The unit cell is the smallest volume that will provide a representative sample of the atomic and molecular groups that comprise a mineral. Dimensions are given along the three crystallographic axes (designated *a*, *b*, and *c*) in angstrom units (one angstrom unit equals 10^{-8} centimetre).

Basic pyroxene chain structure

Crystal structure. The crystalline structure of all pyroxene minerals consists basically of four oxygen atoms arranged as a tetrahedron with a silicon atom at its centre. These tetrahedrons are linked in the form of chains, a structure similar to that of the amphiboles. The linkage is effected by the sharing of two oxygens between neighbouring tetrahedrons in the chain, so that the basic formula becomes one silicon (Si) to three oxygens (0), or SiO_3. Identical units of the chain are repeated at approximately 5.3-A (one angstrom unit [A] equals 10^{-8} centimetre) intervals; this direction defines the c-axis of the crystal, and the distance is the dimension of the unit cell in the c-direction. The linkage of a pyroxene chain viewed from various directions is shown in Figure 1.

The diopside structure consists of only one type of chain, which is also characteristic of jadeite, augite, and protoenstatite. In contrast, the clinoenstatite and pigeonite structures consist of two structurally distinct SiO_3 chains. The orthorhombic pyroxenes have a unit cell that is derived by twinning (repetitive intergrowth) of the latter structure along the a-axis. In spodumene there is a slight deviation whereby two different kinds of SiO_4 tetrahedrons alternate along one kind of chain, and in omphacite this alternation occurs along each of two kinds of chain to produce a low-symmetry structure capable of accommodating a wide variety of cation (positively charged atom) sizes.

The diopside structure

The tetrahedrons may show some substitution of silicon by aluminum and, less commonly, by traces of ferric iron or titanium. Most of the important cations other than silicon, however, occur as layer cations that are linked to the chain oxygens to form cation–oxygen polyhedrons. Thus, the chains themselves are linked together by these cations to form a structure that is weakest parallel to the dominant cleavage direction of all pyroxenes.

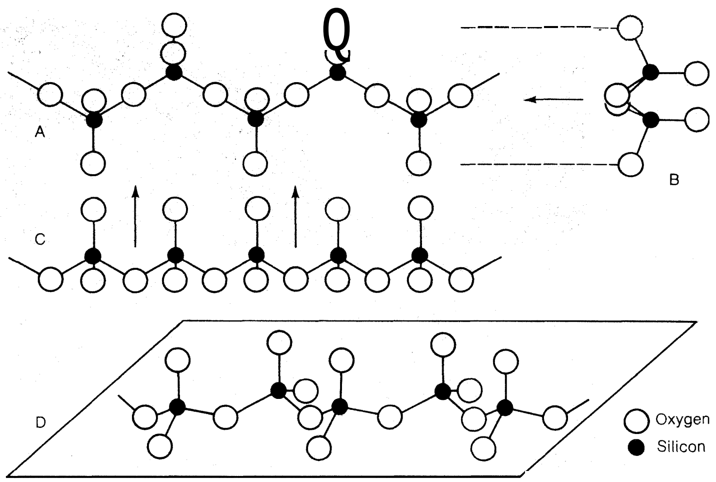


Figure 1: Single pyroxene chain, $(SiO_3)n$, in three projections (A) on (100), (B) along the *z* direction, (C) along the y direction, as well as (D) in perspective.

After Bragg, 1937, Jong, 1959, in Deer, Howie, and Zussman, *Rock Forming Minerals*, vol. 2 (1968); Longman Group Ltd.

After Warren and Bragg, 1928, in Deer, Howie, and Zussman, *Rock Forming Minerals*, vol. 2, fig. 14 (1963); Longman Group Ltd.

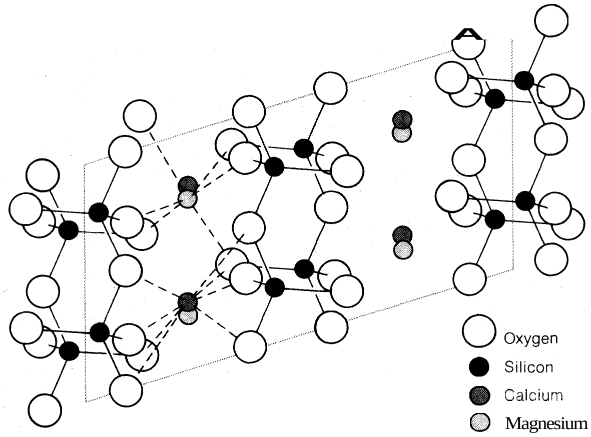


Figure 2: Idealized structure of diopside as viewed along the y direction.

The diopside structure may be taken as an example. Calcium and magnesium are divalent (doubly charged) cations that fulfill charge balance requirements to give the formula $CaMg(SiO_3)_2$. These cations lie in octahedral layers parallel to (100)—*i.e.*, a plane that intersects the a-axis and is parallel to b and c (see CRYSTALLOGRAPHY)—but occur in two crystallographically distinct cation sites. The small magnesium ion (ionic radius = 0.78 A) lies in the M_1 site and is linked with six oxygen atoms to form a fairly regular octahedron. The larger calcium ion (ionic radius = 1.06 A) in the M_2 site, however, is linked with eight oxygens. Chemical substitution within the pyroxene group allows the M_1 site to be occupied by many small cations such as magnesium, ferrous and ferric iron, aluminum, or manganese. The M_2 site is usually occupied by the larger cations such as calcium or sodium, but in calcium-poor pyroxenes the site is occupied by smaller cations such as magnesium or ferrous iron in hypersthene or sodium in jadeite; the M_2 polyhedrons are then distorted octahedrons. In the pigeonite structure, which is somewhat problematical, disordered calcium and ferrous iron probably occupy the M_2 site in irregular 7-fold coordination (*i.e.*, surrounded by and bonded to seven oxygens), whereas M_1 is occupied by disordered magnesium and ferrous iron in regular 6-fold coordination. This condition may, in fact, be an average of randomly arranged domains of diopside-type and clinoenstatite-type structures.

The simpler structure of diopside is illustrated in Figure 2, in which the M_2 site is occupied by calcium in 8-fold coordination and in which the M_1 site is occupied by

Photomicrographs of various pyroxene minerals in thin sections (illuminated with polarized light). (Left) Augite phenocryst (large, individual, gray crystal) in basalt lava, showing characteristic basal octagonal form and square-segmentation cleavage (magnified about 18.5 X). (Centre) Titanaugite crystal (yellow) showing typical hourglass zoning (magnified about 13.2 X). (Right) Gabbro with two pyroxenes, a calcium-rich augite (blue-red and green) enclosed by a calcium-poor pigeonite crystal; clinohypersthene in thin lamellae has separated from the calcium-rich augite, and augite has separated from the pigeonite before it inverted to hypersthene (gray; magnified about 21.1 X).

BY courtesy of G. Malcolm Brown

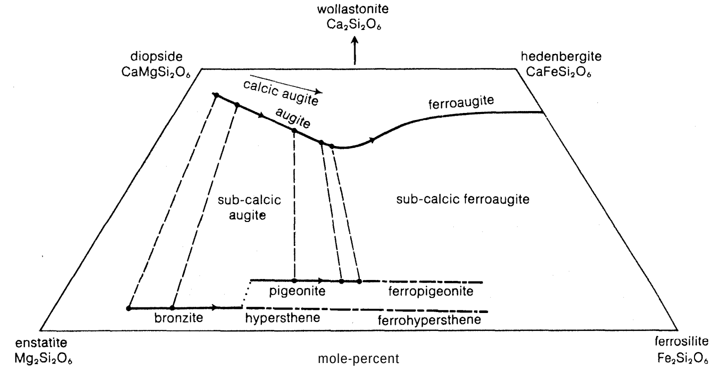
magnesium in 6-fold coordination. Hedenbergite would differ only by the presence of ferrous iron rather than magnesium in the M_1 site, and in the common augite both the M_1 and M_2 sites must be disordered between calcium, magnesium, and ferrous iron.

In the calcium-poor pyroxenes, the structural relations are more complex and lead to polymorphism. Magnesium silicate ($Mg_2Si_2O_6$) may be taken as an example. Three structural forms exist: clinoenstatite, rhombic enstatite, and protoenstatite. In these, successive tetrahedral slabs are related by glideplanes of symmetry parallel to (100), the glide component acting along either direction of the c-axis (see CRYSTALLOGRAPHY).

Chemical composition. The term common pyroxenes is applied to the great majority of minerals with compositions that can be expressed in terms of their calcium silicate ($Ca_2Si_2O_6$), magnesium silicate ($Mg_2Si_2O_6$), and iron silicate ($Fe_2Si_2O_6$) molecular contents. Compositions plotted on a triangular diagram show distinct pyroxene groupings that are of the greatest significance. No pyroxenes plot in that part of the diagram more calcic than the join (boundary line between phases) at $CaMgSi_2O_6$ and $CaFeSi_2O_6$ because of the structural prohibition of calcium from entering the M_1 sites. The common pyroxene quadrilateral is shown in Figure 3.

The pyroxene quadrilateral

The nomenclature of the common clinopyroxenes was once based on subdivision of the quadrilateral, but the rigidity of the nomenclature and the orientation of the boundary lines between each named type proved unsatisfactory. It is now known that relationships within and between the common pyroxenes simplify the diagrammatic representation of their compositional fields within the quadrilateral.



gι re 3: C(non pyr quadrilateral showing the compositional ranges f various py

Pyroxenes plotting close to or on the diopside–hedenbergite join are found only in metamorphosed calcium silicate rocks, for example by the reaction

$$\underset{\text{dolomite}}{CaMg(CO_3)_2} + \underset{\text{quartz}}{2SiO_2} \rightleftharpoons \underset{\text{diopside}}{CaMgSi_2O_6} + \underset{\text{carbon dioxide.}}{2CO_2}$$

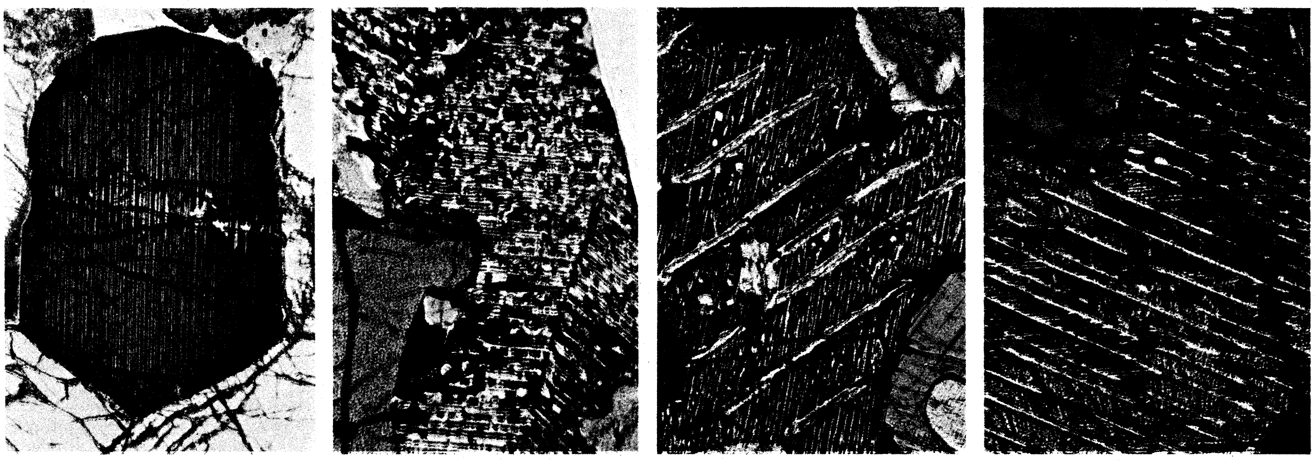
The calcic augites occur in alkali basalts and their derivatives and are usually richer in aluminum and titanium than the augites. The calcic augites do not have iron-rich variants because their parent magmas (molten silicate material from which igneous rocks crystallize upon cooling) generally produce liquids from which aegirine-augite and aegirine precipitate. The titanium-rich varieties are known as titanaugites and often exhibit hourglass zoning. The augites are the most common pyroxenes and show extensive magnesium–iron substitution to give the distinctive augite–ferroaugite trend line. Extreme fractionation of tholeiitic (olivine-poor) basalt magma is depicted by this trend of pyroxenes: diopsidic augite in ultramafic rocks (composed almost entirely of iron magnesium silicate), to augite in less mafic gabbroic rocks, to ferroaugite in intermediate, dioritic rocks (containing some quartz and sodium feldspar), and to ferrohedenbergite in acid, granitic rocks (rich in silica). The subcalcic augites and subcalcic ferroaugites occur as metastable phases (subject to further change under existing conditions) in the quickly cooled groundmass (fine-grained crystalline bulk) of lavas. The tholeiitic suite of differentiated rocks generally contain, in addition to a member of the augite series, a calcium-poor pyroxene with, approximately, either 5 percent (bronzite) or 10 percent (pigeonite) of calcium silicate. The ultramafic rocks usually contain a bronzite, the gabbroic rocks a pigeonite, the intermediate rocks a ferropigeonite, and the acid rocks have no calcium-poor pyroxene. Certain intermediate rocks, notably the andesite and dacite lavas, contain hypersthene or ferrohypersthene instead of a pigeonitic phase.

The enstatite–ferrosilite pyroxene series is not fully represented on the quadrilateral. The nomenclature, based on ferrosilite percentage, is: enstatite (0–10 percent), bronzite (10–30 percent), hypersthene (30–50 percent), ferrohypersthene (50–70 percent), eulite (70–90 percent), ferrosilite (90–100 percent).

Aluminum and other elements

The aluminum (Al) content of the common pyroxenes varies with temperature and pressure of crystallization, and, in cases in which two pyroxenes are present, it is higher in the calcium-rich phase. In ultramafic nodules from basalt, enstatites contain up to 5.5 percent (average about 1.5 percent in lava orthopyroxenes), and calcic augites contain about 6.5 percent alumina (Al_2O_3). Compared with coexisting olivines, pyroxenes show higher chromium and vanadium and lower nickel and cobalt contents. With fractionation, augites show depletion in chromium, vanadium, and nickel and slight enrichment in scandium. Coexisting calcium-rich and calcium-poor pyroxenes have a distribution of magnesium and iron between the phases that may ultimately be a valuable indicator of the pressure and temperature of formation. At present the distribution is too much influenced by other factors, including the oxidation state of the iron and aluminum content, to be dependable.

The less common pyroxenes have a more simple chemistry, which is expressed chiefly by the formulas in the Table. In aegirines the main replacement is sodium and ferric iron by calcium, magnesium, and ferrous iron to give aegirine augites, although varieties with up to 4 percent vanadium oxide and 5 percent manganese oxide have



Photomicrographs of various thin sections containing pyroxene minerals; all show separation within a mineral grain of distinct phases due to further cooling after the grain had solidified (illuminated by polarized light). (Far left) Bronzite crystal from an ultramafic rock; thin lamellae of a calcium-rich species, probably pigeonite, have separated from the bronzite, and the host (grayish) thus has a very low calcium content (magnified about 40 X). (Left) Twinned crystal of inverted pigeonite from a gabbro. Augite, seen as brightly coloured thin lamellae with herringbone texture because of the twinned relationship, has separated from the pigeonite; further cooling has caused the host, gray-coloured hypersthene, to change symmetry (invert; magnified about 22×). (Right) Inve[illegible] it r[illegible] a [illegible] e [illegible] ly cooled gabbro than that at left as a [illegible] l . the ugite lamellae are wider, and after inversion more augite has separated from the hypersthene host (magnified about 70.4 X). (Far right) Complex separation of augite from an inverted pigeonite (magnified about 70.4 X).

BY courtesy of G. Malcolm Brown

been reported. Ferroan johannsenite with about 13 percent iron oxide is transitional to hedenbergite. Omphacite (rich in sodium and aluminum) and fassaite (rich in aluminum and ferric iron) are otherwise of diopside composition.

PHASE RELATIONS AND STABILITY

Pyroxenes can be formed over a very wide range of temperatures, representing their stability in low-grade metamorphic rocks (subject to low temperatures and pressures), at the one extreme, and ultramafic nodules from upper-mantle sources, at the other. Diopside is formed under low carbon dioxide pressures at 300" C (57" F) from dolomite and silica and from a diopside melt at 1,391" *C* (2,536" F). The stability fields of pyroxenes are relatively larger at higher temperatures, however, and the more common phases are found in igneous and high-temperature metamorphic environments. Hydrothermal (referring to hot, water-rich fluids derived from magma) alteration usually produces amphibole or chlorite, but pyroxenes are generally resistant accessories in detrital sediments.

Pressure–temperature conditions

The pyroxene quadrilateral (Figure 3) has been studied extensively. Diopside (di) has no known polymorphs, and its pressure–temperature phase diagram is a simple melting curve: at low pressures the rate of change of temperature with pressure is 15° C (59° F) per kilobar (the bar is a pressure unit equal to 10^6 dynes per square centimetre). Enstatite (en) melts incongruently (accompanied by reaction with the liquid) at low pressures but congruently at about 2.3 kilobars. With respect to the di-en join at atmospheric pressure (Figure 4), reaction and crystallization result in two coexisting pyroxenes at temperatures below 1,400" C (2,550" F), calcium-rich and calcium-poor. The calcium-poor phase at solidus temperatures (the solidus is the locus of points on a phase diagram above which solid and liquid are in equilibrium and below which the system is entirely solid) is a protoenstatite solid solution (single crystalline phase that varies in composition within finite limits), and the calcium-rich phase is an augite solid solution. At subsolidus temperatures each phase exsolves (when solid solutions become unstable because of temperature change, the phases may separate out, or exsolve) the other phase and changes composition along the domed limbs of the curve. At about 1,100" C (2,000" F), the protophase inverts to a slightly more calcic, orthorhombic phase that then continues to exsolve diopside to 800" C (1,500" F) or lower. This bears general comparison with natural pyroxenes, except that the protoenstatite phase is rare. It inverts to clinoenstatite on quenching and ceases to be stable above about eight kilobars.

From F.R. Boyd and J.F. Schairer, "The System $MgSiO_3$-$CaMgSi_2O_6$," *Journal of Petrology*, Vol. 5, no. 2 (1964); Oxford University Press

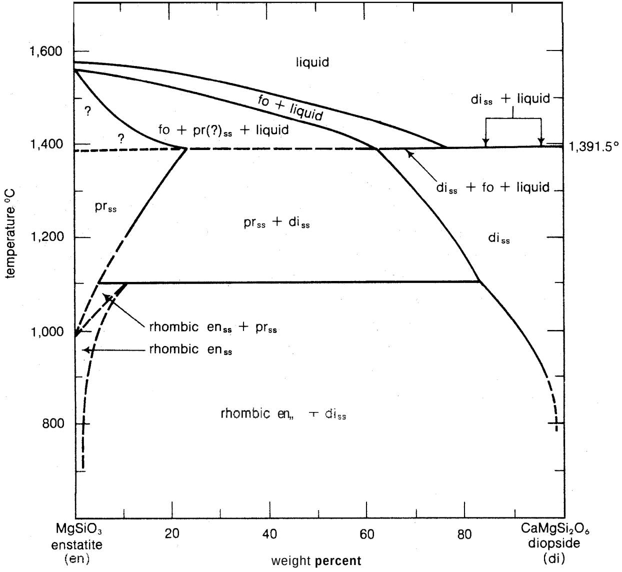


Figure 4: Equilibrium relations along the enstatite–diopside join. The phases of various mixtures of pure enstatite (en) and pure diopside (di) are shown as a function of temperature at atmospheric pressure. Other phases shown are protoenstatite (pr) and forsterite (fo), an olivine with the composition Mg_2SiO_4.

Pyroxenes in the hedenbergite (Hd)–ferrosilite (Fs) system are not stable at liquidus temperatures (above the liquidus temperature a system is completely liquid), the higher temperature equivalents being ferriferous wollastonite, fayalitic olivine, and a silica phase (see Figure 5). Ferrohedenbergite is stable below about 950" C (1,750" F), but ferrosilite is totally unstable at these low pressures. Pressure increases the stability field of hedenbergite, and at about 13 kilobars the iron wollastonite field is eliminated. At about 17 kilobars, ferrosilite is stable and melts congruently, although the high-temperature liquidus phase has a problematical pyroxene structure.

The diopside–enstatite and enstatite–ferrosilite phase relations would be simple solid solutions between each

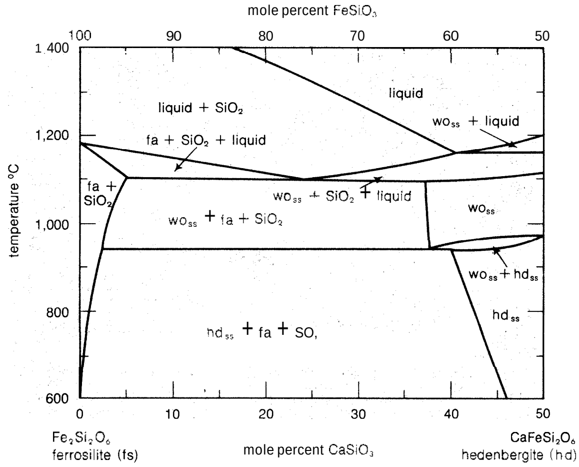


Figure 5: Equilibrium relations for the iron-rich pyroxenes along the ferrosilite–hedenbergite join. The phases of various mixtures of pure ferrosilite (fs) and pure hedenbergite (hd) are shown as a function of temperature at atmospheric pressure. Other phases, the result of the instability of pyroxenes at elevated temperatures, are iron-rich wollastonite (wo), the olivine fayalite (fa) with a composition of Fe_2SiO_4, and silica (SiO_2).

From D.H. Lindsley, G.M. Brown, and I.D. Muir, *Mineralogical Society of America* Special Paper *2* (1969)

end-member if it were not for the complexities shown on Figures 4 and 5. Natural pyroxene occurrences indicate the need for a restudy of the enstatite–ferrosilite system, of the effect of diopside on hedenbergite phase relations, and of compositional joins within the quadrilateral.

Confirmation by mineral synthesis

The broad natural relations shown in Figure 3 are confirmed by synthetic studies. Augite and bronzite coexistence, for example, results from the demonstrated instability of protopyroxene at moderate pressures and iron content. Inversion to clinoenstatite from protoenstatite occurs in a Papuan lava and in enstatite–chondrite meteorites. Pigeonite is stable at high pressures on the diopside–enstatite join but in nature is only found in the composition range shown on Figure 3, coexisting with augite. Inversion of pigeonite occurs in slowly cooled gabbros and dolerites unless they are iron-rich. The inversion product is orthorhombic hypersthene, with augite exsolution lamellae (thin layers) that indicate the previous monoclinic structure. Exsolution at 1,000" C and inversion at 980" *C* have been shown by experimental studies at zero pressure. The cessation of the calcium-poor pyroxene trend (Figure 3) is explained by ferrosilite instability, the end point of the calcium-rich trend coincides with the minimum for Wo,, + Hd,, shown on Figure 5, and the absence of iron wollastonites in all but low-pressure environments is also confirmed.

The occurrence of Tschermak's molecule

Aluminous pyroxenes contain the so-called Tschermak's molecule ($CaAl_2SiO_6$), which alone is stable above 1,150" C (2,100" F) and 11 kilobars. This molecule and the magnesium aluminum silicate molecule ($MgAl_2SiO_6$) are molecular expressions of the aluminum that can enter pyroxene structures at high pressure. In magnesium silicate ($MgSiO_3$), the alumina (Al_2O_3) solubility increases from 2.5 percent at atmospheric pressure to 14–19 percent at 20 kilobars, but it decreases at higher pressures. Diopside, however, can contain 12–15 percent Al_2O_3 at atmospheric pressure. Ferric diopsides contain ferri-Tschermak's molecule ($CaFe_2SiO_6$) produced, with andradite garnet and quartz, by oxidation of augites. More sodium in the magma would lead to formation of aegirine (acmite), which melts incongruently to hematite + magnetite + liquid up to 45 kilobars and from low to high oxygen pressures. Experiments at 40 kilobars suggest complete solid solubility between acmite and jadeite. Jadeite (from albite + nepheline) is only stable above about 17 kilobars (900" C [1,650° F]), and jadeite + quartz (from albite) is stable above about 25 kilobars at similar temperatures. Omphacite is a jadeite–diopside solid solution with a fairly wide pressure-staoility range if viewed in isolation from other phases. At 30 kilobars, it splits into jadeitic and diopsidic phases at about 1,450" C (2,640" F). Spodumene is the high-pressure member of three polymorphs with lithium aluminum silicate compositions ($LiAl\ Si_2\ O_6$). At 10 kilobars and 900" C (1,650" F), it converts to a form called *β*-eucryptite, so that its low-pressure pegmatite environment must imply crystallization well below 900" *C*.

BIBLIOGRAPHY. W.A. DEER, R.A. HOWIE, and J. ZUSSMAN, *Rock-Forming Minerals,* vol. 2, *Chain Silicates* (1963), a comprehensive, modern text dealing with all-aspects of pyroxene mineralogy, with an extensive bibliography—now somewhat outdated because post-1962 research is extensive; A. POLDERVAART and H.H. HESS, "Pyroxenes in the Crystallization of Basaltic Magma," *J. Geol.,* 59:472–489 (1951), a now-classic exposition of the paragenetic relationships among the common pyroxenes; AMERICAN GEOLOGICAL SOCIETY, *Short Course on Chain Silicates* (1966), a useful summary of modern research on pyroxene and amphibole mineralogy, much of which is otherwise unpublished—see especially the sections on pyroxene phase relations (F.R. BOYD), crystal structures (D.E. APPLEMAN), and paragenesis (G.M. BROWN); G.M. BROWN, "Mineralogy of Basaltic Rocks," in H.H. HESS and A. POLDERVAART (eds.), *Basalts* (1967), a modern review of the structures, chemistry, and phase relations of the more common pyroxenes, with references to papers on subsolidus exsolution and inversion (extensive bibliography); *Annual Report of the Director, Geophysical Laboratory, Carnegie Institution, Washington D.C.* (1964–), experimental and crystallographic studies of pyroxene mineralogy described in numerous original articles: B.H. MASON and L.G. BERRY, *Elements of Mineralogy* (1968); F.R. BOYD JR. and J.F. SCHAIRER, "The System Mg Si 0_3-Ca Mg Si_2 0,," *J. Petrology,* 5:275–309 (1964); and D.H. LINDSLEY and J.L. MUNOZ, "Subsolidus Relations Along the Join Hedenbergite-Ferrosilite," *Am. J. Sci.,* 267–A:295–324 (1969).

(G.M.B.)

Pythagoreanism

The philosophical school and religious brotherhood known as Pythagoreanism is believed to have been founded by Pythagoras of Samos, who settled in Croton in southern Italy about 525 BC.

GENERAL FEATURES OF PYTHAGOREANISM

The character of the original Pythagoreanism is controversial, and the conglomeration of disparate features that it displayed is intrinsically confusing. Its fame rests, however, on some very influential ideas, not always correctly understood, that have been ascribed to it since antiquity. These ideas include those of (1) the metaphysic of number and the conception that reality, including music and astronomy, is, at its deepest level, mathematical in nature; (2) the use of philosophy as a means of spiritual purification; (3) the heavenly destiny of the soul and the possibility of its rising to union with the divine; (4) the appeal to certain symbols, sometimes mystical, such as the *tetraktys,* the golden section, and the harmony of the spheres (to be discussed below); (5) the Pythagorean theorem; and (6) the demand that members of the order shall observe a strict loyalty and secrecy.

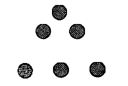


Figure 1: The *Tetraktys* (see text).

By laying stress on certain inner experiences and intuitive truths revealed only to the initiated, Pythagoreanism seems to have represented a soul-directed subjectivism alien to the mainstream of Pre-Socratic Greek thought centring on the Ionian coast of Asia Minor (Thales, Anaximander, Anaxagoras, and others), which was preoccupied with determining what the basic cosmic substance is.

In contrast with such Ionian naturalism, Pythagoreanism was akin to trends seen in mystery religions and emotional movements, such as Orphism, which often claimed to achieve through intoxication a spiritual insight into the divine origin and nature of the soul. Yet there are also aspects of it that appear to have owed much to the more sober, "Homeric" philosophy of the Ionians. The Pythagoreans, for example, displayed an interest in metaphysics (the nature of Being), as did their naturalistic predecessors, though they claimed to find its key in mathematical form rather than in any substance. They accepted the essentially Ionian doctrines that the world is composed of opposites (wet–dry, hot–cold, etc.) and generated from something unlimited; but they added the idea of the imposition of limit upon the unlimited and the sense of a musical harmony in the universe. Again, like the Ionians, they devoted themselves to astronomical and geometrical speculation. Combining, as it does, a rationalistic theory of number with a mystic numerology and a speculative cosmology with a theory of the deeper, more enigmatic reaches of the soul, Pythagoreanism interweaves Rationalism and irrationalism more inseparably than does any other movement in ancient Greek thought (see PHILOSOPHY, HISTORY OF WESTERN).

MAJOR CONCERNS AND TEACHINGS

The problem of describing Pythagoreanism is complicated by the fact that the surviving picture is far from complete, being based chiefly on a small number of fragments from the time before Plato and on various discussions in authors who wrote much later—most of whom were either Aristotelians or Neoplatonists (see below *History of Pythagoreanism*). In spite of the historical uncertainties, however, that have plagued searching scholars, the contribution of Pythagoreanism to Western culture has been significant and therefore justifies the effort, however inadequate, to depict what its teachings may have been. Moreover, the heterogeneousness of Pythagorean doctrines has been well documented ever since Heracleitus, a classic early-5th-century Greek philosopher who, scoffing at Pythagoras' wide-ranging knowledge, said that it "does not teach one to have intelligence." There probably never existed a strictly uniform system of Pythagorean philosophy and religious beliefs, even if the school did have a certain internal organization. Pythagoras appears to have taught by pregnant, cryptic *akousmata* ("something heard) or *symbola*. His pupils handed these on, formed them partly into *Hieroi Logoi* ("Sacred Discourses"), of which different versions were current from the 4th century on, and interpreted them according to their convictions.

Significance of belief in transmigration of souls

Religion and ethics. The belief in the transmigration of souls provided a basis for the Pythagorean way of life. Some Pythagoreans deduced from this belief the principle of "the kinship of all beings," the ethical implications of which were later stressed in 4th-century speculation. Pythagoras himself seems to have claimed a semidivine status in close association with the superior god Apollo; he believed that he was able to remember his earlier incarnations and, hence, to know more than others knew. Recent research has emphasized shamanistic traits deriving from the ecstatic cult practices of Thracian medicine men in the early Pythagorean outlook. The rules for the religious life that Pythagoras taught were largely ritualistic: refrain from speaking about the holy, wear white clothes, observe sexual purity, do not touch beans, and so forth. He seems also to have taught purification of the soul by means of music and mental activity (later called philosophy) in order to reach higher incarnations. "To be like your Master" and so "to come nearer to the gods" was the challenge that he imposed on his pupils. Salvation, and perhaps ultimate union with the divine cosmos through the study of the cosmic order, became one of the leading ideas in his school.

The advanced ethics and political theories sometimes ascribed to Pythagoreanism may to some extent reflect ideas later developed in the circle of Archytas, the leading 4th-century Pythagorean. But a picture current among the Peripatetics (the school founded by Aristotle) of Pythagoras as the educator of the Greeks, who publicly preached a gospel of humanity, is clearly anachronistic. Several of the Peripatetic writers, Aristoxenus, Dicaearchus, and Timaeus, seem to have interpreted some principles—properly laid down only for esoteric use in the brotherhood—as though these applied to all mankind: the internal loyalty, modesty, self-discipline, piety, and abstinence required by the secret doctrinal system; the higher view of womanhood reflected in the admission of women to the school; a certain community of property; and perhaps the drawing of a parallelism between the macrocosm (the universe) and the microcosm (man), in which (for instance) the Pythagorean idea that the cosmos is an organism was applied to the state, which should thus mix monarchy, oligarchy, and democracy into a harmonic whole—these were all universalized.

Metaphysics and number theory. According to Aristotle, number speculation is the most characteristic feature of Pythagoreanism. Things "are" number, or "resemble" number. To many Pythagoreans this concept meant that things are measurable and commensurable or proportional in terms of number—an idea of considerable significance for Western civilization. But there were also attempts to arrange a certain minimum number of pebbles so as to represent the shape of a thing—as, for instance, stars in a constellation that seem to represent an animal. For the Pythagoreans even abstracted things "have" their number: "justice" is associated with the number four and with a square, "marriage" with the number five, and so on. The psychological associations at work here have not been clarified.

The tetraktys

The harmony of the cosmos. The sacred decad in particular has a cosmic significance in Pythagoreanism: its mystical name, *tetraktys* (meaning approximately "fourness"), implies $1 + 2 + 3 + 4 = 10$; but it can also be thought of as a "perfect triangle," as in Figure 1.

Speculation on number and proportion led to an intuitive feeling of the *harmonia* ("fitting together") of the *kosmos* ("the beautiful order of things"); and the application of the *tetraktys* to the theory of music (see below *Music)* revealed a hidden order in the range of sound. Pythagoras may have referred, vaguely, to the "music of the heavens," which he alone seemed able to hear; and later Pythagoreans seem to have assumed that the distances of the heavenly bodies from the earth somehow correspond to musical intervals—a theory that, under the influence of Platonic conceptions, resulted in the famous idea of the "harmony of the spheres." Though number to the early Pythagoreans was still a kind of cosmic matter, like the water or air proposed by the Ionians, their stress upon numerical proportions, harmony, and order comprised a decisive step toward a metaphysic in which form is the basic reality.

The doctrine of opposites. From the Ionians, the Pythagoreans adopted the idea of cosmic opposites, which they—perhaps secondarily—applied to their number speculation. The principal pair of opposites is the limit and the unlimited; the limit (or limiting), represented by the odd (3,5,7, . . .), is an active force effecting order, harmony, "cosmos," in the unlimited, represented by the even. All kinds of opposites somehow "fit together" within the cosmos, as they do, microcosmically, in an individual man and in the Pythagorean society. There was also a Pythagorean "table of ten opposites," to which Aristotle has referred—limit–unlimited, odd–even, one–many, right–left, male–female, rest–motion, straight–curved, light–darkness, good–evil, and square–oblong. The arrangement of this table reflects a dualistic conception, which was apparently not original with the school, however, or accepted by all of its members.

The Pythagorean number metaphysic was also reflected in its cosmology. The unit (1), being the starting point of the number series and its principle of construction, is not itself strictly a number; for, to be a number is to be even or odd, whereas, in the Pythagorean view, "one" is seen as *both* even and odd. This ambivalence applies, similarly, to the total universe, conceived as the One. There was also a cosmogonical theory (of cosmic origins) that explained the generation of numbers and number-things from the limiting-odd and the unlimited-even—a theory that, by stages unknown to scholars, was ultimately incorporated into Plato's philosophy in his doctrine of the derivation of sensed realities from mathematical principles (see PLATONISM AND NEOPLATONISM: *Greek Platonism front Aristotle through Middle Platonism, its nature and history).*

Mathematics and science. Pythagorean thought was scientific as well as metaphysical and included specific developments in arithmetic and geometry, in the science of musical tones and harmonies, and in astronomy.

Arithmetic. Early Pythagorean achievements in mathematics are unclear and largely disputable, and the following is, therefore, a compromise between the widely divergent views of modern scholars.

Use of the *gnōmones*

In the speculation on odd and even numbers, the early Pythagoreans used so-called *gnōmones* (Greek: "carpenter's squares"). Judging from Aristotle's account, gnomon numbers, represented by dots or pebbles, were arranged in the manner shown in Figure 2. If a series of odd numbers is put around the unit as gnomons, they always produce squares; thus, the members of the series 4, 9, 16, 25, . . . are "square" numbers. If even numbers are depicted in a similar way, the resulting figures (which offer infinite variations) represent "oblong" numbers, such as those of the series 2, 6, 12, 20. . . . On the other hand, a triangle represented by three dots (as in the upper part of the tetraktys) can be extended by a series of natural numbers to form the "triangular" numbers 6, 10 (the tetraktys), 15, 21. . . . This procedure, which was, so far, Pythagorean, led later, perhaps in the Platonic Academy, to a speculation on "polygonal" numbers.

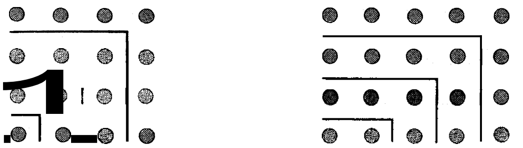
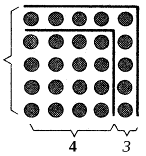


Figure 2: Gnomonsof Pythagorean numbertheory (see text).

Probably the square numbers of the gnomons were early associated with the Pythagorean theorem (likely to have been used in practice in Greece, however, before Pythagoras), which holds that for a right triangle a square drawn on the hypotenuse is equal in area to the sum of the squares drawn on its sides; in the gnomons it can easily be seen, in the case of a 3,4,5–triangle for example (see Figure 3), that the addition of a square gnomon number to a square makes a new square: $3^2 + 4^2 = 5^2$, and this gives a method for finding two square numbers the sum of which is also a square.



Fiaure 3: Gnomon for Pythagorean theorem. The matked off "carpenter's square" —comprising 3 groups of 3 dots each (3 X 3)—thus represents 3^2, which when added to 4^2 yields 5^2 (<he total gnomon).

The problem of irrational numbers

Some 5th-century Pythagoreans seem to have been puzzled by apparent arithmetical anomalies: the mutual relationships of triangular and square numbers; the anomalous properties of the regular pentagon; the fact that the length of the diagonal of a square is incommensurable with its sides—*i.e.*, that no fraction composed of integers can express this ratio exactly (the resulting decimal is thus defined as irrational); and the irrationality of the mathematical proportions in musical scales. The discovery of such irrationality was disquieting because it had fatal consequences for the naïve view that the universe is expressible in whole numbers; the Pythagorean Hippasus is said to have been expelled from the brotherhood, according to some sources even drowned, because he made a point of the irrationality.

In the 4th century, Pythagorizing mathematicians made a significant advance in the theory of irrational numbers, such as the-square-root-of-it ($\sqrt{n}$), it being any rational number, when they developed a method for finding progressive approximations to $\sqrt{2}$ by forming sets of so-called diagonal numbers.

Geometry. In geometry, the Pythagoreans cannot be credited with any proofs in the Euclidean sense. They were evidently concerned, however, with some speculation on geometrical figures, as in the case of the Pythagorean theorem, and the concept that the point, line, triangle, and tetrahedron correspond to the elements of the tetraktys, since they are determined by one, two, three, and four points, respectively. They possibly knew practical methods of constructing the five regular solids, but the theoretical basis for such constructions was given by non-Pythagoreans in the 4th century.

It is notable that the properties of the circle seem not to have interested the early Pythagoreans. But perhaps the tradition that Pythagoras himself discovered that the sum of the three angles of any triangle is equal to two right angles may be trusted. The idea of geometric proportions is probably Pythagorean in origin; but the so-called golden section—which divides a line at a point such that the smaller part is to the greater as the greater is to the whole—is hardly an early Pythagorean contribution. Some advance in geometry was made later, by 4th-century Pythagoreans; *e.g.,* Archytas offered an interesting solution to the problem of the duplication of the cube—in which a cube twice the volume of a given cube is constructed—by an essentially geometrical construction in three dimensions; and the conception of geometry as a "flow" of points into lines, of lines into surfaces, and so on, may have been contributed by Archytas; but on the whole the achievements of non-Pythagorean mathematicians were in fact more conspicuous than those of the Pythagoreans.

Mathematical ratios of tones

Music. The achievements of the early Pythagoreans in musical theory are somewhat less controversial. The scientific approach to music, in which musical intervals are expressed as numerical proportions, originated with them, as did also the more specific idea of harmonic "means." At an early date they discovered empirically that the basic intervals of Greek music include the elements of the tetraktys, since they have the proportions 1:2 (octave), 3:2 (fifth), and 4:3 (fourth). The discovery could have been made, for instance, in pipes or flutes or stringed instruments: the tone of a plucked string held at its middle is an octave higher than that of the whole string; the tone of a string held at the ⅔ point is a fifth higher; and that of one held at the ¾ point is a fourth higher. Moreover, they noticed that the subtraction of intervals is accomplished by dividing these ratios by one another. In the course of the 5th century they calculated the intervals for the usual diatonic scale, the tone being represented by 9:8 (fifth minus fourth); *i.e.,* $3/2 \div 4/3$, and the semitone by 256:243 (fourth minus two tones); *i.e.,* $4/3 \div (9/8 \times 9/8)$. Archytas made some modification to this doctrine and also worked out the relationships of the notes in the chromatic (12-tone) scale and the enharmonic scale (involving such minute differences as that between A flat and G sharp, which on a piano are played by the same key).

Astronomy. In their cosmological views the earliest Pythagoreans probably differed little from their Ionian predecessors. They made a point of studying the stellar heavens; but—with the possible exception of the theory of musical intervals in the cosmos—no new contributions to astronomy can be ascribed to them with any degree of probability. Late in the 5th century, or possibly in the 4th century, a Pythagorean boldly abandoned the geocentric view and posited a cosmological model in which the Earth, Sun, and stars circle about an (unseen) central fire—a view traditionally attributed to the 5th-century Pythagorean Philolaus of Croton (see PHYSICAL SCIENCES, HISTORY OF).

HISTORY OF PYTHAGOREANISM

Literary sources, especially Aristotelian

The life of Pythagoras and the origins of Pythagoreanism appear only dimly through a thick veil of legend and semihistorical tradition. The literary sources for the teachings of the Pythagoreans present extremely complicated problems. Special difficulties arise from the oral and esoteric transmission of the early doctrines, the profuse accumulation of tendentious legends, and confusion caused by the split in the school in the 5th century BC. In the 4th century, Plato's inclination toward Pythagoreanism created a tendency—manifest already in the middle of the century in the works of his pupils—to interpret Platonic concepts as originally Pythagorean. But the radical skepticism as to the reliability of the sources shown by some modern scholars has on the whole been abandoned in recent research. It now seems possible to extract bits of reliable evidence from a wide range of ancient authors, such as Porphyry and Iamblichus (see below Neo-Pythagoreanism).

Most of these literary sources hark back ultimately to the environment of Plato and Aristotle; and here the importance of one of Aristotle's students has become obvious, viz., the musicologist and philosopher Aristoxenus, who in spite of his bias possessed firsthand information independent of the point of view of Plato's Academy. The role played by Dicaearchus, another of Aristotle's pupils, and by the Sicilian historian Timaeus, of the early 3rd century BC, is less clear. Recently, the reliability of Aristotle's account of Pythagoreanism has also been emphasized against the doubts that had been expressed by some modern scholars; but Aristotle's sources, in turn, hardly lead farther back than to the late 5th century (perhaps to Philolaus; see below Two Pythagorean sects). In addition, there are scattered hints in various early authors and in some not very substantial remains of 4th-century Pythagorean literature. The mosaic of reconstruction thus has to be to some extent subjective.

Four chief periods

Early Pythagoreanism. Within the ancient Pythagorean movement four chief periods can be distinguished: early Pythagoreanism, dating from the late 6th century BC and extending to about 400 BC; 4th-century Pythagoreanism; the Hellenistic trends; and Neo-Pythagoreanism, a revival that occurred in the mid-1st century AD and lasted for two and a half centuries.

Background. The background of Pythagoreanism is complex, but two main groups of sources can be distinguished. The Ionian philosophers (Thales, Anaximander, Anaximenes, and others) provided Pythagoras with the problem of a single cosmic principle, the doctrine of opposites, and whatever reflections of Oriental mathematics there are in Pythagoreanism; and from the technicians of his birthplace, the Isle of Samos, he learned to understand the importance of number, measuring, and proportions. Popular cults and beliefs current in the 6th century and reflected in Orphism introduced him to occultism, ritualism, and the doctrine of individual immortality. In view of the shamanistic traits of Pythagoreanism, reminiscent of Thracian cults, it is interesting to note that Pythagoras seems to have had a Thracian slave.

Pythagorean communities. The school apparently founded by Pythagoras at Croton in southern Italy seems to have been primarily a religious brotherhood centred around Pythagoras and the cults of Apollo and of the Muses, ancient patron goddesses of poetry and culture. It became perhaps successively institutionalized and received different classes of esoteric members and exoteric sympathizers. The rigorism of the ritual and ethical observances demanded of the members is unparalleled in early Greece; in addition to the rules of life mentioned above, it is fairly well-attested that secrecy and a long silence during the novitiate were required. The exoteric associates, however, were politically active and established a Crotonian hegemony in southern Italy. About 500 BC a coup by a rival party caused Pythagoras to take refuge in Metapontum, where he died.

During the early 5th century, Pythagorean communities existed in many southern Italian cities, a fact that led to some doctrinal differentiation and diffusion. In the course of time the politics of the Pythagorean parties became decidedly antidemocratic. About the middle of the century a violent democratic revolution swept over southern Italy; many Pythagoreans were killed, and only a few escaped, among them Lysis of Tarentum and Philolaus of Croton, who went to Greece and formed small Pythagorean circles in Thebes and Phlious.

Two Pythagorean sects. Little is known about Pythagorean activity during the latter part of the 5th century. The differentiation of the school into two main sects, later called akousrnatikoi (Greek: akousma, "something heard," viz., the esoteric teachings) and *mathēmatikoi* (Greek: *mathēmatikos,* "scientific"), may have occurred at that time. The acousmatics devoted themselves to the observance of rituals and rules and to the interpretation of the sayings of the master; the "mathematics" were concerned with the scientific aspects of Pythagoreanism. Philolaus, who was rather a mathematic, probably published a summary of Pythagorean philosophy and science in the late 5th century.

Archytas and his school

Fourth-century Pythagoreanism. In the first half of the 4th century, Tarentum, in southern Italy, rose into considerable significance. Under the political and spiritual leadership of the mathematic Archytas, a friend of Plato, Tarentum became a new centre of Pythagoreanism, from which acousmatics—so-called Pythagorists who did not sympathize with Archytas—went out travelling as mendicant ascetics all around the Greek-speaking world. The acousmatics seem to have preserved some early Pythagorean Hieroi Logoi and ritual practices. Archytas himself, on the other hand, concentrated on scientific problems, and the organization of his Pythagorean brotherhood was evidently less rigorous than that of the early school. After the 380s there was a give-and-take between the school of Archytas and the Academy of Plato, a relationship that makes it almost impossible to disentangle the original achievements of Archytas from joint involvements (but see above, Geometiy and Music).

The Hellenistic Age. Whereas the school of Archytas apparently sank into inactivity after the death of its founder (probably after 350 BC), the Academics of the next generation continued "Pythagorizing" Platonic doctrines, such as that of the supreme One, the indefinite dyad (a metaphysical principle), and the tripartite soul (see PLATONISM AND NEOPLATONISM). At the same time, various Peripatetics of the school of Aristotle, including Aristoxenus, collected Pythagorean legends and applied contemporary ethical notions to them. In the Hellenistic Age, the Academic and Peripatetic views gave rise to a rather fanciful antiquarian literature on Pythagoreanism. There also appeared a large and yet more heterogeneous mass of apocryphal writings falsely attributed to different Pythagoreans, as if attempts were being made to revive the school. The texts fathered on Archytas display Academic and Peripatetic philosophies mixed with some notions that were originally Pythagorean. Other texts were fathered on Pythagoras himself or on his immediate pupils, imagined or real. Some show, for instance, that Pythagoreanism had become confused with Orphism; others suggest that Pythagoras was considered a magician and an astrologist; there are also indications of Pythagoras "the athlete" and "the Dorian nationalist." But the anonymous authors of this pseudo-Pythagorean literature did not succeed in re-establishing the school, and the "Pythagorean" congregations formed in early imperial Rome seem to have had little in common with original Pythagoreanism; they were ritualistic sects that adopted, eclectically, various occult practices.

Amalgamation with Neoplatonism

Neo-Pythagoreanism. With the ascetic sage Apollonius of Tyana, about the middle of the 1st century AD, a distinct Neo-Pythagorean trend appeared. Apollonius studied the Pythagorean legends of the previous centuries, created and propagated the ideal of a Pythagorean life—of occult wisdom, purity, universal tolerance, and approximation to the divine—and felt himself to be a reincarnation of Pythagoras. Through the activities of Neo-Pythagorean Platonists, such as Moderatus of Gades, a pagan trinitarian, and the arithmetician Nicomachus of Gerasa, both of the 1st century AD, and, in the 2nd or 3rd century, Numenius of Apamea, forerunner of Plotinus (an epoch-making elaborator of Platonism), Neo-Pythagoreanism gradually became a part of the expression of Platonism known as Neoplatonism; and it did so without having achieved a scholastic system of its own. The founder of a Syrian school of Neoplatonism, Iamblichus, a pupil of Porphyry (who in turn had been a pupil of Plotinus), thought of himself as a Pythagorean sage and about AD 300 wrote the last great synthesis of Pythagoreanisrn, in which most of the disparate post-classical traditions are reflected. It is characteristic of the Neo-Pythagoreans that they were chiefly interested in the Pythagorean way of life and in the pseudoscience of number mysticism. On a more popular level, Pythagoras and Archytas were remembered as magicians. Moreover, it has been suggested that Pythagorean legends also influenced the Christian monastic tradition.

Medieval and modern trends. In the Middle Ages the popular conception of Pythagoras the magician was combined with that of Pythagoras "the father of the quadrivium"; *i.e.*, of the more specialized liberal arts of the curriculum. From the Italian Renaissance onward, some "Pythagorean" ideas, such as the tetrad, the golden section, and harmonic proportions, became applied to aesthetics. To many Humanists, moreover, Pythagoras was the father of the exact sciences. In the early 16th century, Nicolaus Copernicus, who developed the view that the Earth revolves around the Sun, considered his system to be essentially Pythagorean or "Philolaic," and Galileo was called a Pythagorean. The 17th-century Rationalist G.W. Leibniz appears to have been the last great philosopher and scientist who felt himself to be in the Pythagorean tradition.

It is doubtful whether advanced modern philosophy has ever drawn from sources thought to be distinctly Pythagorean. Yet Platonic–Neoplatonic notions, such as the mathematical conception of reality or the philosopher's union with the universe and various mystical beliefs are still likely to be stamped as Pythagorean. Even today uncritical admiration of Pythagoreanism is common.

EVALUATION

The history of the projection of Pythagoreanism into subsequent thought indicates how fertile some of its core concepts were. Plato is here the great catalyst; but it is possible to perceive behind him, however dimly, a series of Pythagorean ideas of paramount potential significance: the combination of religious esoterism (or exclusivism) with the germs of a new philosophy of mind, present in the belief in the progress of the soul toward the actualization of its divine nature and toward knowledge; stress upon harmony and order, and upon limit as the good; the primacy of form, proportion, and numerical expression; and in ethics, an emphasis upon such virtues as friendship and modesty. The fact that Pythagoras, to later ages, also became alternatively a Dorian nationalist, a sportsman, an educator of the people, or a great magician is a more curious consequence of the productivity of his teaching.

BIBLIOGRAPHY

Fragments and texts: The collection of the fragments in HERMANN DIELS and WALTHER KRANZ, *Die Fragmente der Vorsokratiker,* 6th ed., vol. 1 (1951), is insufficient; additions are given in MARIA TIMPANARO CARDINI (ed.), *Pitagorici: Testimonianze e frammenti,* 3 vol. (1958–64); and in CORNELIA J. DE VOGEL, *Pythagoras and Early Pythagoreanism* (1966). For the pseudo-Pythagoreans, see HOLGER THESLEFF (ed.), *The Pythagorean Texts of the Hellenistic Period* (1965).

Early Pythagoreanism: The best comprehensive introduction to Pythagoreanism is the long chapter "Pythagoras and the Pythagoreans," in W.K.C. GUTHRIE, *A History of Greek Philosophy,* vol. 1, pp. 146–340 (1962). Somewhat different approaches have been taken by DE VOGEL (*op. cit.*); and JAMES A. PHILIP, *Pythagoras and Early Pythagorearzism* (1966), works that demand more active criticism by the reader. Fairly full references to the discussion of Pythagoreanism up to 1960 are in WALTER BURKERT, *Weisheit und Wissenschaft: Studien zu Pythagoras, Philolaos, und Platon* (1962; Eng. trans., *Lore and Science in Ancient Pythagoreanism,* 1972), a highly technical and at times rather overcritical work. Among later technical discussions, likely to become influential, are the articles "Pythagoras" and "Pythagoreer" in *Pauly-Wissowa Realencyclopadie,* vol. 47, (1963), and suppl. vol. 10 (1965)--of the contributors, KURT VON FRITZ and H. DORRIE arrive at less controversial conclusions than B.L. VAN DER WAERDEN.

Hellenistic Pythagoreanism: HOLGER THESLEFF, *An Introduction to the Pythagorean Writings of the Hellenistic Period* (1961); additions and corrections in *Entretiens Fondation Hardt,* vol. 18 (1972).

Neo-Pythagoreanism: PHILIP MERLAN, *From Platonism to Neoplatonism* (1953).

For up-to-date bibliographical accession, see *L'Année philologique* (annual), under the subject heading "Pythagorica" and the various Pythagoreans.

(H.T.)

Quarantine and Isolation

The words quarantine and isolation, which by popular usage can be employed almost interchangeably in reference to disease control, in the present article are used in the following technical senses: Quarantine is the detention or restraint of human beings or other creatures that may have come into contact with communicable disease until it is deemed certain that they have escaped infection, while isolation is the separation of an infected individual from the healthy until he is unable to transmit the disease.

Early practices with respect to leprosy and the plague

History of quarantine and isolation. The earliest recognition that diseases might be communicable led to extreme measures designed to isolate infected persons or communities. Fear of leprosy, a slowly progressive mutilating disease, caused wide adoption of the control measures set out in Leviticus 13, namely, isolation of the infected and the cleansing or burning of his garments. The leper, until only recently, was considered an outcast, and no healthy person would dare have any communication with him. Against acute, highly fatal diseases like bubonic plague, which spread rapidly throughout populations, attempts were made by healthy communities to prevent entry of goods and persons from infected communities; in the 7th century AD, for example, armed guards were stationed between plague-stricken Provence and the diocese of Cahors. Over a thousand years later, in 1720, when Marseilles was suffering a severe epidemic of the plague, a ring of sentries was placed around the city to prevent any person's escaping. "Cordons sanitaire" of this kind were probably an ineffective means of control.

In the 14th century the growth of maritime trade and the recognition that plague was introduced by ships returning from the Levant led to the adoption of quarantine in Venice. It was decreed that ships were to be isolated for a limited period to allow for the manifestation of the disease and to dissipate the infection brought by persons and goods. Originally the period was 30 days, *trentina,* but this was later extended to 40 days, *quarantina.* The choice of this period is said to be based on the period that Christ and Moses spent in isolation in the desert. In 1423 Venice set up its first lazaretto, or quarantine station, on an island near the city. The Venetian system became the model for other European countries and the basis for widespread quarantine control for several centuries.

In the 16th century the system was extended by the introduction of bills of health, a form of certification that the last port of call was free from disease; a clean bill, with the visa of the consul of the country .of arrival, entitled the ship to free pratique (use of the port) without quarantine. Quarantine was later extended to other diseases besides plague, notably yellow fever, with the growth of American trade, and cholera, which was particularly associated with the pilgrimages to Mecca.

By the mid-19th century the practice of quarantine had become a considerable nuisance. The periods of quarantine were arbitrary and variable from country to country, and there were instances of perverse and bureaucratic application of the quarantine regulations. The disinfection of letters and rummaging of papers could be an excuse for political espionage, and the opportunities for bribery and corruption were frequently exploited. Great discomfort and delay was caused to travellers; the prison reformer John Howard had, in 1786, deliberately sailed from Smyrna to Venice in a ship with a foul bill of health so that he could gain firsthand experience of lazarettos; his account (*An Account of the Principal Lazarettos in Europe* [1789]) presents a depressing picture.

First international sanitary conference

General dissatisfaction with quarantine practice led to the convening of the first international sanitary conference in Paris in 1851. The arguments were conducted at two levels. Commercially, the conflict was between the countries with considerable vested interests in quarantine and the major maritime nations, which favoured its abolition; medically, the opposition was between the "contagionists," who believed that diseases like cholera and plague are transmitted from person to person, and the "miasmatists," who thought that they are caused by infected atmosphere and that the remedy is sanitation, not quarantine. Despite these differences, agreement was reached on some important general principles for the standardization of quarantine procedures. Since the convention and regulations were not generally ratified, there was little immediate impact from it.

In the next fifty years a succession of sanitary conferences, with better understanding of the epidemiology of communicable disease, reached some agreement on the maximum permissible measures of control and on the removal of the most irksome restrictions of quarantine practice, but the accord reached by the 11th conference, at Paris in 1903, was the first really effective measure to be signed. Out of it came in 1907, the Office International d'Hygiène Publique ("International Office of Public Health"), the forerunner of the World Health Organization. (The forerunner of the Pan American Sanitary Bureau had been established five years earlier, in 1902).

Present practices. Today, isolation of persons is practiced much less rigidly or extensively than formerly in the control of communicable disease. It is clearly appropriate, and adopted, in dangerous diseases spread directly from person to person, such as smallpox or typhoid. It is recognized, however, that isolation may fail for a variety of reasons—it depends, for example, on complete and immediate recognition of infective persons, which is rarely achieved even for those with open signs of disease. It is ineffective in diseases that are transmitted by an intermediate carrier—*e.g.*, the mosquito in yellow fever and malaria. In plague, isolation is important to prevent man-to-man spread but would not have any effect on the main route of infection—by bites of the rat flea. It is inappropriate to isolate human cases of a disease, such as brucellosis, that is usually acquired by contact with infected farm animals or their products. Even for diseases in which it may protect individuals, isolation will often have little effect on the general epidemic; this may be, as in measles, because infectivity precedes the appearance of the characteristic feature, the rash, by a few days, or, as in polio virus infections, because a number of persons are carriers, harbouring the disease agent without discernible illness. The difficulty of recognizing potentially infective persons often makes isolation impracticable even in situations in which it could be appropriate.

Current use of quarantine

Quarantine is much modified in modern practice because of the better understanding of how communicable disease is transmitted. In its purest form it is applied to animals, as in the control of rabies; recent experience in Great Britain has revealed difficulties because dogs have developed rabies after a quarantine period of six months, generally accepted as a maximum reasonable incubation period. In man it is now more usual to apply *surveillance* of contacts, with, possibly, daily reporting to a doctor to get prompt recognition of illness but without restricting movement; such a policy, coupled with other control measures, is now accepted even in such diseases as smallpox. In some instances modified quarantine is imposed; adult contacts of typhoid should be excluded from food handling until repeated bacteriological examination of feces and urine has shown them to be free of the disease; susceptible children exposed to measles have sometimes been excluded from school, but the practice was already declining even before the widespread use of measles vaccines. Quarantine and exclusion of plants and of plant products are still widely practiced in accordance with international agreements (see DISEASES OF ANIMALS; DISEASES OF PLANTS).

BIBLIOGRAPHY. WORLD HEALTH ORGANIZATION, *International Sanitary Regulations,* 3rd annotated ed. (1966), lists the provisions designed to combine the minimum interference with world traffic with the maximum security against international spread of the six quarantinable diseases: plague, cholera, yellow fever, smallpox, typhus, and relapsing fever. These regulations have been superseded in part by the *International Health Regulations,* adopted by the 22nd World Health Assembly, 1969. These new regulations apply to cholera, plague, smallpox, and yellow fever, but not to typhus and relapsing fever.

AMERICAN PUBLIC HEALTH ASSOCIATION, *Control of Communicable Disease in Man,* 10th ed. (1965), a widely used reference work on most diseases of public health importance, including those occurring outside the Western Hemisphere; N.M. GOODMAN, *International Health Organizations and Their Work* (1952), a good summary of the history of quarantine and its influence on the development of international health work; C.E.A. WINSLOW, *The Conquest of Epidemic Disease* (1943), a wide-ranging and scholarly historical account; H. ZINSSER, *Rats, Lice and History* (1935), a classic history of typhus and plague; G. SMITH, *Plague on Us* (1941), a popular account of the epidemiology of communicable disease.

(J.Kn.)

Quasi-stellar Sources

Quasi-stellar sources (also often called quasi-stellar objects, quasars, or QSO's) form a class of astronomical objects with starlike point images that show large shifts to the red of the lines in their optical spectra. (A spectrum is the band of colour formed when light passed through a slit is then broken up instrumentally into its constituent wavelengths and is a continuous and overlapping series of images of the slit; dark or bright lines in a spectrum give clues to the chemical composition of the object and also to its temperature and other physical conditions.) No final agreement has been reached regarding the name of these objects and in this article they will be referred to as QSO's. As existing distance estimates are not fully reliable, the amount of light and in many cases radio energy released is not finally known either. QSO's could be among the intrinsically brightest objects in the sky; or the energy release might be more moderate if they were relatively near objects rather than among the most distant as was at first thought.

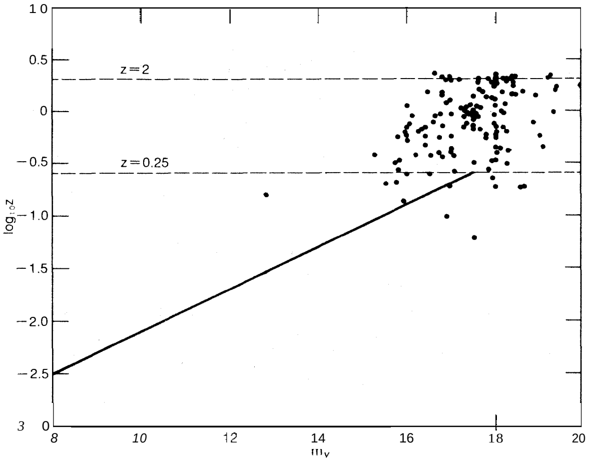
Discovery. The first QSO's were discovered by radio astronomers in 1960. Until then all radio sources outside the Galaxy that had been identified with optical objects had been found to have their origin in external galaxies. Accurate radio positions for three such radio sources, designated in the third Cambridge radio source catalog (3C) as 3C 48, 3C 196, 3C 286, led to the discovery of starlike objects, at those positions. The object 3C 48 was studied extensively by optical astronomers, but its line spectrum could not be understood. In 1962, a much brighter starlike object, 3C 273, was identified by the use of a radio telescope in Australia. The line spectrum of this optical object was finally identified as due to lines found in laboratory spectra in the ultraviolet and normally not visible in the spectrum of a star. The lines in the spectra of nearly all galaxies are shifted toward the red end of the spectrum; the red shift, discussed below, is defined as the observed wavelength minus the laboratory wavelength divided by the laboratory wavelength. The red shift for 3C 273 was 0.158. The red shift of 3C 48 was determined to be 0.367. Following this breakthrough, several quasi-stellar radio sources were identified and their red shifts measured.

In general, the red shifts turned out to be far larger than any that had been detected for ordinary galaxies. By 1965, at least one object with a red shift approximately equal to two had been discovered. The early—and possibly too quick—explanation was that the red shifts, like the red shifts of many galaxies, were due to expansion of the universe and would be proportional to their distance from the observer. It was therefore concluded that the quasi-stellar radio sources must be enormously far away and intrinsically much more luminous, up to 100 times brighter than the brightest galaxies. The fact that they emitted much more energy in the ultraviolet part of the spectrum than normal stars led to the conclusion that such objects might be found by optical survey methods alone. The search proved successful. By 1970 it had been shown that, down to the faintest objects that can be detected with the large telescopes, there are something like 200 quasi-stellar objects per square degree. If they are uniformly distributed, some 10,000,000 must exist, of the 22nd magnitude or brighter. (The visually brightest stars are of about magnitude −2. A difference of 5 magnitudes is equivalent to 100-fold difference in brightness, and the notation is such that the brighter object has the smaller magnitude.) Only a fraction of the QSO's known are radio sources.

Probable numbers of QSO's

Observed properties. *Emission-line red shifts.* By 1970 about 200 QSO's had been studied and their red shifts determined. The optical spectra in general show broad emission lines. (When the light from a hot incandescent gas is recorded directly, the lines in the spectrum are bright and are called emission lines; when the light from a hot source passes through a cooler region of the same gas before it is analyzed, some dark lines called absorption lines are formed at the same wavelengths.) The QSO lines are superposed on a continuous bright spectrum of light from the source, called a continuum, which has a nonthermal origin—that is, it is not simply due to radiation from a hot gas cloud. The optical emisson lines of the QSO's are mostly the strongest lines of some of the common elements. These include lines that would be produced in the laboratory at wavelengths of 2796 and 2803 angstroms by once-ionized magnesium (atoms with one electron missing); at 1548 and 1551 angstroms by three-times-ionized carbon; at 1909 angstroms by twice-ionized carbon; and the two strongest lines of the so-called Lyman series of hydrogen. None of these lines can be normally seen in the spectra of astronomical objects because their wavelengths, in the ultraviolet range, are absorbed by the Earth's atmosphere. In the QSO's the red shift moves them into the visible part of the spectrum. The range of the red shifts known at present is from approximately 0.06 to approximately 2.89. A plot of apparent brightness against red shifts (see the Figure) gives what is practically a scatter diagram; that is, there is very little correlation of optical brightness with red shift. The only striking feature appears to be the dearth of objects with red shifts larger than about 2.2. This cutoff is thought to be real, but the cutoff in apparent brightness—there are practically no QSO's identified that have apparent brightness less than magnitude 19.5—is due to the extreme difficulty in identifying or observing QSO's fainter than this.

From G.R. Burbidge and E.M. Burbidge. "A Progress Report," *Nature*, vol. 224



Plot of apparent brightness, m_V, in magnitudes against the logarithm of the red shift, *z*, for 136 QSO's. The line shows the relation between red shift and apparent brightness found for standard galaxies (see text).

Absorption-line red shifts. Since 1966 a considerable number of QSO's with absorption lines in addition to the emission lines in their spectra have been found. One QSO in which only absorption lines are seen has been found. In most of these it is found that the red shift values calculated from the emission and the absorption lines are about the same; there are several cases in which they differ by small amounts. Several have been found, however, to have many different values of the absorption red shifts, some much smaller than the emission red shifts; in one case seven have been identified. The QSO's with rich absorption-line spectra have large emission-line red shifts.

Flux variations and sizes. A general property of the QSO's studied so far is that they show time variations in the optical and radio energy flux; in the optical region, the energy flux is the brightness of the object measured in magnitudes. In some cases large changes amounting to factors of ten or more have taken place in less than a year's time. Measurable variations are known to occur on time scales as short as days. At the time that these variations were discovered, no other extragalactic objects were known to be variable, but later it was shown that the nuclei of some galaxies that are very similar in their radiating properties to the QSO's are also variable. From the time scale of the variation an upper limit to the size can be found, provided it is assumed that the radiating surface is not expanding relativistically—that is, at speeds approaching that of light. This limiting value is the time scale multiplied by the velocity of light. Thus, the sizes of the radiating regions are estimated to be no larger than light-years (10^{13} kilometres) or light-days (2.5 × 10^{10} kilometres). The blurring of the optical images caused by atmospheric effects means then that with optical telescopes on the surface of the Earth, man cannot measure the diameters of starlike objects to be less than about 0.5 seconds of arc. At quite modest distances in the universe this angular size corresponds to a linear size much bigger than light-years (*e.g.*, for a galaxy 30,000,000 light-years away, in the local Virgo cluster of galaxies, 0.5 seconds of arc will equal about 75 light-years). Radio astronomers can now measure angular sizes as small as 10^{-3} second of arc and have found that many QSO's and nuclei of galaxies that emit measurable radio-wavelength radiation contain very small nuclear emitting regions. For the nuclei of nearby galaxies, the sizes that are measured are compatible with the upper limits to the sizes set by the variations.

Size of QSO's

The radio variations are restricted to high frequencies (radio wavelengths less than about 20 centimetres) and are compatible with a model in which it is supposed that successive bursts of relativistic electrons expand outward from a central object. In general, no obvious pattern can be discerned in the optical variations though it is suspected that, in one or two QSO's, a component is periodic.

The nature of the red shift. Since the discovery of the QSO's there has been uncertainty concerning their distances. Though the simple explanation of the large red shifts were that they are due to the expansion of the universe, and that the objects are therefore exceedingly luminous and very distant, the discovery of the variable nature of the QSO's first led some workers to question this interpretation. The reasons why variations caused some doubts about the cosmological nature of the red shifts were in part intuitive—the sense that very distant objects 100 times more luminous than galaxies could not be so small that they could vary significantly in times as short as years or less. When translated to physical terms, these arguments led to the conclusion that there were severe difficulties in deriving self-consistent physical explanations for the phenomena if the distances were great, because the very high radiation density that must be present in the very small source must react strongly with its sources. If the QSO's are much closer than their red shifts would imply, their luminosities are much smaller, and the problem is much less severe. It is known also, however, that the nuclei of galaxies that are known to be close also vary, and the same problem, though in much milder form, arises.

Doppler red shifts

If the QSO's are much closer than the distance calculated from these red shifts, their large red shifts must be due to something other than the expansion of the universe alone. The only known mechanisms that could explain the red shifts are the Doppler effect or the presence of very strong gravitational fields. Any other explanation would be outside the laws of physics as they are now understood. Attempts to explain the red shifts by Doppler effect —*i.e.*, by arguing that the objects have all been thrown out of galaxies at very high speeds and are receding from the Galaxy so that large red shifts are seen (an object with a red shift of 2 must be receding with a velocity 80 percent that of light) have not been convincing. Since no blue shifts have been seen, there is no evidence that any objects are coming toward the Galaxy. It is known that large gravitational red shifts can arise if atomic transitions occur in regions in very strong gravitational fields, but no realistic models have been proposed on the assumption that the line spectra observed in the QSO's are produced in this way. Thus, without appealing to new effects altogether, theoretical arguments suggest that the red shifts are definitely cosmological in origin.

The distribution of the red shifts does give some indication that causes other than the expansion of the universe are at work. There is preliminary evidence of a sharp peak in the red shifts both of emission and of absorption lines at a value of 1.95. Like the scarcity of red-shift values above 2.2, the peak at 1.95 may have either a cosmological (caused by distance) or intrinsic origin. If position of the peak is a result of an intrinsic property of QSO's, then the width of the peak gives an indication of their distances. Specifically, the distance-caused component of their red shifts must be less than or about equal to 0.01; they must be closer than about 200,000,000 light-years (see Figure).

The problem is that there is no way of estimating distances of QSO's other than by assuming that the red shifts have an origin similar to that of the red shifts of galaxies that obey the Hubble expansion law; *i.e.*, the red shift is proportional to the distance from the observer. There is, however, no discemable Hnbble relation for the QSO's alone. Though this does not mean that their red shifts are not cosmological in origin, it does mean that there is at present no direct proof that the red shifts do have this origin. Apart from using the red shift, the only other known method that can be used to measure the distance of an extragalactic object is to establish that it is physically associated with another object whose distance is known. Attempts to measure distances of QSO's using this method have been tried. One QSO has recently been found to have the same red shift as an abnormal galaxy, and they both appear to lie in a cluster of faint galaxies. Three QSO's with large red shifts have been found to lie very close in angular measure to the centres of galaxies with very small red shifts. In each case the a priori probability that the QSO's are not associated with the galaxies is rather small. These latter results, however, have not been taken as strong evidence that the qso's are local. Investigations of the distribution of radio sources, some of which are identified with QSO's, in regions of the sky very close to nearby peculiar galaxies (galaxies with unusual. characteristics; see GALAXIES: EXTERNAL) have led to the conclusion that these radio sources and peculiar galaxies are physically associated. This evidence is not yet generally accepted as sufficient, and the result is not statistically sound.

It was suggested at an early stage that if the QSO's lie at cosmological distances, absorption features might be seen in the spectra because of absorption of some of the light by intergalactic gas lying in the light path between the solar system and the qso's. The discovery of such absorption would have established that the qso's lie at cosmological distances and would also have provided a method of investigating the intergalactic medium. The observations have shown, however, that no detectable continuous absorption caused by smoothly distributed neutral atomic hydrogen is present, meaning that either the density of this gas in intergalactic space is less than one atom per 10'' cubic centimetres — a fantastically low value — or else that the QSO's are local. A variety of spectroscopic and statistical arguments also suggest that the sharp absorption lines seen in some QSO spectra mentioned earlier arise in the objects and not in the intergalactic medium. Thus, this test also has failed.

At present it must be concluded that the distances of the QSO's are not known.

Physical models of QSO's. It is generally agreed that a QSO consists of an extremeiy small, very massive nucleus, with a total size of less than a light-year, which is surrounded by an extended halo of gas that is being excited by the energy radiated by the central object. The line spectra, both emission and absorption, arise in different parts of this cloud. The electron temperature, the temperature calculated from the physical condition of the electrons in the cloud, and electron densities in the gas are approximately 30,000" K, and approximately 100 to 1,000,000 electrons (and an equal number of protons) per cubic centimetre. Calculations show clearly that there is a widespread range of values in different QSO's. The composition of the gas regarding those elements that can be analyzed appears to be similar to that in gaseous nebulae in the Galaxy. The multiple absorption red shifts are probably to be associated with shells of gas that are moving at speeds ranging from about 1,000 kilometres per second to about 150.000 kilometres per second relative to the central object.

This central object emits radiation over a wide spectral range and the basic radiation mechanism is thought to be the synchrotron process; that is, radiation by relativistic electrons moving in magnetic fields that also occurs in artificial particle accelerators called synchrotrons. The intensity of the radiation varies as a power of the wavelength. A graph of this relationship, in many cases, steep-*ens* from the optical into the infrared region of the spectrum. The bulk of the power is radiated in the optical and the infrared regions. Scattering of the photons by electrons·— the Compton effect.—may be important in some situations.

Energy released by QSO's

It is generally agreed that the energy emitted by QSO's, by strong radio sources, and by galactic nuclei is gravitational energy and not thermonuclear in origin. It has been argued that it is generated in multiple supernova outbursts, each supernova collapsing and releasing a large amount of gravitational energy, or in collisions between stars, or in the gravitational collapse of a single supermassive star, or, related to this, that it is derived from the rotational energy of a supermassive star. The total energy released is determined by the luminosity (the total brightness) and the lifetime of the object. The luminosity is uncertain because it depends on the distance assumed. If the red shifts are cosmological in origin the luminosities are 10^{46}–10^{47} ergs per second, while if the objects are only at distances 100,000,000 light-years, the luminosities are 10^{42}–10^{43} ergs per second. The lifetimes in the luminous phase are estimated to be 10^{6}–10^{7} years, so that the total energy released lies in the range 10^{56}–10^{62} ergs or the rest-mass energy (the amount of energy equivalent. to a certain mass; mass and energy are sometimes interchangeable) corresponding to masses of from 100 to 10^{9} times the mass of the Sun. In the case of the strongly radiating radio galaxies, in which the distances are not in doubt, the total energy released is known to be 10^{60}–10^{62} ergs. Since according to general relativity only a few percent of the rest-mass energy is likely to be released even in the most favourable situations, it is clear that very large masses must be present in the very small volumes observed. Theoretical investigations of this problem have not led to any detailed acceptable theory. There are two basic problems: to understand how very large mass concentrations are produced and how they evolve, and then to understand the details of. the mechanisms by which a large part of the available energy is transformed into relativistic particles, which then radiate their energy via the synchrotron process and Compton scattering.

Violent activity in galaxies

It is now realized that the QSO phenomenon is closely related to violent activity that takes place over a wide range of power levels in the nuclear regions of many, and perhaps all, galaxies. This activity is manifested by the generation of nonthermal radio, infrared, and optical radiation; by the excitation of large masses of gas; and by the ejection of comparatively cool matter and relativistic plasma, a gas formed of electrically charged particles moving at speeds approaching that of light. The latter gives rise to nonthermal radio emission far from the centres of many galaxies. The nucleus of the Galaxy, of which the solar system is a part, shows some of these violent characteristics, as do many nearby galaxies, but the power level is low, so that they would remain undetected if they were at great distances. The fact that so many galaxies are known to behave in this way suggests that the phenonienon takes place fairly continuously over much of the life of a galaxy. An understanding of the sequence of events that leads to this situation will only come when it is understood how galaxies form and evolve, and this problem cannot be disentangled from cosmology. If the universe is evolving, and if the galaxies formed from diffuse matter at an early stage soon after a very dense beginning phase (a view that is taken by many), then it is necessary to understand how a very compact nucleus is formed through successive evolutionary stages going to higher and higher densities so that eventually gravitational energy release leads to the activity that is now seen. An alternative possibility advocated originally in another context by Sir James Hopwood Jeans in the 1920's, and later by some advocates of the steady-state (nonevolutionary) cosmology, is that the nuclei of galaxies are the places in which matter is being created, and that the activity in QSO's and in nuclei in general is a manifestation of this. The Soviet astronomer and academician V.A. Ambartsumian, who already proposed in 1958 that this type of activity would turn out to be of the greatest importance, has argued, within the framework of an evolving universe, that the violent activity is caused by small regions in the initial dense configuration in which expansion has been delayed by varying amounts.

Plainly the problems are of great significance since this area may be one in which astronomy will have a direct impact on fundamental physics, but no clear picture has emerged so far.

There is a clear continuity between the QSO phenomenon and the activity in the nuclei of galaxies. If it is accepted that the red shifts are all due to the expansion of the universe, the QSO's are some of the most energetic events of the type seen in galactic nuclei, and at a time in the past corresponding to red shifts of approximately two, the universe went through a phase at which there was much more activity than there is at present. If the QSO's are local objects, their energetic properties are no greater than those of many galactic nuclei, and the powerful radio galaxies are the most energetic manifestations of nuclear activity so far discovered. In this picture, however, the existence of large intrinsic red shifts must be related to the physics of the nuclei themselves.

BIBLIOGRAPHY. G.R. and E.M. BURBIDGE, ***Quasi-Stellar Objects*** (1967); and F.D. KAHN and H.P. PALMER, ***Quasars*** (1967), provide comprehensive reviews of the subject to 1966. Review articles of more recent developments may be found in journals. See esp. ***A. Rev. Astr. Astrophys.***, 5:399–452 (1967), 7:527–552 (1969), 8:369–460 (1970); ***Nature***, 224: 21–24 (1969); and ***Scient. Am.***, 215:40–52 (1966), 223:22–29 (1970).

(G.B.)

Quebec (Province)

One of the 10 provinces of Canada, Quebec is bounded on the north by Hudson Strait and Ungava Bay, on the east by Labrador, on the southeast by the Gulf of St. Lawrence and New Brunswick, on the south by the United States (Maine, New Hampshire, Vermont, and New York), and on the west by Ontario and Hudson Bay. Its 594,860 square miles (1,540,680 square kilometres) of land and water make it the largest Canadian province in area, while its more than 6,000,000 inhabitants make it the second most populous, after Ontario. Its capital, Quebec city, is the oldest city of Canada, and its metropolis, Montreal, is the largest city in Canada.

Historical and contemporary complexitiesof Quebec

To understand present-day Quebec, however, one must see the province against a background that goes back to the creation of the French colony in North America during the 16th century. Most observers would agree that the single most important theme in Quebec's history since the British acquisition of New France in 1763 has been the continuous attempt to achieve an accommodation between the numerically dominant French-speaking population and the economically dominant English-speaking one. Whatever changes in geographical size or political institutions have taken place in the province, life in Quebec has always been marked by a collective effort to maintain a distinct society. This characteristic in the second half of the 20th century was at the core of debates over the future of the federal structure of all Canada.

The present Province of Quebec was created in 1867, after being the colony of New France for over two centuries until it was ceded to Britain in 1763. Named the Province of Quebec between 1763 and 1791, it then became the Province of Lower Canada until 1841, and then the District of Canada East until 1867. During these earlier periods its geographical boundaries were changed arbitrarily, and only in the 20th century, with the reacquisition of the northern part of Quebec, did it acquire its present size. Even today, however, there are problems about the eastern boundaries, because no Quebec government has accepted a 1927 decision of the British Privy Council to award Labrador to Newfoundland.

Quebec's size and its boundaries are not the most important influences on its life. Profoundly marked by 18th-century wars between France and Britain over their North American territories and by difficulties between the two linguistic groups since 1763, the social, economic, and political institutions of Canada and of Quebec have been unable to solve these tensions. Because only in Quebec and at the level of the federal government are French and English on an equal footing as official languages, French-Canadians have felt that they are threatened as a minority group in Canada. In Quebec, where they comprise about 80 percent of the population, they maintain that the situation remains discriminatory. Control by the English minority in Quebec of most economic activities of the province has generally led to exclusion of the French-Canadians from opportunities of economic advancement.

Although men of goodwill on both sides have tried to find a lasting solution, the economic and social inequalities have created a growing nationalism among French-Canadians and a feeling among some of them that only the separation of Quebec from Canada can solve their problems. Events since the late 1960s have shown, however, that neither extreme nationalism nor separatism is yet accepted by the majority of those who live in Quebec. For information on related topics, see the articles CANADA and CANADA, HISTORY OF. Features of the province are covered in MONTREAL; QUEBEC (CITY); SAINT LAWRENCE RIVER; and SAINT LAWRENCE SEAWAY.

THE HISTORY OF QUEBEC

Antecedents of crisis. When in 1534 the French explorer Jacques Cartier landed at present-day Gaspé and took possession of the land in the name of the King of France, he brought with him the traditions of mercantile expansion of 16th-century Europe to this land where Indians and Eskimos had been living for some thousands of years. That some school books start the "real" history of Quebec with the founding of Quebec city by Samuel de Champlain in 1608 reflects the mercantilist attitudes of the 16th century. The debate among historians as to what comprises the real history of Quebec, however, is not limited to the question of who were the first occupants. Because the Province of Quebec as a political and geographical entity was created by the proclamation of 1763, the notion is sometimes also advanced that its real history should start with the capitulation of the French Army in 1760.

Problems of cultural and national identity

The various definitions given by historians are not simply semantic questions, for they contain diverse assumptions concerning the political identity of the Quebec government. For example, there is a political tradition among French-Canadians that the government of Quebec is also the government of the French-Canadian people, and, therefore, they are heir to what was New France. The 1966–67 ***Annuaire du Québec*** ***(Que'bec Yearbook)*** states this claim most clearly:

> Quebec is a state with limited responsibilities that belongs to the Canadian Federation as a province. It is also the national state of the French-Canadians and exercises its governmental prerogatives, in the areas of its responsibilities, on the majority of the heirs of those who colonized New France.

To this, some authors have replied that the territories covered by New France and those now included in the Province of Quebec cannot be equated. Although New France began with the founding of three cities—Quebec city in 1608, Trois-Rivières in 1616, and Montreal in 1642—it finally included territories that extended west in what is now the United States to the Ohio and Mississippi rivers. Even if the British government, by the Quebec Act of 1774, did in fact include practically all the territories of New France in the new Province of Quebec, this situation lasted only briefly.

Some English-speaking historians assert that the Quebec Act created what is now Quebec as well as the practice of trying to fuse British and French institutions in the new political entity. The new British colony was, thus, to be governed by a governor and a council, using British criminal law and French civil law. Whatever the British government intended, however, when the composition of Quebec's population gradually changed as a result of increasing English-speaking immigration, it became increasingly difficult to carry out a policy that could give satisfaction to both the English- and French-speaking groups. In 1775, the year the American Revolution broke out, Quebec city was besieged by American troops, and Montreal was occupied. When peace was restored in 1783, the Loyalists who had fled from the United States were settled west of Ottawa River, in what became the Province of Ontario. This was the beginning of the basic geographical dichotomy in Canada between French and English. In 1791, Canada was split into Lower Canada (Quebec) and Upper Canada (the future Ontario), roughly following the Ottawa River boundary.

Beginnings of ethnic inequities

Furthermore, although throughout the province the rural population remained overwhelmingly French, Montreal became the domain of the English merchants, who were bitterly anti-French. The metropolis of Canada was to have an English-speakmg majority until the middle of the 19th century, and, even after that, French-Canadians never achieved control of its economic life. In this way, the ethnic-economic stratification was maintained within the Province of Quebec. Discrimination existed between the two linguistic groups not only in economic, political, and religious activities but also in such other fields as education. Gradually, two different educational systems came into being. English-speaking McGill University was opened in 1821, but it was not until 1852 that the French-speaking Quebec Seminary, founded in 1668, became Laval University.

Continuing symptoms of internal malaise. During the first part of the 19th century, the causes for conflict between the two groups increased with the rapid growth of the English-speaking population in Canada. The English merchants of Montreal tried in 1822 to obtain an Act of Union that would have united Lower and Upper Canada and given them an English-speaking majority in the country as a whole. The reaction against this attempt among French-Canadians was strong and prepared the way for the 1837 rebellion. This rebellion, the first major manifestation of political nationalism among French-Canadians, was led by Louis-Joseph Papineau, whose Patriote Party became a centre for radical politics. After the rebellion was put down, the British government sent out the Earl of Durham to investigate; his report, published in 1839, offended French-Canadians by referring to them as a people without a history or culture and by characterizing the situation in Lower Canada as "a war between two races." The report also suggested the setting up of responsible government in Canada as a solution to the tensions between the two groups. In 1841 a new Act of Union joined the provinces of Upper and Lower Canada, and, in 1867, the British North American Act created the confederation of Canada by the federation of the four provinces of Nova Scotia, New Brunswick, Quebec, and Ontario.

From then on, French-Canadian nationalism became a permanent feature of Canadian as well as Quebec politics. Doctrines of papal supremacy over national authority introduced the idea of the religious mission of French-Canadians in North America. Under the leadership of such men as Henri Bourassa and the abbé Lionel Groulx, the province evolved its special vocation as the "political home" of French-Canadians, and the government of that province assumed special responsibility for the defense of French culture. This situation also resulted in the doctrine of provincial autonomy that was used by Prime Minister Maurice Duplessis during his various terms of office between 1936 and 1957.

French-Canadian nationalism also led to the "quiet revolution" of the Liberal government under Jean Lesage, who took office in 1960, and to the not-so-quiet revolution of a terrorist group known as the Front de Libération du Québec (FLQ), which was responsible for sporadic violence and the murder in 1970 of Quebec's labour minister, Pierre Laporte. The creation of the Parti Québecois in 1970 brought into being a new form of Quebec nationalism, one that is no longer strictly French-Canadian: it has English-speaking members as well as members of other ethnic groups, and its advocacy of separation from Canada is based on issues of economic and social development.

Revival of nationalism and separatism

THE NATURAL AND HUMAN LANDSCAPE

Quebec is a sort of "land's end," almost separated from the rest of Canada by Hudson Bay and Hudson Strait. For a number of years, provincial maps issued by the Quebec government have reflected the long-standing dispute by not showing the Labrador boundary, thus including the whole of the North Atlantic Coast down to the Gulf of St. Lawrence as part of the province.

Physical environment. ***Geographical regions.*** Quebec usually is divided into three major geographical areas that reflect its main geological structure: the Canadian Shield (aiso called the Laurentian Shield); the populated lowlands of the St. Lawrence Plain; and the Appalachians. The Canadian Shield covers 95 percent of Quebec, comprising the entire region north of the St. Lawrence Plain. Beginning at the foothills of the Laurentians, the oldest mountain range in the world, it runs northward to the Hudson Strait. The Shield is composed of three main subdivisions: the Laurentian Hills, covered with trees, have become a natural playground for summer and winter activities; the taiga, a region of stunted trees farther north; the tundra, in which a continuously frozen ground, or permafrost, allows no trees to grow and where summer enlivens only caribou moss and a few dwarf birches.

The fertile zone of Quebec and the most densely populated is the lowlands of the St. Lawrence Plain, while the Appalachian region is composed of hills and plateaus, more or less undulating and rising to the higher mountain ranges in the United States. This region includes the Eastern Townships (the area south of the St. Lawrence Plain) and the Gasp6 Peninsula.

Waters. The landscape of Quebec is laced with thousands of lakes and rivers: if the St. Lawrence is included, the freshwater area amounts to some 110,000 square miles. The St. Lawrence, which cuts across southern Quebec from west to east, is one of the world's greatest waterways. Among its tributary rivers are the Ottawa, the Saguenay, the Saint-Maurice, and the Manicouagan.

Climate. Quebec's climate is often extreme, sometimes severe. In Fort-Chimo, on Ungava Bay, the temperature ranges from an average of —11° F (—24" C) in January to 52" F (11° C) in July. In the south in Sherbrooke, it varies from a January average of 15° F (—9" C) to a July average of 68° F (20" C). Temperature changes may be as great as 30° F (17" C) in less than 24 hours. The period in which the snow remains on the ground differs from an average of 12 to 13 weeks on the Montreal plain to 23 weeks on the north coast of the St. Lawrence. The same variations exist for days without frost, Montreal having an average of 140 days without frost, the far north fewer than 80 days.

Patterns of settlement

Human imprints. For all practical purposes, the inhabited part of Quebec is limited to the St. Lawrence Plain and parts of the Appalachians and the Laurentian Hills. Over 80 percent of its population lives within an area 200 miles long and 60 miles wide, one of the highest concentrations in Canada. Almost 80 percent lived in towns in 1970, and fewer than 500,000 were classified as rural farmers; the balance of the population was scattered in forestry, fishing, mining, and other types of activity in small settlements. The historical movement of the population in Quebec has been from large numbers in scattered, diversified settlements to an increasing concentration in urban areas. The shortage of fertile land suitable for cultivation prevented the development of a truly agricultural economy. Contradicting some long-held beliefs, sociologists have advanced the idea that Quebec never developed a peasant society on European models.

The rural settlements that existed during and after the period of French control were limited to the shores of the St Lawrence, forming a continuous line between the urban centres of Montreal and Quebec city. Outside of Quebec city, Trois-Rivières, and Montreal, the land was divided into long, narrow, individual strips. A house was built at the end of each of these strips, on the side of the road that led to the towns, thereby forming a type of ribbon development. Rural Quebec had no village until the end of the 18th century, and most rural paiishes were created during the 19th. As each road became fully settled, a parallel road was opened further inland, a process that was repeated until the whole of the St Lawrence Plain was occupied This system cf colonization. in which houses were located equidistantly along a road, allowed for maximum density in settlement, but it was not suitable for agricultural activities.

The very rapid urbanization of the province in the 20th century did not imply, therefore, the breakdown of a traditional peasant society. Although relatively isolated rural and fishing settlements existed on the north coast of the St. Lawrence and mining and foresting settlements existed in the Laurentian Hills or in northern Quebec, these communities did not create any distinctive “folk” type of society that would have grven a dominant cultural orientation to French Canada. As research has shown, the rural mode of life remained within the general cultural norms of Quebec as a whole.

Until the 1940s, when industries dislodged agriculture from its economic dominance, most French Canadians lived in rural areas, receiving their income from mixed farming, forestry, fishing, poultry raising, livestock, and dairying. In the St. Lawrence Plain, the favourable soils and climatic conditions added a certain diversity of crops —tobacco in Joliette, sugar beets in the Richelieu Valley, and apples in the Montreal plain. The early-springtime collection of sap from the maple trees and the "sugaring-off" process is a pleasant feature of Quebec agriculture.

Importance of the Canadian Shield area

The future of Quebec is, however, now linked with the development of the Canadian Shield, which contain? one of the world's largest reserves of minerals. The first area to be developed in the 20th century was around the Noranda–Rouyn area, which is noted for copper, zinc, lead, gold, and silver. Continued demand for copper resulted in the opening up of new mines in Chibougamau. At present, the main attention is on Labrador, where large deposits of iron are located. Outside the shield. copper is mined at Murdochville in Gaspé, while the Eastern Townships supply nearly two-thirds of the Western world's output of asbestos.

Cheap waterpower, abundant raw materials, relatively iow labour cost,, and dense concentration of population in the St. Lawrence Plain have been the main factors behind the development of industries in Quebec. The process began with the building of textile mills in the Eastern Townships, but industry gradually was concentrated in the Greater Montreal area, in which today some 5,000 plant? account for 60 percent of Quebec'? manufacturing output. Montreal is a leading producer of aircraft, railway locomotives, leather footwear, and chemical and pharmaceutical products; other industries include flour milling, tobacco, biewing, and distillery. Montreal also has oil refineries that supply nearly one-third of the petroleum requirements of Canada.

Because of the present distribution of the population. Quebec can be presented as a dichotomy between Greatei Montreal, with more than 40 percent of the population, and the rest of the province. Although other large cities in the province include Quebec, Chicoutimi, Hull, and Trois-Rivières, this division represent3 the basic structure of Quebec society. In the rural backwaters, most of the population ekes out a meager living by cultivating relatively small farming lots and by cutting wood In the smaller cities and in Greater Montreal are to be found all the characteristics of modern industrial urban concentrations, with their pollution and their social pathology. The dichotomy is most evident when outlying areas of Quebec are compared with Montreal. The Gasp6 Peninsula, the poorest and least urbanized region, has over 240,000 inhabitants, but it is only about one-third urban. Nearly 20 percent of its population earns its living through agriculture. and, in the summer months, tourism becomes the major source of income for the population.

THE PEOPLE OF QUEBEC

The ethnic mélange. Because the distribution of the various ethnic groups is the single most important factoı m Quebec's social, economic, and political life, past and future trends must be taken into account. In the years of French colonization, from 1608 to 1750. only some 10,000 French came over to settle on the shores of the St. Lawrence. The approximately 60,000 persons who were "ceded" to Britain in 1763 were practically all Canadian born, and most of them traced their descent through several generations of habitants, the name given to those born in New France

Impact of changing birth and death rates

The high birth rate among habitants explains the growth of French Canada. Between 1700 and 1760, the average yearly rate of birth was between 54 to 62 per 1.000 inhabitants, while the death rate war relatively low for the time: between 24 and 40 per thousand. After 1763, this extremely rapid natural growth of the population cootinued, and from 1800 to 1850 the total number of French-Canadians increased from about 100,000 to 1,000,000.

The very favourable conditions for population growth ceased gradually to operate during the 19th century, however. for the fertile lands around the St. Lawrence were fully occupied, and increasing difficulties were encountered in maintaining satisfactory economic conditions for family life. French-Canadians began a migration to the United States and other parts of Canada that between 1840 and 1940 is said to have taken 1,000,000 persons from Quebec. Although efforts were made to open new rural areas in more northerly Quebec, the harsher conditions of life limited the new settlements. Trends were, in fact, away from rural settlements, and the new industries attracted an ever larger number into the towns. By the beginning of the 20th century, the birth rate had been reduced to about 38 per 1,000, while the death rate still hovered around 20 per thousand From then on, both figures declined rapidly, especially the death rate. In 1963 the birth rate was about 16 per 1,000, the lowest in Canada, and the death rate less than 7. By 1970 the birth rate had fallen to 15.3, bringing about what some authorities called a “crisis of the birth rate ”

As a result, the overall population distribution in Quebec is changing. Predictions made by demographers from the Université de Montréal suggest that by the year 3000 the oresent 68 percent of Montreal that is French-speaking may decline to 50 percent, while the overall French population in the province may drop to 70 percent. With nearly half of Quebec's population in the Greater Montreal area, what is happening there to population distribution is of considerable political importance The cosmopolitan character of Montreal makes it a “pluricultural” city, with a diversified population of persons of British origin (close to 12 percent in 1961), Italian (7 percent), German (1 percent), and many others (12 percent).

At a time when there are demands for change in the electoral map of Quebec so as to give better representation to the urban population, ethnic pluralism has become a decisive factor of Montreal politics and is gradually acquiring more importance in the Province of Quebec. The long-term trend in population change is making for an ever greater differentiation between Montreal and the rest of Quebec, so that most of the cultural legislation prepared by the Quebec government will become less and less acceptable for Montreal. This fact became evident when, at the end of 1971, a bill to reorganize the educational system of Montreal was withdrawn from the National Assembly of Quebec because of attacks on it from nationalistic French-Canadian groups and from other ethnic groups who, for diverse reasons, felt their linguistic rights threatened.

The Indian population of Quebec has grown to more than 25,000, distributed among 39 different reserves and

Northern coast of the Gasp6 Peninsula.
BY courtesy of Information Canada Phototheque

Indian communities. Some 3,300 Eskimos live in northern Quebec. Eskimo and Indian affairs in Canada are under federal jurisdiction, but, with the transfer of certain responsibilities to the province, a Direction Générale of northern Quebec has been created by the Quebec government to provide services for the Eskimo population. This creation has led to a debate between the federal and provincial governments as to their exact responsibilities in this matter.

Religion. Another distinctive characteristic of Quebec has been its religious homogeneity. During the period of New France, Catholicism was the official religion. After 1760, freedom of religious practice was authorized bv the British government. As the English population grew, so did Protestantism. Following the potato famine in Ireland in the mid-19th century that forced many Irish to migrate to Canada, the number of Catholics in Quebec also increased. Many marriages took place between the Irish and the French-Canadians, and today many persons in Quebec who have French as their mother tongue have Irish names. With the coming of such new immigrant groups as the Italians after World Wars I and II, the proportion of Catholics again increased and stood at 88 percent in 1961. For a long period, religion was a basic mode of differentiation among the various groups; today this is not so. Religion is no longer a major aspect of Quebec's political life.

Increasing proportion of non-French Catholics

Current immigration. During the 1960s, Quebec received about 20 percent of all immigrants entering Canada, some 30,000 each year; the largest group, about 5,000 a year, was from Italy. Immigrants from France were the second most important group, while immigrants from French-speaking countries made up between 25 to 30 percent of the yearly average for those years. The tendency, however, has been for the majority of immigrants to settle in Montreal, to choose English as their working language, and to send their children to English-speaking schools. To solve certain problems in social and linguistic integration, the Quebec government in 1968 created a Department of Immigration.

THE PROVINCIAL ECONOMY

The most distinctive feature of Quebec's economy is clearly shown by its present distribution of labour, which in 1967 was about 61 percent in tertiary, or service, industries; 30 percent in secondary, or manufacturing–processing, industries; and the balance in primary, or agricultural–extractive, industries. Although until recently Quebec was considered to have a mainly agricultural vocation, the occupational concentration that now exists in industrial and service activities shows that Quebec has followed the general trend of industrialization in Canada.

Quebec's economic growth, however, has done little to change the economic position of French-Canadians. Within the major corporations, banks, or branches of multi-national companies that operate in Quebec, few French-Canadians have risen to the top. Promotion has been very slow for them, even in the nationally owned Canadian National Railway or Air Canada, both of which have their headquarters in Montreal. This problem has provided an increasing point of contention in the province's social and political unrest.

Economic resources. Because of continuing difficulties in surveying the land, it is not yet known how extensive Quebec's natural resources are. What is known to exist is quite considerable, especially as far as mining is concerned. Between 1910 and 1970, mineral production increased 100 times in gross dollar value (not accounting for inflation). This rather spectacular development is due to the discovery of extremely large iron-ore deposits in northern Quebec. Geologists estimate that these deposits are among the largest in the world. Two new towns have been created in the north as a result, Schefferville and Gagnon, and a large port, Sept-fles, has been developed. The government of Quebec has created two corporations to survey resources and stimulate exploitation. Furthermore, since only a small part of Quebec's iron is processed in the province, it has created another corporation, the objective of which is to develop an integrated iron-and-steel industry in Quebec. Another major development has been the establishment of the General Investment Corporation. The aim of this corporation is not only to finance economic growth by acting as a broker in the amalgamation of small firms but also to lend money to start new industries.

Minerals, water-power, and forestry

Another of Quebec's major economic resources is electric power. With the nationalization of all electric power in Quebec in 1963, Hydro-Québec became the largest producer of electricity in Canada, its installed capacity representing over one-third of all electricity produced in Canada. Besides its more than 50 hydroelectric plants, Hydro-Québec has 17 thermal plants, and a nuclear plant opened near Trois-Rivières in 1971. With the government of Newfoundland, Hydro-Québec is engaged in developing the hydroelectric resources of Churchill Falls in Labrador. The James Bay project, under discussion in the early 1970s, will increase electric resources further. In February 1967, the Research Institute in Electricity, the first of its kind in the world, was created near Montreal, employing more than 200 research workers.

Forestry is the third great economic resource of Quebec, It is estimated that forests cover close to 400,000 square miles, of which some 221,000 are actually exploited, contributing about 25 percent of the gross industrial production. About two-thirds of the annual forest harvest is used by the more than 50 pulp and paper plants in Quebec, which employ more than 28,000 employees and produce nearly one-half of the Canadian and over one-tenth of the world's pulp and paper products.

Production sectors. The gross domestic product of Quebec is growing, but its rate of increase is less than that for Canada as a whole. A number of unfavourable factors affect Quebec's economy, and such areas as textiles were definitely in regression in the early 1970s, causing a high rate of unemployment. Slowdowns in investments have been reported for the pulp and paper industries, food and beverages, construction, and other related services. The per capita income has also lagged significantly behind that of Ontario, the other highly industrialized province.

Transportation. Quebec is fully integrated in the general transportation system of Canada and of North America. Except in northern Quebec, neither climate nor distance are serious problems. The major handicap in the development of transportation is the low population density of the regions outside the urban area of the St. Lawrence Valley and the restricted flow of goods and persons to the north. With nearly 57,000 miles of modern highways, 73 airports, more than 5,000 miles of railroads, and numerous ports on the St. Lawrence, however, transportation facilities play a paramount role in the growth of industries. Yet, with about 2,000,000 road vehicles, Quebec has a lower per capita ratio of vehicles than the Canadian average.

The transportation system is largely oriented on the basis of the geographic position of Montreal, the major crossroad for moving persons and goods in and out of Quebec, whether by road, water, or air. Nearly 30 international airlines serve Montreal. Montreal International Airport, in Dorval, ranks second only to Toronto's airport in flights and passengers. The inner air traffic for Quebec province is provided by two companies. Montreal is both a major ocean port and, by virtue of the St. Lawrence Seaway, which allows ships to travel a total of 2,300 miles from the North Atlantic to the ports of the Great Lakes, a major inland port as well.

The railway system of Quebec is practically restricted to the St. Lawrence Plain, with a few branch lines of the two major Canadian companies. Three privately owned railways transport iron ore from northern Quebec.

In urban transportation, the subway system in Montreal has 16 miles of track that supplement the Montreal Transportation Commission surface routes. Quebec has 25 other urban transportation services and more than 50 transportation companies carrying passengers to and from various towns.

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of government. *Provincial level.* Quebec's administrative system, created by the British North American Act of 1867, can be defined as government through parliamentary democracy. An elected unicameral National Assembly is the equivalent of the parliamentary institutions of other Canadian provinces; its second chamber was abolished in 1968. A lieutenant governor represents the British monarch. The Executive Council, or Cabinet, of 23 ministers is headed by a prime minister who is responsible to the National Assembly for all legislation within provincial competence. The Executive Council has responsibility for preparing legislation for the National Assembly. Adoption is assured through the parliamentary principle that a prime minister and his Cabinet will remain in power as long as the prime minister is able to command a majority in parliament.

Legislative, executive, and judicial branches

An unusual characteristic of Quebec is its administration of justice. Although its Department of Justice, as in other Canadian provinces, has a dual responsibility for criminal and civil laws, the civil law of Quebec is different from that in the other Canadian provinces. It also has its own provincial police, La Sûreté Provinciale du Québec, who have taken over this responsibility from the federal Royal Canadian Mounted Police.

Politics. To elect the National Assembly for its four-year mandate, Quebec is divided into 108 electoral districts, each returning a representative on the basis of a majority vote. There is a universal franchise for all persons over the age of 18. The law limits electoral expenses, and the state pays a major part of each candidate's costs. An imbalance exists in the representation between rural and urban districts, a problem that was especially important in the 1970 election, when the Parti Québecois claimed that the present system was undemocratic. In that election, more than 80 percent of the 3,500,000 registered voters went to the polls. Since then, a parliamentary commission has been created to propose a new electoral map of Quebec.

Provincial–municipal interactions. The Quebec government long centralized the greater part of its administrative activities in Quebec city, but, to regionalize its responsibilities, 10 regional administrative centres were created in 1966. Certain difficulties have been encountered in developing this new administrative structure, however, because there are more than 1,600 municipalities in Quebec, each with its own form of local administration. Each municipality consists of a council and a mayor. The councils administer revenues, which come mainly from diverse local taxes as well as f om government grants. In defense of their autonomy, local governments have resisted the efforts of the provincial government to integrate them into some form of regional administration, even when the need for such coordination has become most urgent—as in the development of police, transport, and similar services for Greater Montreal. Although regional administration has been in existence for Greater Montreal since 1970, the conflicting claims of the separate municipalities present constant difficulties.

Revenues. Another problem of the provincial administration has been that of finding ways to increase revenues. In Canada, taxes are divided into the four levels of administration: federal, provincial, municipal, and school board. Quebec, unlike the other Canadian provinces, also levies direct taxes on personal and corporate incomes and on estates, which together represent about 50 percent of the total taxes paid in the province. Other sources of income are an 8 percent tax on retail sales, a provincial lottery, and a provincial monopoly on the sale of alcoholic beverages. Municipal corporations and school boards also impose real-estate taxes and some taxes on business and on the retail value of premises. All these sources of revenue have been insufficient, however, to finance the large-scale health, social, and educational investments that have been made by the Quebec government, and one of the main sources of conflict between the federal and the Quebec governments has been the redistribution of federal tax revenues.

Problems of revenues and social services

The social milieu. Social conditions in Quebec differ from those in the rest of Canada in a number of respects.

Services. Although sharing with the rest of Canada one of the world's highest standards of living, Quebec has a larger proportion of unemployed and persons on social security than the Canadian average. The cost of social aid is one of the highest among provincial governments in Canada. This situation is explained by the fact that age, health, level of education and training, and similar factors place a higher percentage of the population below the requirements of a modern industrialized society. The Quebec government began a large-scale re-evaluation of social assistance in the early 1970s and was trying to develop minimum-wage legislation. It also completely reorganized its health service and introduced universal medicare. These steps meant stretching resources to the limit, however, and serious difficulties were being encountered. Nationalist French-Canadians also point out that, because the basic resources in Quebec are in the hands of the English-speaking minority, the inequalities between the two linguistic groups cannot be solved until economic decisions are completely controlled by the Quebec government.

Thus the relation among economic development, social conditions, and educational level remains one of the basic causes of the political tension between French and English, on the one hand, and between the federal and the provincial governments, on the other. Although Quebec has transformed some of its institutions in order to meet the difficult demands of an industrial society, the changes are yet too recent to be able to determine whether they will improve the present situation permanently. The ethnic division of labour between French and English in Quebec, according to all the published research on the subject, shows no improvement, and the various reports published by the Royal Commission on Biculturalism and Bilingualism in Canada indicate that the goal of equality between the two ethnic groups is as distant as it was at the beginning of the 1960s.

Education. Nothing shows more clearly the complex nature of Quebec than its educational system. Organized originally along religious and linguistic lines and largely privately financed, it has become, for all practical purposes, a public system since 1964 under the provincial Department of Education. Larger sums are spent each year on education (in 1970–71 more than 30 percent of the province's expenditures). Although the religious dichotomy between Protestants and Catholics has been maintained, it is largely along linguistic lines that the separation continues. A major problem is the gradual growth in the number of English-speaking pupils, while, because of the low birth rate, the overall number of French-speaking pupils at the elementary level is decreasing. Some tension exists over the fact that, because of the law that permits parents to choose the language of education, increasing numbers of persons are choosing English.

Organized on the normal structure of pre-school, elementary, secondary, and college–university levels, the new system has created Collèges d'Éducation Général et Professionnel, which act as two-year pre-university-level institutions. There are three English-speaking and four French-speaking universities; of the latter, the Université du Québec is organized in three campuses, in Montreal, Trois-Rivitres, and Chicoutimi.

CULTURAL LIFE AND INSTITUTIONS

Ethnicity within the cultural milieu. In many ways, Quebec is a smaller plural society within the larger pluralism of Canada: that is to say, it is nearly as difficult to define the cultural identity of Quebec society as it is to define that of Canada as a whole. Although a minimum of overall cultural identity does exist in French-Canadian life in Quebec, there are also many cultural differences between its working class groups in east Montreal and similar working class groups in such smaller centres as Abitibi, Lac-Saint-Jean, Gaspé, or the Eastern Townships. Regional variations do exist, and they produce a sociocultural fragmentation that in part explains the differences in political votes, religious behaviour, and even the quality in the use of the French language. Furthermore, there are class–cultural differences among French-Canadians, a strong elitist tradition in French Canada explaining the high social status of such professions as medicine, law, and the clergy. Even the new middle class of French Canada, which has appeared since the development of urban industrial society, is more politically aggressive in Montreal than in Quebec city, for the obvious reason that in Montreal their socio-economic status is most difficult. Because of the cultural variations within Quebec, it is often difficult and sometimes impossible to obtain unanimity in political decisions that touch cultural or educational questions.

Ramifications of cultural pluralism

Governmental activities. To ensure that French-Canadians are not placed at too great a cultural disadvantage in their own society, the Quebec government has created a number of institutions that aim to foster cultural life. Foremost among these institutions is the Department of Culture, which is responsible for improving the quality of the language used and for stimulating cultural, literary, and other artistic activities. Created in 1961, it was the first of its kind in North America. It not only gives direct financial aid to such state cultural bodies as museums and helps more than 60 theatrical, ballet, and musical companies but it also has contributed for a number of years to book publishing and to public libraries for book buying. One of the major responsibilities of the department is to develop cultural links with other French-speaking countries, and, in 1965, a cultural entente was signed between Quebec and France that set up cultural- and educational-exchange programs. Another of its activities has been the development of regional cultural centres within Quebec to foster regional cultural life. Another institution created by the Quebec government in its efforts to stimulate French-language cultural development is Radio-Québec, founded in 1968 to develop cultural and educational programs. In 1969 its scope was broadened to take in television as well, and it was renamed Quebec Broadcasting Bureau.

The arts. Besides the official cultural institutions, Quebec possesses an extremely large number of private artistic organizations, ranging from theatre companies to film making. The most spectacular cultural development in Quebec, however, has been that of the ***chansonniers,*** who represent a cross between poets and songwriters. Their popularity, especially among the younger generation, arises from the fact that their songs reflect the present search for cultural and political identity. The *chansonniers* are involved frequently in political activites and are identified largely with the nationalist movement. Music and painting also share in this artistic revival, as does literature. In the area of book publishing, nearly 1,000 books a year are published in Quebec, in both English and French. This activity, together with 19 television stations, 73 radio stations, and 14 daily newspapers, gives an intense and varied cultural life to the province.

PROBLEMS AND PROSPECTS

When in 1967 Gen. Charles de Gaulle, the visiting president of France, made his controversial declaration "Vive le Quebec libre" ("Long live free Quebec") in Montreal, he marked the beginning of a new wave of rising separatist feeling in Quebec. Although an ever larger number of French-Canadians were attracted to extreme nationalism, this renewed wave did not create the required unanimity among Quebecers that was expected by separatists. On the contrary, not only did it sharpen the political differences between English and French and provoke most of the new immigrant groups in Quebec society to make a stand on the question of separatism, it also split French-Canadians into two opposing camps. The new urban, middle class French-Canadians, who form the majority of the population, became divided over separatism, and even the labour unions did not achieve unanimity on this question.

As the chaotic events of the early 1970s have shown, there is no present converging national interest between lower class and middle class French-Canadians. It would seem, therefore, that unless extensive social and economic deprivation among all classes develops in the future to create such a political unity, the present fragmentation among French-Canadians will continue. Although it would be premature to speak about a decline in the present wave of nationalism in Quebec, there is no doubt that a decline in separatism as a political choice has occurred. In many ways, the present social pluralism of Quebec society, with its division between Greater Montreal and the rest of the province, and the new class stratification that is to be found in the cities preclude any extreme solution acceptable to all groups. These facts do not mean absence of conflicts or even of violence, however, and in all probability the present difficulties will continue.

BIBLIOGRAPHY. An excellent short bibliography on historical and literary aspects of Quebec may be found in ***The Oxford Companion to Canadian History and Literature,*** pp. 679–687 (1967); while PHILIPPE GARIGUE, ***A Bibliographical Introduction to the Study of French Canada*** (1956) and ***Bibliographie dzt Québec, 1955–1965*** (1967), cover all areas. The best source of statistical and other basic information is the ***Qztebec Statistical Year-Book*** (issued annually since 1914). On geographical questions, see the chapter on Quebec in LUDGER BEAUREGARD, ***Le Canada*** (1970). No recent single study of Quebec presents the overall situation regarding Quebec's

unique status within the Canadian confederation. MASON WADE, *The French-Canadians, 1760–1945*, rev. ed. (1968), remains the best simple introduction in English. MARCEL RIOUX and YVES MARTIN, *French-Canadian Society* (1964); and *Facets of French Canada*, prepared by the Association Canadienne des Éducateurs de Langue Française (1967), are both collections of essays on social issues. On the recent political events, large documentation has yet to be analyzed properly. Among them the following may be consulted: RAMSAY COOK (comp.), *French Canadian Nationalism* (1969); PIERRE ELLIOTT TRUDEAU, *Le Fédéralisme et la Société Canadienne-Française* (1967; Eng. trans., *Federalism and the French Canadians*, 1968); RENE LEVESQUE, *Option Quebec* (1968); and GERARD PELLETIER, *The October Crisis* (1971).

(P.Ga.)

Quebec (City)

The oldest city of its nation and one of the most picturesque on the North American continent, Quebec (French Québec) is the capital of the province of Quebec and, to its residents, the spiritual capital of French Canada. The thick stone walls of its Citadel, topped by the guns of an earlier day, lie astride Cape Diamond, which rises precipitously for 333 feet (101 metres) out of the north bank of the St. Lawrence River. This "Gibraltar of America" was built to protect the water entrance to New France, at the point where the river suddenly widens on its eastward course to the Gulf of St. Lawrence and the Atlantic Ocean. The name Quebec is derived from an Algonkin Indian word meaning "narrowing of the river," a phenomenon noticed by the explorer Jacques Cartier when he sailed in from the Atlantic in 1535.

European atmosphere

Quebec is often referred to as the most European of cities in the Americas. Among the elements contributing to its distinctive character are its old wards cut through by narrow, winding streets, the architecture of its houses, an overall atmosphere redolent of a bygone France, its relics from the past, and its slow rhythm of life—all existing alongside the institutions of a modern metropolis. The Quebecois, almost 95 percent of them of French ancestry, are a people proud of their past and of the authentic cultural life that has been retained. Vast amounts of money are spent annually on the preservation and restoration of the old city, while, elsewhere within the city limits, demolition is constant for the widening of streets and the construction of other aspects of modern commercial and domestic life.

Present-day Quebec is a major inland seaport open to ocean navigation throughout the year. It is also the centre of provincial administration, the commercial hub of eastern Quebec province, and an industrial city with rail, highway, air, water, and electronic connections to the nation and the world. Its landscape and life tumble across the hills from Cape Diamond to the tablelands around the confluence of the Saint-Charles and St. Lawrence rivers, from the governmental and cultural centres of the upper city to the commercial centres and slums in the lower city. Amid its geographical complexity and in its sometimes jarring juxtapositions of centuries, Quebec is both a contemporary city growing vertically and a living museum of man's aspirations to re-create the forms of his civilization in a wilderness. (For information on related topics, see QUEBEC; CANADA; CANADA, HISTORY OF.)

History. The origins of Quebec coincide with the discovery of Canada. Jacques Cartier, the first European to sail up the St. Lawrence River in search of the Northwest Passage to the Orient, landed at the Stadacona settlement of the Huron Indians on September 7, 1535. Chief Donnacona welcomed him favourably. After a hard winter encampment at Cap-Rouge, on the St. Lawrence, during which illness killed many of his men, Cartier returned to France in the spring with Chief Donnacona and five Hurons. Donnacona never returned.

In 1608 Samuel de Champlain revived Cartier's colonial plan, installing the first permanent base in Canada at Quebec. At the foot of Cape Diamond, he built a combination dwelling place, store, and fortress. From here he undertook several expeditions to the interior and made alliances with the Indians he found in the area.

French foundations

The first economic activity of the settlement was the fur trade, but Champlain understood that this alone would not assure the survival of his young settlement. The land had to be tilled, and the Indians had to be "civilized." In 1615 the first missionaries, the Récollets, reached Quebec, followed by the Jesuits in 1625. The first settler, a French apothecary named Louis Hébert, landed in 1617 and built a small house atop Cape Diamond.

In 1628 Quebec was exposed to a long blockade by a British expedition and surrendered in 1629. Though 36 settlers had confidence in the future and remained, Champlain returned to France until 1632, when the Treaty of Saint-Germain-en-Laye returned Quebec to France. The colony was then able to develop rapidly and on more solid bases. Quebec, a colony of men, greeted the return of its Jesuit missionaries and welcomed its first nuns. In 1639 Madame de la Peltrie and two other Ursulines opened the first school, and three Hospitalières established the first hospital. With the arrival of Monsignor François-Xavier de Montmorency de Laval in 1659, Quebec became the first episcopal seat in North America. Laval founded the first professional school and, in 1663, the Quebec Seminary.

During the following years, the first efforts at industrialization emerged. Jean Talon opened the first brewery and set up the leather and hosiery industries. He also encouraged the wood trade and stimulated shipbuilding. New settlers and a regiment of soldiers sent to fight the threatening Iroquois swelled the population of Quebec to about 500; yet it was smaller than that of Montreal, its sister city, 150 miles up the St. Lawrence.

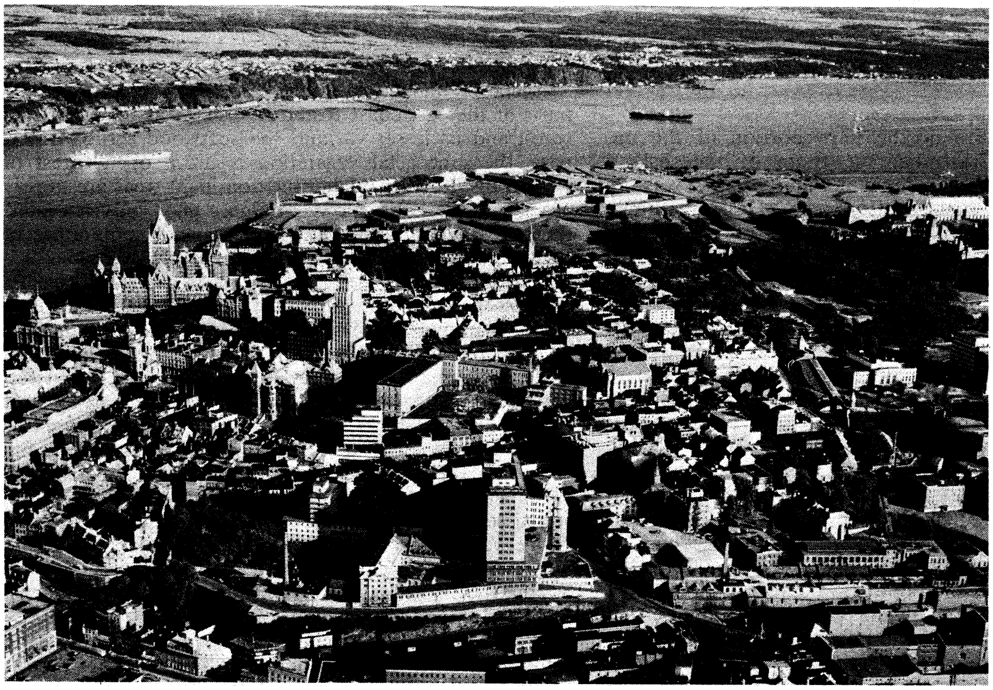
At the same time, the English colonies of America were growing rapidly, and America was beginning to experience the old European rivalries between France and England. A confrontation was inevitable. In October 1690 the fleet of Sir William Phipps attempted to take Quebec but was beaten back by its governor, the comte de Frontenac. Having lost 400 men, Phipps sailed back to Boston. In 1711 a second attempt to take the city also failed when a British armada crashed on the reefs of the St. Lawrence before reaching Quebec.

The traditional enemy came back, however, during the summer of 1759. The expedition included 76 warships, 150 transports, 13,500 sailors, and 9,000 soldiers. It submitted the city to a violent bombardment, but the colony held out. Unable to take it by front, the English troops moved around it; and, on the morning of September 13, they deployed on the heights of the hill to the west of the fortified enclosure. The marquis de Montcalm accepted the battle. Fifteen minutes later the French troops were in full flight. James Wolfe, the English commander, was killed in action, and Montcalm died of his wounds. The city fell on September 18 and finally was surrendered to England by the Treaty of Paris in 1763. A new life under an English governor began for the nearly 8,000 inhabitants of Quebec. In 1775–76, during the American Revolution, the Americans, under Richard Montgomery and Benedict Arnold, failed in an attempt to take the city.

British capture and acquisition

In 1791, after the creation of the provinces of Upper and Lower Canada, Quebec was designated as the capital of Lower Canada, now the province of Quebec. It was incorporated in 1832 and was given its actual charter in 1840. In 1864 Quebec was the seat of the conference of British North American colonies to plan the confederation of Canada. In 1917, during World War I, the city suffered from violent riots in protest over the decision of the Canadian government to enforce military conscription. During World War II, Pres. Franklin D. Roosevelt of the United States and British prime minister Winston Churchill twice met in the Citadel to plan the invasion of Normandy.

The contemporary city. The upper city is built on top of Cape Diamond, the lower city beneath it, most of it lying at the bottom of the cliff and on the foothills of the Laurentides, in the valley of the Saint-Charles. The lower city, by far the largest and most populous, includes the principal commercial centre and densely populated wards with their poverty zones and industrial sections. The upper city is the heart of the artistic and cultural life, of



Quebec city, overlooking the St. Lawrence River. In the background are Chateau Frontenac (left), the Citadel on Cape Diamond (centre), and the Plains of Abraham (right).

By courtesy of the Canadian Government Travel Bureau, Ottawa

provincial and municipal administration, and of the more affluent residential quarters.

Points of interest. Quebec presents many points of attraction for the many thousands of tourists who visit it annually. The best known is the Dufferin Terrace, crowned by the impressive Château Frontenac, a hotel behind which rises the Citadel. From the terrace are visible the Heights of Lévis on the south shore, Île d'Orléans, which divides the St. Lawrence into two branches, and, beyond, the coast of Beaupré, backed by the peaks of the Laurentides. The largest park is the Plains of Abraham, the site of the great battle of 1759. It is located on Cape Diamond, alongside the river, and extends from the west bounds of the city to the bottom of the Citadel walls. A promenade hung on the slope of the rock joins the park with the terrace. Many other parks with various monuments are scattered throughout the city.

The principal relics of the past are religious, many dating from the 17th century. On the Place Royale stands the modest church of Notre-Dame des Victoires (1688). Other old buildings include the Ursuline monastery, the seminary, the Anglican cathedral (the first of that confession in Canada), and the Catholic basilica, where many of the bishops of Quebec are buried.

The Quebecois. Quebec is the nerve centre of an urban agglomeration of almost 500,000 persons, most of them Canadian born and of French ancestry. Slightly more than 5 percent are of other descent, the largest group being Irish. The vast majority are Roman Catholics and French speaking. Most of the English-speaking population is bilingual and descended from old Quebec families and, as a consequence, blends readily into the city's general life.

The port of Quebec

Economic life. The service industries provide jobs for about 80 percent of the working population. Another large group includes provincial, municipal, and federal employees, including those in schools and hospitals. Some 500 manufacturing plants in the Greater Quebec area employ about 25,000 people. This aspect of industrial life is strongly influenced by the port, where each year more than 7,000,000 tons of goods are handled. In order of importance, the major fabrication industries are newsprint, grain milling, cigarettes, garments, and shipbuilding. Factories are situated largely in the vicinity of the port and along the Saint-Charles. The port is 35,000 feet long, and the harbour has a depth varying between 35 and 50 feet at high tide, allowing the accommodation of vessels of up to 50,000 tons. About 1,000,000 visitors a year yield an annual tourism income of $45,000,000.

Government. The administration of Quebec lies in the hands of a municipal council consisting of a mayor and 15 councillors from various districts, each elected for four-year terms. Every citizen 18 years of age and older is entitled to vote. There is also an executive committee presided over by the mayor and responsible to the municipal council. Since 1970 Quebec has been the seat of the Quebec urban community, a body coordinating services for the 23 municipalities within the metropolitan area. It is responsible for all the land valuation, the promotion of industry and tourism, and regional management and long-range planning. The governing body consists of the mayors of the 23 municipalities and an executive council.

Education and community activity. A system of dual school boards——ne for Catholics, one for Protestants—is maintained, with the Catholic schools outnumbering the Protestant schools by 68 to 3. Instruction is carried on in French and English, respectively.

Quebec's cultural life is concentrated on Université Laval and its affiliated teaching institutions. Prior to the early 1960s, higher education was the privilege of a small elite, but since then numerous reforms and financial grants have opened it to students from all social classes. Night courses reach a large adult population, while summer sessions in French attract persons from around the Western Hemisphere.

The concert hall and small theatre of the Grand Théâtre, opened in 1971, became a focus of Quebec's artistic life. The modern building is decorated with a mural by the sculptor Jordi Bonet, a work that caused considerable controversy. Many museums and libraries are available among the public and private institutions. Sports are exceedingly popular, especially hockey, baseball, racing, golf, and skiing in the many centres in the Laurentides only a few miles from the city. The Mont Sainte-Anne centre has been the scene of world-cup skiing tournaments. Among the principal local events are the three-week-long winter carnival ending on the night of Mardi Gras and the Provincial Exhibition of late August.

(Ra.D.)

BIBLIOGRAPHY. MAZO DE LA ROCHE, ***Quebec, Historic Seaport*** (1944), is an account from a historical standpoint. See also WILLA CATHER, ***Shadows*** on ***the Rock*** (1931, reprinted 1971); the more economically slanted ***Natural Resources of Quebec,*** issued by the Canadian Department of Natural Resources (1923); and A.G. DOUGHTY and WILLIAM WOOD, ***The***

King's **Book of Quebec,** *2* vol. (1911), a memorial volume; GILBERT PARKER and C.G. BRYAN, ***Old Quebec*** (1903), is still useful for the historical background.

Queensland

Comprising the entire northeastern portion of the nation-continent of Australia, the State of Queensland reaches well north of the Tropic of Capricorn and was the first successful settlement of European peoples within a tropical climate. The 1,000 miles (1,600 kilometres) of its eastern coastal region are separated by the northern reaches of Australia's Great Dividing Range from the vast inland plains. Extending for 1,250 miles off this coast is the Great Barrier Reef, one of the most remarkable coral formations in the world.

Queensland's 667,000 square miles (1,730,000 square kilometres) make up nearly one-quarter of all Australia and nearly one-third of the occupied part of the continent. It is bounded on the north and east by the Pacific Ocean, on the south by New South Wales, on the southwest by South Australia, and on the west by the Northern Territory. Although Brisbane, the capital, contains slightly less than one-half of the 1,800,000 residents of the state, smaller but sizable cities, especially along the coast, have dispersed the population across a larger area than in any other state of the commonwealth. Settled originally on the basis of the grazing potentiality of its great grasslands, Queensland in the 1970s was becoming increasingly diversified in its economy, a significant part of which lay in the allure of its tropical resorts. (For information on related topics, see the articles AUSTRALIA, COMMONWEALTH OF; AUSTRALIA, HISTORY OF; AUSTRALIA; AUSTRALIAN ABORIGINAL CULTURES; and GREAT BARRIER REEF.)

THE NATURAL AND HUMAN LANDSCAPE

Physical environment. ***Surface features.*** The inland two-thirds of Queensland comprises a vast plain, broken in a few places by low ranges and hills, whereas the eastern (coastal) third is hilly, occasionally mountainous, with wide valleys and plains between the ranges. The Great Dividing Range runs the entire length of the state, separating east-flowing from west-flowing rivers. In South Queensland it lies close to the coast, and peaks reach to 4,500 feet; but it becomes a low range 250 miles inland at the Tropic of Capricorn, and again nears the coast in North Queensland. Between the Dividing Range and the coast are many other ranges, more conspicuous than the Divide in Central and North Queensland, where the state's highest peak is Bartle Frere, at 5,287 feet (1,611 metres).

Mountains, rivers, and soils

The rivers of Inland Queensland flow only after heavy rains, when they flood widely across the flat plains. In the arid southwest, the rivers branch into thousands of distributaries, and big floods are followed by lush growth of clover and other herbage in what is known as Channel Country. Most coastal rivers are also intermittent in flow and shallow, navigable only in the tidal reaches.

Most coastal soils have low fertility, though occasional alluvial and other richer soils occur. Large areas of rich black, gray, and gray-brown soils are found inland, forming the basis for one of the greatest natural grasslands of the world. The Great Artesian Basin underlying the plains provides water for livestock.

Climate. The rainfall of Queensland ranges from 180 inches (4,500 millimetres) a year on the wettest part of the northeastern coast to five inches (125 millimetres) in the southwest. All Queensland receives more summer rain than winter rain, but the latter is important in the southeast, in which large areas of wheat and other winter crops are grown. Rainfall decreases in quantity as one moves inland, where droughts are frequent.

Queensland's summer temperatures are warm to hot, humid on the coast but usually dry inland. Winters are mild and sunny, with frosts occasional on the southern coast and frequent inland. Average maximum summer temperatures, registered in January for the southern coast, the northern coast, and the central inland area, are about 85" F (29" C), 88° F (31" C), and 100" F (38° C), respectively; comparable winter minimums in July are 49" F (9° C), 63" F (17" C), and 45" F (7" C). Humidities are greatest in the north, lowest inland.

Vegetation and animal life. The wetter or more fertile parts of the east coast foster several areas of dense rain forest and its lesser variant, the softwood scrub. Otherwise, the subcoastal vegetation is more open forest of eucalyptus. Further inland, the dominant trees are mostly acacia species—brigalow, mulga, and gidyea—interspersed with large grassy plains and with some areas of desert spinifex, a spiny grass growing in clumps.

Tropical rain forests and exotic animals

Queensland has a great variety of animals—50 species of marsupials, 400 of birds, and 1,600 of fish. Two egg-laying mammals, or monotremes, are the spiny anteater and the platypus. Marsupials range from the large red and gray kangaroos to koala bears, opossums, cuscusses, and marsupial mice. Birds include the flightless emu of the plains, great flocks of coloured parrots, and songbirds of the forests. Fish include the freshwater Queensland lungfish, eating and gamefish, and beautiful reef fish.

Human imprints. ***Traditional regions.*** The strong regional sentiment within Queensland arises from the size of the populated areas, some 1,000 miles from south to north, 800 miles east to west. The main regions are South Queensland, Central Queensland, North Queensland, and Inland Queensland.

South Queensland extends north to Bundaberg and some 200 miles inland, including the Darling Downs farming subregion. Some 1,300,000 of Queensland's 1,800,000 inhabitants live in South Queensland. Rockhampton, Central Queensland's largest city, is the outlet and commercial centre for the beef cattle and wool produced in its hinterland. In recent years this region's vast coalfields have been developed, and a very large alumina-refinery has been built at Gladstone.

North Queensland is identified with sugarcane, producing about three-quarters of Australia's crop. Subregions are Mackay and the Atherton (Atherton Tableland). Townsville serves as the commercial centre for the pastoral hinterland and has a copper refinery nearby. Coalfields lie inland from Mackay and Bowen. Inland Queensland is the great pastoral area of the state. Popularly called the West, it has an identity overriding its 900-mile spread and is sharply different from the coast.

Patterns of settlement. Almost all of Queensland has been occupied by pastoralists for 80 to 100 years, but large areas of the state are little changed from the original landscape. In western Queensland, the flocks of sheep and herds of cattle mainly graze the natural grasslands, the interspersed belts of trees providing only light grazing. In southern Inland Queensland the mulga is a useful drought reserve, since the leaves are edible by livestock. It is cut or rolled down for this purpose, and it regenerates.

Nearer the coast, some large areas of open eucalyptus forest have been "ringbarked," a process that kills the trees and allows grass to grow. A large belt of brigalow timber, an acacia looking much like an eucalyptus, in the 20-to-30-inch-rainfall belt of Central and South Queensland has been partially cleared for farming and grazing in recent years, since the presence of brigalow indicates good soils.

In the crop-growing and dairying region of the southeast and on the sugar-growing north coast, timber has been completely cleared from the better soils, mainly river alluviums and basalt areas. Elsewhere, much original timber remains, and there are many national parks and hardwood forests.

Farm holdings and township patterns

The outstanding feature of land settlement in Queensland is the large size of the grazing holdings and farms. Most of the western grazing country is held in leasehold from the crown, and some 2,000 pastoral holdings average over 100,000 acres in area. On the better grassed country the holdings are 15,000 to 25,000 acres, carrying 5,000 to 10,000 sheep. In the grain-farming areas much of the land is freehold, and most farms are between 600 and 3,000 acres in size. Large holdings, with a high degree of mechanization and a small labour force, suit the rather unreliable rainfall pattern of Inland Queensland.

In the coastal sugar areas, farms range from 50 to 100 acres and are highly mechanized.

The large holdings in rural Queensland have produced a sparse pattern of villages, or townships. In the far west and in the rougher rangy areas of the subcoastal belt, the grazing properties, or stations, are very large, employing six to 20 people and carrying more supplies than a village store. Hence, townships often lie as far apart as 30 to 100 miles.

The last decade has seen a decrease of population in most of western Queensland. In the agricultural areas containing mainly single-family farms, the small towns are closer, about 10 miles apart; but with the improvement of roads and vehicles some small towns are disappearing, and towns of more than 1,000 people and lying 40 to 60 miles apart are serving the rural areas.

Queensland has a more decentralized population than any other state. Apart from Brisbane, with more than 750,000 persons, there are five cities with more than 40,000 and five others with more than 20,000. The development of ports and commercial centres along the Queensland coast and the growth of the sugarcane industry are in large part responsible for this widespread settlement.

THE PEOPLE OF QUEENSLAND

Ethnic groups. Nearly 90 percent of Queensland's population is Australian born, and, as in other states, it is predominantly of English, Scottish, or Irish ancestry. Important minorities came from Germany in the 19th century and from Italy after 1920, the latter frequently becoming cane farmers in North Queensland. After 1947, a wider range of European immigrants arrived from The Netherlands, West Germany, and southern and eastern Europe, as well as some from the United States.

The assimilation of Europeans has been rapid once some fluency in spoken English is achieved, and religion has not proven to be a barrier. The assimilation of the Australian Aboriginal, on the other hand, has been slower. Persons having more than 50 percent Aboriginal blood total about 30,000, including some Torres Strait Islanders, who are Papuan in ancestry. Efforts are being made to give the Aboriginal children the same range of job training as other Queenslanders.

Contemporary ***demography.*** Queensland's present-day birth and death rates are close to the Australian average. The annual birth rate is 21 per 1,000 persons, the death rate nine per 1,000. Throughout its history, Queensland's birth rate has been significantly higher than the Australian average, reflecting a more rural population. The margin has narrowed in recent years with increasing urbanization.

Scourge of tropical diseases in the 19th century

From 1860 to 1890 the death rate was substantially higher than the Australian average. During this period, Queensland's future was threatened by tropical diseases brought in by troops from India, by Chinese miners, and by South Pacific islanders. Almost every tropical disease was introduced, and Queensland seemed likely to suffer the same fate as previous European attempts to settle in the tropics. Intense and continued effort by the medical profession and the Queensland health services, however, progressively wiped out the diseases. Australia's policy of restricted immigration was strongly influenced by Queensland's experience of unrestricted immigration, and after about 1890 the state was developed almost entirely by European labour.

The number of rural workers has declined since 1965, with increasing mechanization of farming and the abandonment of some small farms and dairy farms. These trends have far more than offset the increased population in those few rural areas that still are expanding, notably the brigalow lands.

Although Queensland is the most decentralized of Australia's states, the trend toward increased urbanization is strong. Brisbane's population as a percentage of Queensland's increased from about 39 in 1961 to 45 in 1971. The development of Queensland's great mineral resources, contributing to the growth of mining towns and ports, is both a decentralizing and urbanizing force.

THE STATE'S ECONOMY

Queensland's economic pattern is similar to that of the other states, but its manufacturing is less developed than that of New South Wales and Victoria in such areas as steel, chemicals, and transport equipment. It is more developed in the production of certain minerals, in tropical crops, and in extensive cattle raising.

Components. ***Employment patterns.*** Some 580,000 Queenslanders were employed at the end of 1971, excluding employees in agriculture and private domestic service. The largest groups were in manufacturing, commerce and finance, and professional and personal services. Seasonal unemployment is more noticeable in Queensland than in other states, mainly because of the seasonal nature of sugar harvesting and cattle slaughtering and the decline in construction work in the wet months from December to April. Unemployment ranges from about 1 percent of the work force in the busiest months to 3 percent in the slackest.

Agriculture. The original attraction of Queensland is the great natural grasslands of its interior. The grasslands remain the state's largest single resource, even with the growth of manufacturing, mineral exploitation, and agriculture.

Domination of agriculture by livestock

Wool production is almost all of the fine-textured merino variety. Beef cattle are tending to replace sheep in some areas and already have replaced many dairy cattle. Some of the best croplands in Queensland are fully developed: for sugar production on the coast and for grains on the subcoastal Darling Downs. Other crops include peanuts, tobacco, cotton, and many tropical and temperate-climate fruits and vegetables. Queensland has a greater area of land suitable for further development than any other state, especially brigalow and similar lands suitable for beef grazing and summer grain crops.

Manufacturing. Queensland's manufacturing and processing industries now exceed the output of the rural industries in their contribution to the state's income. Brisbane, which employs more than 60 percent of the factory workers, has by far the greatest range of manufacturers. Elsewhere in the state, most manufacture comprises the processing of such primary products as sugarcane, meat, timber, and alumina and various base metals, as well as in industry supplying local needs—printing, baking, vehicle repair, and the like.

Mining. The most spectacular development in Queensland in the late 1960s was in minerals, including alumina and coal. The value of overseas exports of these rose fourfold between 1966–67 and 1970–71. The great bauxite field at Weipa on Cape York Peninsula exports bauxite overseas and supplies the big alumina refinery at Gladstone. The Mt. Isa mines in the northwest greatly expanded their production of copper and of silver, lead, and zinc. Several large coal mines for the export market have been opened up in the hinterland of the central coast. Natural gas from the Roma field supplies Brisbane, and there is a small oilfield at Moonie.

Tourism. A major Queensland industry, the tourist facilities of the state are serving a rapidly increasing number of Australian and overseas visitors attracted by the mild, sunny winters and moderate summers. Prime attractions include the surfing beaches and the Great Barrier Reef islands, as well as the rain-forested national parks and tours of the eastern coast. The surfing beaches are in South Queensland. Gold Coast, 20 miles of beaches near the southern border, is built on the tourist and holiday industry. It has some 70,000 permanent residents and accommodations for 120,000 annual visitors.

Federal and local services

Role of government. The state government in Queensland regulates those parts of the economy not under federal control and provides directly some services, especially railways, main roads, education, health services, police and law, some forestry and irrigation, and some housing. Operating under state legislation, local authorities provide minor roads, some bus services, water supply, drainage and sanitary services, and town planning. Statutory bodies provide electricity, harbours, slaughterhouses, and marketing of some primary produce.

State industrial awards (wages and hours fixed by the state-appointed arbitration authority) cover some 65 percent of the employees in Queensland, a larger percentage than other states. Notable among employee unions is the large Australian Workers Union, covering shearing, grazing, canefarm, and mill employees and construction workers.

Transportation. Queensland developed as a series of ports, with railways tapping the inland, but now coastal shipping is used only for heavy cargoes such as steel and sugar. Brisbane is the largest receiving port, while Gladstone (alumina and coal) and Weipa (bauxite) are the largest shipping ports. Some 6,000 miles of railway operate throughout the state. Closure of some branch lines in agricultural areas has occurred; but two new railways have been built for coal exports, and other lines have been improved. Air services are important in Queensland, especially in getting about the large area of the inland.

A steadily improving road network links all parts of the state, but maintenance of unpaved roads is costly and difficult with heavy summer rainfall. A paved road reaches Cairns and is partly constructed to the main inland towns. Paved beef roads have been built to get fat cattle to railheads from the southwestern Channel Country and the northwest.

Brisbane is the only city with serious traffic congestion, and a 2.5-year program of freeway construction is under way. Most provincial cities also have planned similar but less extensive systems.

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of government. The state government follows the usual British pattern of a legislature, an executive, and a judiciary. Queensland is the only state with a unicameral parliament, the upper house having voted itself out of existence in 1922. Local authorities are established under the Local Government Acts; they are elected through adult franchise, and voting is compulsory. Unlike the situation in other state capitals, most of the metropolitan area of Brisbane is controlled by a single city council. Public justice is vested in the Supreme Court and district courts and, for civil jurisdiction, in lower courts as well. The Supreme Court holds periodical sittings in centres throughout the state.

The Australian Labour Party controlled the state for two long periods, 1915–29 and 1932–57. Since 1957 a coalition of the Country and Liberal parties has governed, and Queensland is the only state in which the Country Party is the major partner in the coalition.

The social milieu. Social and economic divisions are even less marked in Queensland than in the rest of Australia, which itself makes up the world's largest area of uniform language, social customs, and standards of living.

Education. About 80 percent of primary school children attend state-government schools, and the rest are in schools run by religious and other bodies. About 75 percent of secondary school children also attend state high schools. Education is compulsory between ages six and 15.

Private schools receive governmental subsidies, as well as grants for libraries and science laboratories. Correspondence schools always have been important to the remote parts of the state, and more than 10,000 children, including apprentices, are educated in this way. Regular school assessments of pupils are made, and university-admission examinations are held.

Universities are located in Brisbane and Townsville and institutes of technology in Brisbane, Toowoomba, and Rockhampton. An agricultural college, a conservatory of music, and some 20 technical colleges complete the state's offerings in higher education.

Health and housing. Queensland is the only state with a public hospital system that provides free wards as well as the usual subsidised private wards that charge fees. Some 150 public hospitals treat 230,000 patients and 40,000 maternity cases each year. Among the medical-research institutes is the Radium Institute, which successfully deals with skin cancers, a high-incidence affliction in Queensland. The Flying Doctor Service, which commenced in Queensland in 1928, now makes over 400 flights a year to remote farms or settlements, especially in the inland region.

Prevalence of single-family homes

Some 17,000 dwelling units a year are built in Queensland, of which 14,000 are separate houses. With a warm climate and ample hardwoods, relatively cheap timber houses have been the most common type in Queensland. Older houses often were erected on timber piles to give room below the house for laundry, automobiles, etc. and to keep houses airy and free from termites. The trend in recent years is to build low-set houses, as in other states, and brick or brick veneer is now very common.

CULTURAL LIFE AND INSTITUTIONS

A feature of recent cultural development in Queensland is the greater opportunity now enjoyed by people in the nonmetropolitan areas. A network of radio stations, both governmental and commercial, extends over the whole state, except in the most remote areas, which must depend on shortwave transmissions of national programs. Television also covers most of the state. Several companies sponsored by the Arts Council make continuous tours of music, ballet, and theatre into the provincial cities and inland country centres and schools, covering a far greater rural area than is done in any other state. Provincial newspapers are strong in Queensland, and the larger daily papers carry international news.

Local participation in press, radio, television, theatre, and vocal, orchestral, and band music is growing in provincial and rural Queensland, as is the time spent on the arts in school courses. At a higher level, the new James Cook University at Townsville offers arts courses at both undergraduate and graduate levels.

Thus, the earlier isolation and lack of cultural opportunities of rural dwellers in Queensland is, to an extent, being overcome. The state government has established a Cultural Activities section in the Department of Education to advise and coordinate cultural work and to help remote areas. In the early 1970s more than 125 bodies were receiving government grants, and a number of cultural centres were set up and vacation courses held.

Brisbane, with a population 10 times that of any provincial city, is naturally the strongest cultural centre. It has a history of good performance in orchestral music, and the Queensland Symphony Orchestra, though small, is accorded high rank. There is an excellent junior orchestra as well. Vocal music also has been strong, especially in the city of Ipswich, where many Welsh coal miners settled. In theatre, Brisbane is notable for its number of little theatres, run for many years by competent amateurs and now receiving financial aid; some groups have become professional companies. Several new theatres have been built recently, and playwriting has shown strength in Queensland. Brisbane also has many privately owned art galleries, in addition to the state art gallery and museum, and interest in the visual arts has been increased by leading Australian painters working in Queensland at various times. On the more popular level, annual competitions of fence painting are held in Brisbane and elsewhere, which attract many entries, and there are competitions for works of sculpture. The state Public Library, with its associated Oxley Memorial Library of specifically Queensland items, is in Brisbane. Libraries are maintained by 86 of Queensland's 131 local governments, some having several branches.

Popular culture shows little that is distinctively Queensland in origin. One widely enjoyed form of popular culture, however, is the appreciation of the many miles of unspoiled landscapes and shorelines within the state. This is reflected back in the popularity of separate houses that offer outdoor living amid trees and gardens.

PROSPECTS

Queensland, long regarded by Australians as "the state of the future," achieved full and early development only on the great inland grasslands. There and nearer the coast, beef cattle, fodder crops, and grains are now the major expanding industry. The present rapid mineral expansion is based on enormous deposits of high-grade coal and bauxite and on the refining of alumina. The tourist industry is strong and growing, especially on the surfing beaches and islands near the Barrier Reef. The early handicap of great distances has become the asset of great stretches of unspoiled landscape with a well-decentralized population. Manufacturing is expanding, particularly in Brisbane, the population of which faces most of the current urban problems, although not acutely and with a degree of subtropical relaxation.

BIBLIOGRAPHY. GOVERNMENT STATISTICIAN, *Queensland Yearbook* (annual) and current bulletins; PREMIER'S DEPARTMENT, *The Queensland Scene* (1971), a pictorial account, with brief text; DEPARTMENT OF INDUSTRIAL DEVELOPMENT, *Investment Queensland, Australia* (1969); R.H. GREENWOOD, *Regions of Queensland* (1971); R.W. CILENTO (ed.), *Triumph in the Tropics* (1959), a historical study; STANLEY and KAY BREEDEN, *Tropical Queensland* (1970), on the flora, fauna, and landscape; GEORGE FARWELL, *The Sun Country* (1970).

(H.W.H.)

Quintilian

The author of a treatise in Latin called *Institutio oratoria* ("The Training of an Orator"), Marcus Fabius Quintilianus was an experienced and successful teacher whose work embodies the best educational thought of late antiquity. The core of the *Institutio* is a treatise on the systematic study of oratory (rhetoric), and Quintilian's discussions of the minute technicalities of his subject properly belong to an age in which the primary aim of education was to produce skilled public speakers. But his general observations on education are of permanent value. He advises the teacher to apply different teaching methods according to the different characters and abilities of his pupils; he believes that the young should enjoy their studies and knows the value of play and recreation; he warns against the danger of discouraging a pupil by undue severity; he makes an effective criticism of the practice of corporal punishment; he depicts the schoolmaster as taking the place of a parent. "Pupils," he writes, "if rightly instructed regard their teacher with affection and respect. And it is scarcely possible to say how much more willingly we imitate those we like."

Quintilian was born *c.* AD 35 at Calagurris (modern Calahorra) in northern Spain. He was probably educated in Rome, where he afterward received some practical training from the leading orator of the day, Domitius Afer, and then practiced for a time as an advocate in the law courts. He left for his native Spain sometime after 57 but returned to Rome in 68 and began to teach rhetoric, combining this with advocacy in the law courts. Under the emperor Vespasian (ruled 69–79) he became the first teacher to receive a state salary for teaching Latin rhetoric, and he also held his position as Rome's leading teacher under the emperors Titus and Domitian, retiring probably in 88. Toward the end of Domitian's reign (81–96) he was entrusted with the education of the Emperor's two heirs (his grandnephews); and through the good agency of the boys' father, Flavius Clemens, he was given the honorary title of consul (*ornamenta consularia*). His own death, which probably took place soon after Domitian's assassination in 96, was preceded by that of his young wife and two sons.

The systematic study of oratory

His great work, the *Institutio ouatoria,* in 12 books, was published shortly before the end of his life. He believed that the entire educational process, from infancy onward, was relevant to his major theme of training an orator. In book I he therefore dealt with the stages of education before a boy entered the school of rhetoric itself, to which he came in book II. These first two books contain his general observations on educational principles and are notable for their good sense and insight into human nature. Books III to XI are basically concerned with the five traditional "departments" of rhetoric: invention, arrangement, style, memory, and delivery. He also deals with the nature, value, origin, and function of rhetoric, and with the different types of oratory, giving far more attention to forensic oratory (that used in legal proceedings) than to other types. During his general discussion of invention he also considers the successive, formal parts of a speech, including a lively chapter on the art of arousing laughter. Book X contains a famous and much-praised survey of Greek and Latin authors, recommended to the young orator for study. Sometimes Quintilian agrees with the generally held opinion of a writer, but he is often independent in his judgments, especially when discussing Latin authors. Book XII deals with the ideal orator in action, after his training is completed: his character, the rules that he must follow in pleading a case, the style of his eloquence, and when he should retire.

Quintilian's concept of a good orator

The *Institutio* was the fruit of Quintilian's wide practical experience as a teacher. His purpose, he wrote, was not to invent new theories of rhetoric but to judge between existing ones, and this he did with great thoroughness and discrimination, rejecting anything he considered absurd and always conscious that theoretical knowledge alone is of little use without experience and good judgment. The *Institutio* is further distinguished by its emphasis on morality, for Quintilian's aim was to mold the student's character as well as to develop his mind. His central idea was that a good orator must first and foremost be a good citizen; eloquence serves the public good and must therefore be fused with virtuous living. At the same time, he wished to produce a thoroughly professional, competent, and successful public speaker. His own experience of the law courts gave him a practical outlook that many other teachers lacked, and indeed he found much to criticize in contemporary teaching, which encouraged a superficial cleverness of style (in this connection he particularly regretted the influence of the early-1st-century writer and statesman Seneca the Younger). While admitting that stylish tricks gave an immediate effect, he felt they were of no great help to the orator in the realities of public advocacy at law. He attacked the "corrupt style," as he called it, and advocated a return to the more severe standards and older traditions upheld by Cicero (106–43 BC). Although he praised Cicero highly, he did not recommend students slavishly to imitate his style, recognizing that the needs of his own day were quite different. He did, however, appear to see a bright future for oratory, oblivious to the fact that his ideal—the orator-statesman of old who had influenced for good the policies of states and cities—could not exist in times when there was no longer any political freedom.

Declamations attributed to Quintilian

Two collections of declamations attributed to Quintilian have also survived: the *Declamationes majores* (longer declamations) are generally considered to be spurious; the *Declamationes minores* (shorter declamations) may possibly be a version of Quintilian's oral teaching, recorded by one of his pupils. A speech that he is known to have published early in life and a work called *De causis corruptae eloquentiae* ("Reasons for the Decline of Oratory") have not survived. The text of his *Institutio* was rediscovered by a Florentine, Poggio Bracciolini, who, in 1416, came across a filthy but complete copy of it in an old tower at St. Gall, Switzerland, while he was on a diplomatic mission there. Its emphasis on the dual importance of moral and intellectual training was very appealing to the 15th and 16th centuries' Humanist conception of education. Although its direct influence diminished after the 17th century, along with a general decline in respect for the authority of classical antiquity, the modern view of education as all-round character training to equip a student for life follows in a direct line from the theories of this 1st-century Roman.

BIBLIOGRAPHY. The best text of Quintilian is that edited by M. WINTERBOTTOM (Oxford Classical Texts, 2 vol., 1970). There is an English translation by H.E. BUTLER (with text) in the Loeb Classical Library, 4 vol. (1921–22). G.L. SPALDING'S edition with Latin commentary, 6 vol. (1798–1834; vol. 6, index by E. BONNELL), is still useful. Three individual books have been edited with full commentary and introductions: book I by F.H. COLSON (1924); book X by W. PETERSON (1892; 2nd ed., 1939); and book XII by R.G. AUSTIN (1948). See GEORGE KENNEDY, *Quintilian* (1969).

(M.L.C.)

Qur'ān

The Qur'ān (Arabic "reading," "recitation"; often spelled Koran), the holy book of Islām, regarded by believers as the true word of God, was revealed to the Prophet Muhammad and collected in book form after his death. It is accepted as the earthly reproduction of an uncreated and eternal heavenly original, according to the general view referred to in the Qurʾān itself as "the well-preserved tablet" (al-awh *al-maḥfūẓ;* Qurʾān 75:22). The word *qurʾān* is derived from the verb qara'a "to read," "to recite," but there is probably also some connection with Syriac *qeryānā*, "reading," used for the scriptural lessons in the Syrian Church. In the Qurʾān itself the word is not used with reference to the book as a whole but only as a term for separate revelations or for the divine revelation in general. The Qurʾān is held in high esteem as the ultimate authority in all matters legal and religious and is generally regarded as infallible in all respects. Its Arabic language is thought to be unsurpassed in purity and beauty and to represent the highest ideal of style. To imitate the style of the Qur'Bn is a sacrilege.

FORM AND CONTENT

Division into chapters and verses

Form. In length the Qur'Bn is approximately comparable with the New Testament. For purposes of recitation during the holy month of Ramadan it is divided into 30 "portions" (juz', plural *ajzāʾ*), one for each day of the month. Its main division, however, is into 114 chapters, called *sūrah*s, of very unequal length. With the exception of the first *sūrah*, the so-called *fātiḥah* ("opening" of the book), which is a short prayer, the *sūrah*s are arranged roughly according to length, siirah 2 being the longest and the last two or three the shortest. Because the longest *sūrah*s generally derive from the latter part of Muhammad's activity, the consequence of this arrangement is that the oldest *sūrah*s are generally to be found toward the end of the book and the youngest generally appear at its beginning.

In the accepted version now in use, each siirah has a heading containing the following elements: (1) a title, which is usually derived from some conspicuous word in the siirah, such as "The Cow," "The Bee," "The Poets," but usually not indicating the contents of the whole chapter; (2) the *basmalah*; *i.e*., the formula "In the name of God, the Merciful, the Compassionate"; (3) an indication of whether the siirah was revealed at Mecca or at Medina and of the number of its verses; and finally (4) in some cases one or more detached letters; *e.g.*, *tāʾ* sin, *tāʾ* sin mim, or alif *lām mīm*, the meaning of which has not been satisfactorily explained, though it is thought that they might stand for abbreviated words, indicate certain collections of siirahs, or have a magical significance.

The verses in the Qurʾān are called ayah (plural *āyāt*, literally "signs") and vary considerably in length. The shortest verses generally occur in the earliest *sūrah*s, in which the style of Muhammad's revelation comes very close to the rhymed prose (*sajʿ*) used by the *kāhin*s, or soothsayers, of his time. As the verses get progressively longer and more circumstantial, the rhymes come farther and farther apart. There is also a change of linguistic style: the earlier *sūrah*s are characterized by short sentences, vivid expressions, and poetic force; and the later ones become more and more detailed, complicated and, at times, rather prosaic in outlook and language. As a result, it is sometimes difficult to decide whether or not a rhyme is intended to indicate the end of a verse; and consequently, there are variations in the numbering of verses (*e.g.*, between the European editions long used by Western scholars and the official Egyptian edition that has now replaced them in most scholarly works).

The Qurʾan generally appears as the speech of God, who mostly speaks in the first person plural ("we"). When the prophet Muhammad is speaking to his compatriots, his words are introduced by the command, "Say," thus emphasizing that he is speaking on divine injunction only. At times the form is also dramatic, bringing in objections by Muhammad's opponents and answering thern by counter-arguments. Narrative passages are mostly brief. Stories of prophets and biblical persons are often alluded to as though they are known to the audience. The stress is not on the narrative but on its didactic uses.

Hetero-geneous style

On closer analysis very few of the siirahs turn out to be uniform in style or content. The longest text dealing with one subject is *sūrah* 12, which retells the story of Joseph, adding to the biblical account a great many legendary details, most of which seem to be drawn from Jewish sources. Otherwise the longer *sūrah*s are composed of several brief sections dealing with a variety of topics. Thus the Qurʾān often gives the impression of having been produced by a rather haphazard method of composition, an impression that is further heightened by the fact that certain favourite phrases such as "but God is forgiving, compassionate," "God is knowing, wise," "most of them know nothing" often have little or no connection with the immediate context and seem to have been added in order to produce a needed rhyme.

It is often emphasized that Muhammad brought to his people "an Arabic Qurʾān"; *i.e.*, a book in the Arabs' own language comparable to the holy books of Judaism and Chtistianity. Also the vocabulary of the Qur'Bn is overwhelmingly of Arabic origin, but there are, nevertheless, loan words, mostly from Hebrew and Syriac, bearing witness to Muhammad's debt to Judaism and Christianity. These loan words are primarily technical terms such as injil, "gospel" (Greek evangelion); *taurāt*, "the law, or Torah" of Judaism; *Iblīs*, "the devil" (Greek diabolos); or translations or adaptations of theological terms such as *amanā*, "to believe" (Hebrew or Aramaic); *salāt*, "prayer" (probably Syriac). Such explanations are usually regarded with suspicion by Muslims, since orthodox doctrine is that the language of the Qurʾān is the purest Arabic.

Content. It is difficult to classify the contents of the Qurʾān. If the material is arranged chronologically, certain patterns appear since the predominant interest is different in various periods.

God: His nature *and* design for creation. The earliest siirahs concentrate on God as the creator of the world, whose beneficence should arouse the gratitude of mankind and who recompenses or punishes man according to his attitudes toward him. References to the sudden advent of the last judgment and descriptions of the bliss of paradise and the torment of hell complete the picture. Strangely enough, there is no reference to the oneness of God in these early chapters. According to one tradition, on one occasion Muhammad even acknowledged the relative authority of three goddesses, al-Lāt, Manāt, and al-ʿUzzā, but later on abolished the passage in which this reference occurred. There are also a few allusions to the ritual of prayer.

Tradition of previous prophets

Later *sūrah*s place much emphasis on the doctrine that there is but one God, while the other gods of the Arabs are said to be only powerless idols. The references to the last judgment, to paradise and hell are fewer and shorter. On the other hand, there are many polemic utterances: against the idolaters, against those who are ungrateful and do not believe in Muhammad's message. In this connection there are several references to previous prophets, who bad warned their people but were met with disbelief, and thus catastrophe befell the unfaithful. These prophets serve as examples; their lack of success also reflects the experience of Muhammad, and it is implied that the outcome would be similar in his case as well. One implication of this is that Muhammad is one in a long series of prophets who have been sent by God to warn their peoples against the imminent judgment, or, to be more exact, the last and final link in the chain of prophets with a divine message, much in the same way as Mani (3rd century AD Iranian reformer of Zoroastrianism who advocated that matter is evil) regarded himself as the last in the row of revealers of divine truth. It is to be noted that some of the prophets referred to are biblical persons (Noah, Moses, Abraham, Jesus), while others seem to be derived from native Arabic traditions (Hūd, Ṣāliḥ). Major Christian and Jewish figures are common; there is frequent mention of Mary, Zacharias, and John the Baptist, as well as of David, Solomon, Job, and Jonah.

Toward the end of Muhammad's activity in Mecca the earlier mentioned change in style occurs: the verses grow longer, poetic, and often elliptic language is exchanged for a much calmer and prosaic style. Several parables occur in this period; *e.g.*, the rain reviving vegetation is used to illustrate God's resurrecting the dead; the story of seafarers who are surprised by a strong wind and pray to God for help but then forget him as soon as they are saved exemplifies the fickleness of human nature. Verses from earlier revelations are often repeated with additional elaboration. The power of God and the wonders and wisdom of creation are the themes that are elaborated upon. The descriptive element is less pronounced in the eschatological passages (about the end of time); the emphasis is on the fact of intervention by the Lord of Justice. The references to earlier prophets are further developed but Jesus is mentioned less frequently. The oneness of God is emphasized more than ever, and it is emphasized that false gods will not be able to help their worshippers on the day of judgment. In addition to God's omniscience, the problem of his omnipotence then comes to the fore.

Human destiny. Man's destiny is entirely in God's hand, even faith and disbelief are dependent on his will. "They would not believe unless God willed (Qur'Bn 6:111). There is no freedom of will, nor is the Prophet to blame for disbelief, for in the last recourse the decision rests with God in his eternal predestination. But other passages fail to press the point and appear to leave man some freedom to listen to the Prophet's preaching and make his own choice for good or for evil. Muhammad's role as a warning prophet is emphasized. In this connection the references to the row of earlier prophets are elaborated and systematized. It is emphasized that Muhammad's preaching confirms earlier revelation. Abraham appears as the founder of Arabian monotheism. In a way Muhammad is his successor, and there are obvious efforts to establish relations with the Jewish tradition.

Eihical and ritual guidance. In this period there is some interest in ethical commandments. The duty of almsgiving is inculcated—as for ritual practices only prayer seems to be mentioned. On the other hand certain rules concerning forbidden food appear.

In the siirahs revealed at Medina the abovementioned stylistic development is continued. The practical interests of the new Muslim community come into focus. Several revelations deal with various episodes in Muhammad's military operations, encouraging the brave and faithful and blaming the hesitant. Ritual and legal prescriptions are common and detailed. Questions concerning the organization of the community are dealt with, rules of conduct in intercourse with the Prophet are given, laws of matrimony and inheritance and the ritual practices of fasting and pilgrimage are regulated, and so forth. The hostility of the Jews is met by accusations that they have altered the scriptures and abandoned the religion of Abraham, the founder of the *Kaʿbah* (a cube-shaped Muslim holy place in Mecca).

The revelation of the various portions of the Qur’ān met the needs and answered the questions of each period. Sometimes, it even dealt with the personal affairs of Muhammad and his contemporaries. There is no doubt of the Prophet's sincere conviction that he had received the word of God on every occasion.

ORIGINS OF QUR’AN

Revelation of the Qur'Bn to the Prophet

According to Muslims. According to Muslim tradition the Qur'Bn was revealed to Muhammad in separate pieces over some 20 years. On such occasions, Muhammad, it is said, was in a kind of trance or ecstasy, during which the revelations were brought to him by the angel Gabriel. On his return to normal consciousness he recited the words of revelation to those present. There are many traditions about the occasions on which a certain siirah or part of a shrah was revealed. Thus the revelation of the Qur’ān is connected with events in the life of the Prophet. Even the traditional recension (version) of the Qur'iin itself classifies the shrahs as Meccan or Medinan.

Obviously, many people learned the words of the revelation by heart, but there are also traditions that, at the time of their revelation, Muhammad had them written down on "pieces of paper, stones, palm-leaves, shoulder-blades, ribs, and bits of leather," *i.e.,* whatever writing-material there was at hand. It is believed that the Prophet indicated to the scribes the context in which a certain passage should be placed.

Establishment of an authoritative text

After the Prophet's death, and especially after the battle of Yamāmah (633), in which a great number of those who knew the Qur'Bn by heart had fallen, fear arose that the knowledge of the Qur'Bn might disappear. So it was decided to collect the revelations from all available written sources and, as Muslim tradition has it, "from the hearts [*i.e.,* memories] of people." A companion of the Prophet, Zayd ibn Thābit, is said to have copied on sheets whatever he could find and to have handed it over to the caliph 'Umar. After ʿUmar’s death the collection was left in the care of his daughter Ḥafṣah. Other copies of the Qur’ān appear to have been written later, and different versions were used in different parts of the Muslim empire. So that there would be no doubt about the correct reading of the Qur'Bn, the caliph ʿUthmān (644–656) is reported to have commissioned Zayd ibn Thābit and some other learned men to revise the Qur'iin using the "sheets" of Ḥafṣah, comparing them with whatever material was at hand, and consulting those who knew the Qur'Bn by heart. It was decided that in case of doubt about the pronunciation, the dialect of Quraysh, the Prophet's tribe, was to be given preference. Thus an authoritative text of the Qur'Bn (now known as the ʿUthmānic recension) was established.

These traditions may have been reworked and changed to some extent to suit certain dogmatic theories concerning the Qur’ān, but in the main they reflect historical truth. It is obvious that the description of the method of revelation has been somewhat simplified. The Qur’ān itself states (42:50–52) that God spoke to Muhammad "by suggestion, or from behind a veil, or by sending a messenger to suggest what he pleases." The first term (Arabic *waḥy*) denotes a "suggestion" or "inspiration" of the kind that is well known by many poets; the Qur'Bn also uses a term meaning "it was sent down." The second term seems to suggest some kind of imaginative locution without any accompanying vision. Only the third expression alludes to an angel but without mentioning the name of Gabriel.

According to orientalists. The chronology of the *sū*rahs is a much debated problem. The existing traditions concerning the occasions for the revelation of certain passages cannot always be controlled and may or may not be reliable. European scholars have applied the criteria of style and contents to establish the relative order of the shrahs or parts of *sūrahs*. From the time when Theodor Noldeke published his *History* of the *Qur’ān* (1860), it has been common to arrange the *sūrahs* in four groups, deriving from three subsequent periods at Mecca and from Medina. The above exposition of the content of the Qur'Bn roughly follows this arrangement.

In the Muslim view, Muhammad received every word of the Qur'Bn directly from God. The Qur'Bn describes, and indignantly rejects, accusations that the Prophet had reproduced things that he had drawn from other sources.

Question of influence of oral traditions on text

Western scholars who have analyzed the contents of the various revelations have shown that much of the narrative material concerning biblical persons and events is not derived from the Bible, but from later Christian and, above all, from Jewish sources (*e.g.,* Midrash). Other motifs, such as the idea of the impending judgment and the descriptions of paradise agree with standard topics in the missionary preaching of the contemporary Syriac church fathers. The dependence need not, however, be of a literary kind, but might be due to influence from oral traditions.

It would appear that learning the words of the revelation by heart was the normal way of preserving them, and that only on special occasions were the words written down immediately. The existence of various early collections of Qur’ānic material seems to be a warranted fact, although their nature and contents cannot be determined. Some of the siirahs beginning with separate letters (*al-fawātiḥ*)—certain consonant combinations detached from the main text (mentioned above)—occur together in the present Qur'iin and in the order of decreasing length in such a way as to suggest that they once formed separate collections. The establishment of a vulgate recension (a standard version) was not sufficient to secure the uniform and correct leading of the Qur'Bn in all details. The Arabic script was incomplete; several consonants were easy to confuse, and there was no way of indicating the vowels to differentiate the variety of possible meanings inherent in a particular combination of consonants. To assure the correct recitation, therefore, it was necessary to know the text more or less by heart. In this way, differing variant readings arose, warranted by this or that "reader" of the Qur'Bn. The recorded variations, however, turned out to be remarkably few, and though no complete listing of the textual variants exists, it can safely be said that the textual tradition of the Qur'Bn is much firmer and more uniform than that of the New Testament. The Arabic script was gradually improved. Diacritical signs were introduced to distinguish the letters that were similar in form, and long vowels were indicated by the letters *alif* (for *ā*), *wāw* (for *ū*), and yd (for **i**). It is known that this vowel system was still disputed at the beginning of the 9th century. The special vowel signs placed above or beneath the letters were added in a different colour and did not count as part of the text itself.

Interpretations. The "readers" (qurrd', singular *qārī'*) were the specialists of the text of the Qur'ān. They were at the same time philologians, and it was to a great extent from their dealings with the language of the Qur'Bn that the science of Arabic grammar grew. Two schools developed, one at Baṣra (in present-day Iraq), which was especially interested in systematizing and ordering the material to set up the rules governing the language, and a rival one at Kiifa (also in Iraq), which took more interest in the exceptional. It was theorized that several variant readings could be accepted only if they were based on the ʿUthmānic recension (version). It was also important that a reading be based on the authority of some renowned reader.

There was also theological speculation as to the true nature of the Qur'Bn. In the discussions initiated by the Mu'tazilites (lit. "those who stand apart"; a group that sought to introduce principles from Greek rationalism into Islāmic thought) the question of the eternity of the Qur'Bn (*i.e.,* of its heavenly prototype) was one of the main points. The Muʿtazilites, who wanted to avoid everything that might encroach upon the oneness of God, denied the doctrine that the Qur'an was uncreated and eternal, because this would mean that something else besides the God of eternity would exist eternally and thus create an eternal and irreconcilable "dualism." Consequently they asserted that the Qur'Bn was created by God. This doctrine, however, was rejected by orthodox Islām. In popular belief, the reverence for the Qur'Bn is often directed toward the visible book or parts of it. Oaths are taken on it, passages are copied for magical purposes.

Qur'Bnic exegesis

In these and other doctrinal disputes the parties sought support for their opinions in the sayings of the Qur'ān, since it was considered as the ultimate authority in all legal and religious questions. The correct interpretation of the Qur'Bn became the object of a special branch of learning, the so-called tafsir, or Qur'Bnic exegesis. All kinds of resources were utilized in order to elucidate the meaning of a Qur'Bnic passage. Traditions concerning the circumstances surrounding the revelation of certain passages or containing interpretative utterances of the Prophet that had been transmitted orally were recorded and collected, together with other traditions deriving from and concerning the Prophet (Hadith). At times, in order to provide authority for a certain theory, traditions were simply invented. Any interpretation of a Qur'ānic passage that could not be supported by a Hadith was originally rejected. The results of the study of grammar and lexicography were also utilized; examples from contemporary poetry were often quoted in order to elucidate the grammatical structure or the lexical meaning of a passage. Thus, work on the Qur'Bn, whose ultimate goal was the correct understanding and application of its teachings, went hand in hand with the development of Arabic grammar and lexicography.

Two works are especially renowned in the field of tafsir, namely the commentary of aṭ-Ṭabarī (839–923), a huge encyclopaedic collection that sums up everything that had been done so far in the field, and the *Kashshāf* of Zamakhshari (1075–1143), which has gained almost canonical reputation, though its author was a Mu'tazilite and began his work with the words, "Praise be to God who created the Qur'Bn." A handy commentary of Bayḍāwī (d. *c.* 1280), which is often quoted as authoritative, is merely an abridged revision of the latter work.

The theological schools of medieval Islām all sought to support their doctrines with the aid of Qur'Bnic exegesis, and each of them produced their own commentaries. There are also examples of allegorical interpretation (ta'wil) especially in Ṣūfī (Islāmic mystical) literature, in which the doctrines of mysticism are found to be hidden behind the literal sense of the Qur'Bnic word.

Qur'Bnic exegesis gained new significance with the appearance of modernism toward the end of the 19th century. The modernists, who sought to revive Islām from its degradation and to reconcile it with what they found valuable in Western scientific traditions, set up the principle of returning to the pure and uncorrupted Islām of the "ancestors." As a consequence, the interpretation of the oldest and original source of Islām was regarded as imperative, and attempts were made to establish the principles necessary for a correct understanding of the Qur'Bn. Traditional exegesis was accused of having introduced Israelite legends and false traditions that had nothing to do with the original teachings of the Prophet. On the other hand, the authority of the Qur'Bn was never called in question.

Modern commentaries

Muhammad 'Abduh, the founder of modernism in Egypt, for several years published exegetical lectures in the journal al-Mandr; and they were later published in book form by his Syrian disciple Rashīd Riḍā. In them he accepts the Qur'Bn as the literally inspired word of God, in which there can be nothing false or antiquated, and tries to show that the results of modern science and many modern views are already present in the Qur'Bn. This is often achieved by twisted interpretations, reading modern ideas into the words of the Qur'Bn. For instance, the *jinn* (genii) of *sūrah* 2:176 that cause disease are interpreted as "microbes," and the words in 2:250, "How often a little company has overcome a numerous company; and God is with those who endure," is taken to refer to ideas reminiscent of Darwin's theory of the struggle for life and the survival of the fittest. Allegorical interpretation is also used when it can serve the purpose of the author. Other modernistic interpreters of the Qur'Bn have continued along the same lines. The Qur'Bn is, however, left untouched by criticism; as the infallible word of God it cannot have been influenced by the circumstances under which it was revealed, it can contain no mistake, and it cannot be superseded by any new discovery.

The latest development, though, has brought some new ideas to the fore. In an Urdu commentary on the Qur'Bn, which has in part been made available in English, Maulana Abul Kalam Azad (1888–1958), an Indian Muslim scholar (minister of education of the Republic of India at the time of his death), develops some new principles for the interpretation of the Qur'Bn. He argues that it is necessary to interpret the Qur'Bn against the background of its environment; therefore it is necessary to study the cultures and the languages of ancient Arabia and other Semitic peoples. Study of the historical circumstances in which the Qur'Bn came into being is said to facilitate the understanding of what it meant to those who received the revelation.

Scholars have no doubt, however, that something new is entering the field of Qur'Bnic exegesis. D. Rahbar, in his study The God *of* Justice (1960), argues that in order to elucidate a passage in the Qur'Bn one should quote traditional exegesis and medieval dogmatics and, above all, use other Qur'Bnic passages for comparison, letting one passage throw light on another. Though such ideas are looked upon with suspicion by orthodox Muslims and are fervently rejected by most Muslim leaders, they may indicate the inception of a more historical view of the

Qur'ān, one that tries to distinguish between central religious ideas and those outward things that are dependent on the historical environment.

TRANSLATIONS

The Qur'ān was revealed to Muhammad as "an Arabic book" or an Arabic reading (*qur'ān*), to provide the Arabs with a holy book in their own language, comparable with the Scriptures of Judaism and Christianity. As has been noted, the language of the Qur'ān is regarded as surpassing everything that can be written in Arabic. The Qur'Pn itself is a miracle and cannot be imitated by man.

As a consequence of this, it is regarded as unfitting to translate the Qur'ān. In countries in which other languages are spoken, the Qur'ān is still recited in Arabic. There exist Muslim translations of the Qur'Pn; *e.g.*, into Turkish, Urdu, and English (the latter during the Aḥmadiyah movement founded in 1889 by Mirza Ghulam Aḥmad in the Punjab region of India), but on principle these are regarded as paraphrases. not as translations that can be used for ritual purposes.

The Qur'Pn was first printed in Arabic at Rome by Pagninus Brixiensis (1530), but the edition was never circulated. A. Hinckelmann published an Arabic text at Hamburg in 1694. Since then several European editions have appeared; one of the best was that of G. Fliigel (1834), the first critical edition, often reprinted. It is from this edition that Western scholars have usually quoted the Qur'ān. Several editions are today printed in Muslim countries, and an official Egyptian edition is gaining more and more ground among Western scholars.

First critical edition

The first Latin translation was made in 1143 at the request of an abbot of the monastery of Cluny and was published at Basle in 1543 by Theodor Bibliander and afterward rendered into Italian, German, and Dutch. The first French translation was by A. du Ryer (1647); it was translated into English by Alexander Ross (1649–88). G. Sale's English translation first appeared in 1734 and has passed through many new editions. It has become something of a classic and can still be useful in many respects. A translation by J.M. Rodwell, with the *sūrah*s arranged in chronological order, appeared in 1861. E.H. Palmer's translation was published in Sacred Books of the East in 1880. Bell's translation "with a critical rearrangement of the sūrahs" (1937–39) tries to analyze the *sūrah*s into their smallest units and show how these were joined together to form the present Qur'Pn. The translation (1955) of A.J. Arberry, distinguished British scholar of Islām, is well known for its literary qualities, and is highly esteemed, especially by Muslims, for its rendering of Qur'ānic style.

The Qur'ān has also been translated into most other European languages. Special mention should be made of R. Blachère's French translation (1949–50) because of its rather detailed notes, and of R. Paret's German rendering (1962), which is very accurate and makes extensive use of parallel passages within the Qur'ān itself, but is rather dry in its style.

BIBLIOGRAPHY. The basic work is T. NOLDEKE, *Geschichte des Qorans* (1860), *2nd* ed. by F. SCHWALLY (1919–38). Less comprehensive but more modern are R. BELL, *Introduction to the Qur'an* (1953); and R. BLACHERE, *Introduction au Coran* (1947). The history of Qur'ānic interpretation is set forth in I. GOLDZIHER, *Die Richtungen der islamischen Koranauslegung (1920).* It should be supplemented by J.M.S. BALJON, *Modern Muslim Koran Interpretation, 1880-1960* (1961, reprinted 1968). A. JEFFERY, The *Qur'ān as Scripture* (1952), deals with the Qur'ān's view of its own function.

(H.R.)

Rabelais, François

A Franciscan priest as well as an eminent physician, François Rabelais is noted primarily as the author of the comic masterpiece *Gargantua and Pantagruel.* The four novels composing this work are outstanding for their rich use of Renaissance French and for their comedy—ranging from gross burlesque to profound satire—covering the major intellectual and moral preoccupations of the day: legal, medical, political, religious, philosophical. The novels exploit popular legends, farces, and romances, as well as classical and Italian material, but were written, at least in part, for a court public and a learned one. The adjective Rabelaisian applied to scatological humour is misleading; Rabelais uses scatology aesthetically, not gratuitously, for comic condemnation. His creative exuberance, colourful and wide-ranging vocabulary, and literary variety continue to ensure his popularity.

Cliche Musees Nationaux, Paris

Rabelais, oil painting by an unknown artist, 17th century. In the Musée National de Versailles et des Trianons.

Early life. Details of Rabelais's life are sparse and difficult to interpret. He was probably born about 1483 (some believe *c.* 1494, though this is less likely), the son of Antoine Rabelais, a rich Touraine landowner and a prominent lawyer who deputized for the *lieutenant-général* of Poitou in 1527. After apparently studying law, Rabelais became a Franciscan novice at La Baumette (1510?) and later moved to the Puy-Saint-Martin convent at Fontenay-le-Comte in Poitou. By 1521 (perhaps much earlier) he had taken holy orders, a step that presupposes serious scholastic studies in the Franciscan tradition, particularly of John Duns Scotus and St. Bonaventura. The eminent French Hellenist Guillaume Budé, in a letter to a mutual friend, praised Rabelais's legal learning and his mastery of Greek and Latin. At Fontenay-le-Comte, Rabelais frequented legal Humanists, including Amaury Bouchard and André Tiraqueau, a friendship that survived a quarrel between the two concerning the dignity of women. In 1524 Rabelais wrote a eulogistic poem in Greek that Tiraqueau prefixed to some editions of *De legibus connubialibus.* Rabelais's *Tiers Livre* (1546) makes use of legal erudition borrowed from Tiraqueau's *De Nobilitate,* a work he must have consulted in manuscript, as it was not published until much later. The prologue to Rabelais's *Quart Livre* (1552) was to contain a eulogy of Tiraqueau, though it is possible that this friendship had cooled and that Tiraqueau supported the banning of the *Tiers Livre.*

Friendship with leading Humanists

Rabelais was closely associated with Pierre Amy, a liberal Franciscan Humanist of international repute. In 1524 the Greek books of both scholars were temporarily confiscated by superiors of their convent, because Greek was suspect to the hyperorthodox as an "heretical" language: it opened up the original New Testament and encouraged the study of Greek Fathers of the Church, such as Origen and Clement of Alexandria. Both scholars may have been temporarily imprisoned in their convent. Rabelais then obtained an indult (a special privilege) from Pope Clement VII and was removed to the Benedictine house of Saint-Pierre-de-Maillezais, the prior of which was his bishop, Geoffroy d'Estissac. He never liked his new order, however, and later satirized the Benedictines in the person of Frère Jean, though he passed lightly over Franciscan shortcomings. His only known works of this period are a (lost) Latin translation of the first book of Herodotus and of a dialogue of Lucian (also lost), a French poem addressed to Jean Bouchet (an orthodox poet and historian), and the Greek eulogy of Tiraqueau.

Rabelais studied medicine, probably under the aegis of the Benedictines in their Hôtel Saint-Denis in Paris. In 1530 he broke his vows—though later in his *Supplicatio pro apostasia* ("Supplication for Apostasy") he claimed to have worn the habit of a secular priest and to have regularly celebrated mass. He studied medicine at Montpellier, probably with the support of his patron, Geoffroy d'Estissac. Graduating within weeks, he lectured on the works of distinguished ancient Greek physicians, publishing his own editions of Hippocrates' *Aphorisms* and Galen's *Ars parva* in 1532. As a doctor he placed great reliance on classical authority, siding with the Platonic school of Hippocrates but also following Galen (when not opposed to Hippocrates) and Avicenna (980–1037), the Persian philosopher and physician.

During this period an unknown widow bore him two children (François and Junie), who were given their father's name and were eventually legitimated by Pope Paul IV in 1540. A third child, Théodule, was recognized by Rabelais but died in infancy.

First novels. ***Pantagruel.*** After apparently practicing medicine briefly in Narbonne, Rabelais was appointed physician to the hospital of Lyons, the Hôtel-Dieu, in 1532. In the same year he edited the medical letters of Giovanni Manardi, a contemporary Italian physician, as well as a Latin will that later proved to be a fabrication. On November 30 he wrote an enthusiastic letter to Erasmus, the leading Humanist of the time. It was during this period that he discovered his true talent. Fired by the success of an anonymous popular chapbook, *Les Grandes et inestimables cronicques du grant et e'norme géant Gargantua,* he published, pseudonymously, his first novel, *Pantagruel* (now usually read second). *Pantagruel* is slighter in length and intellectual depth than his later novels, but nothing of this quality had been seen before in French in any similar genre. Rabelais displayed his delight in words, his profound sense of the comedy of language itself, his mastery of comic situation, monologue, dialogue, and action, and his genius as a storyteller who was able to create a world of fantasy out of words alone. Within the framework of a mock-heroic, chivalrous romance, he laughed at many types of sophistry, including legal obscurantism and hermeticism, which he nevertheless preferred to the Scholasticism of the Sorbonne. One chapter stands out for its sustained seriousness, praising the divine gift of fertile matrimony, a compensation for death caused by Adam's fall. Rabelais maintains that the Christian father, having enjoyed the privilege to beget legitimate heirs, must educate his sons to become mirrors of his own body and soul conjoined. The work borrows openly from Sir Thomas More's *Utopia* in its reference to the war between Pantagruel's country, Utopia, and the Dipsodes but gaily preaches a semi-Lutheran doctrine—that nobody but God and his angels may spread the gospel by force. *Pantagruel* is memorable as the book in which Pantagruel's companion, Panurge, a cunning and witty rogue, first appears.

The genius of *Pantagruel*

This successful novel was followed in 1532 by the *Pantagrueline Prognostication,* a parody of the almanacs, astrological predictions that exercised a growing hold on the Renaissance mind, particularly after the great conjunction of planets in 1524. As a doctor, Rabelais could (and did) compose almanacs and prognostications. He disbelieved in "judicial" (essentially fortune-telling) astrology, however, and joined forces with authors who, in the name of divine Providence, attacked judicial astrology. His satirical form here owes much to German Latin models but surpasses them in quality. Suddenly, in 1534, Rabelais left the Hôtel-Dieu to travel to Rome with the bishop of Paris, Jean du Bellay. He returned to Lyons in May of that year and published an edition of Bartolomeo Marliani's description of Rome, *Topographia antiquae Romae.* He returned to the Hôtel-Dieu but left it precipitately in February 1535, perhaps because of the persecution following the Affaire des Placards (October 1534), an incident of the Reformation in which France was placarded with posters that openly attacked the "idolatry" of the mass. Even Catholic evangelicals, who sought truth in religion through recourse to scriptural authority, were accused of "Lutheranism" and were forced to flee.

Gargarztua. *Gargantua* belongs to this period. The second edition is dated 1535; the first edition lacks the title page in the only known copy; it probably also dates from 1535, though many believe it was first published in 1534. This masterpiece was, in part, designed to support royalist causes espoused by Jean du Bellay, who had been created cardinal, essentially by King Francis I, in May 1535. The King, the Cardinal, and his brother Guillaume du Bellay, seigneur de Langey, were in contact with the German theologian Philip Melanchthon, hoping to win his support for an ecumenical effort to end the Lutheran schism and to counter-reform the church in France. Meanwhile, diehard Scholastic theologians of the Sorbonne, such as Noel Beda, were kept in prison. *Gargantua* advances these causes, preaching steadfastness in face of persecution but condemning the persecution itself. The Sorbonne is mocked, evangelism propagated. In *Gargantua,* Rabelais continues to exploit the romances mock-heroically, telling of the birth, education, and prowesses of the giant Gargantua, Pantagruel's father. Much of the satire—for example, mockery of the ignorant trivialization of the mystical cult of emblems and of erroneous theories of heraldry—is calculated to delight the court; much also aims at delighting the learned reader—for example, he sides with Humanist lawyers against legal traditionalists and doctors who accepted 11-month, or even 13-month, pregnancies. Old-fashioned Scholastic pedagogy is ridiculed and contrasted with the Humanist ideal of the Christian prince, widely learned in art, science, and crafts, and skilled in knightly warfare. In politics, Rabelais champions appeasement against bellicosity, though he insists that the Christian prince should, as God's agent, follow unsuccessful appeasement with military might. The war between Gargantua and his neighbour, the "biliously choleric" Picrochole, is partly a private satire of an enemy of Rabelais's father, partly a mocking of Charles V, the Holy Roman emperor, and the imperial design of world conquest. Gargantua commands the operations, but some of the exploits are carried out by Frère Jean (the Benedictine). Lean, lecherous, dirty, and ignorant, Frère Jean is a "true monk if ever there was one, since the world started to monk," but he is redeemed by jollity and active virtue; for his fellow monks are timorous and idle, delighting in those "vain repetitions" of prayers that Christ condemned. Gargantua's last major episode centres around the erection of the Abbey of Thélème, a monastic institution that rejects poverty, celibacy, and obedience in the name of self-discipline. It welcomes wealth and the wellborn, praises the aristocratic life, rejoices in good marriages. Probably as an afterthought, it is also a refuge for persecuted evangelicals.

Themes of *Gargantua*

After *Gargantua,* Rabelais published nothing new for 11 years, though he prudently expurgated his two works of overbold religious opinions. He continued as physician to Cardinal Jean and his powerful brother Guillaume. After leaving the Hôtel-Dieu in February 1535, he eventually accompanied the Cardinal to Rome. There he regularized his position by making a "supplication" for his "apostasy" (*i.e.,* his unauthorized departure from the Benedictine monastery), arranging to enter the Cardinal's Benedictine convent at Saint-Maur-les-Fossés. The convent was secularized six months later, and he became a secular priest, authorized to exercise his medical profession.

In 1537 he joined Budé, Clément Marot, and others in a famous banquet celebrating the release from prison of a fellow Humanist, Étienne Dolet. Rabelais may later have quarrelled with Dolet, who pirated his works without the expurgations. In May of that year he was awarded the doctorate of medicine of Montpellier and subsequently dissected at least two corpses, a fairly rare practice at that time. At Montpellier he delivered, with considerable success, a course of lectures on Hippocrates' *Prognostics.*

He was at Aigues-Mortes in July 1538 when Charles V met Francis I, but his movements are obscure until he followed Guillaume du Bellay to the Piedmont, visiting Turin and Ferrara. Late in 1542 Guillaume set out for Lyons but died in January 1543, in the presence of Rabelais and other doctors, before he reached his destination. Rabelais twice alludes in his works to this important event. He believed that God sent special portents and that Guillaume was granted on his deathbed the gift of prophecy.

Condemnation of the novels

Guillaume's death meant the loss of a hero and the loss of a patron. The same year Geoffroy d'Estissac died as well, and Rabelais's novels were condemned by the Sorbonne and the Parlement of Paris, Rabelais sought protection from the King's sister Margaret, queen of Navarre, dedicating to her his ***Tiers Livre*** (1546). Despite its royal *privilège* (*i.e.*, license to print), the book was immediately condemned for heresy by the Sorbonne, and Rabelais fled to Metz (an imperial city), remaining there until 1547. As counsellor to the city, he earned 120 livres a year but wrote, pleading desperate poverty, to Cardinal Jean:

> If you do not have pity on me, I know not what I should do, other than put myself into service with someone over here . . . with harm and evident loss to my studies. It is not possible to live more frugally than I do.

Later work. ***Tiers Livre.*** The ***Tiers Livre*** is Rabelais's most profound work. Pantagruel has now deepened into a Stoico-Christian inerrant sage; Panurge, a lover of self and deluded by the devil, is now an adept at making black seem white. Panurge hesitates: Should he marry? Will he be cuckolded, beaten, robbed by his wife? He consults numerous prognostications, both good Platonic ones and less reputable ones favoured by such authorities as H.C. Agrippa and Girolamo Cardano—all to no effect because of his self-love. His ideas of human sexuality are "erroneous" Galenic ones, not "sound" Hippocratic ones. He consults a good theologian, a Platonic doctor, and a Skeptic philosopher approved of by the learned giants, but his problem is not treated by the "foolish" Bridoye, a judge who—like Roman law in cases of extreme perplexity—trusted in Providence and decided cases by casting lots. Panurge trusts in no one, least of all in himself. It is therefore decided to consult the oracle of the Dive Bouteille (Sacred Bottle), and the travellers set out for the temple. With Bridoye the novel deepens into a paradoxical defense of Christian "folly," owing something to Erasmus' ***Praise of Folly*** but more to St. Paul as Rabelais understood him. The ***Tiers Livre*** was condemned probably for its mockery of penance, its subordination of canon law to civil law in matrimonial matters—Rabelais denied the validity of marriages contracted without parental consent—and for its threat of civil intervention into monastic affairs. The work ends enigmatically with a mock eulogy (a highly appreciated genre) in which hemp is eulogized for its myriad uses, thus artistically balancing Panurge's perverse eulogy on debts at the beginning of the book.

Quart Livre. From 1547 onward, Rabelais found protection again as physician to Cardinal Jean and accompanied him to Rome via Turin, Ferrara, and Bologna. Passing through Lyons, he gave his printer his incomplete ***Quart Livre,*** which, as printed in 1548, finishes in the middle of a sentence. It contains some of his most delightful comic storytelling but also defends a "synergistic" theology (*i.e.*, men must "work together" with grace for salvation). In Rome he sent to the Cardinal de Guise a description of the "Sciomachie" ("Simulated Battle") organized by Cardinal Jean and his allies to celebrate thk birth of Louis of Orléans, second son of Henry II of France. He published the account in 1549 on his return to Lyons with his patron, whom he tended in an illness at Saint-Maur-les-Fossés.

In January 1551 Cardinal Jean presented him with two benefices at Meudon and Jambet, though Rabelais never officiated or resided there. In 1552, during the height of the Gallican crisis, when France nearly broke with Rome over the dispute on the monarch's jurisdictional rights within the French church, he published—with a new prologue—the full ***Quart Livre,*** his longest book. A wealth of both serious and comic writing, it propagates synergism against Calvinistic predestinationalism, mocks bloodthirsty bishops, dismisses the Council of Trent as a non-Catholic council of nitwits, and champions a "syncretistic" Christianity (a welding together of the best of ancient wisdom and scriptural Christianity). This Christian classicism is symbolized by Plutarch's dying Pan, which Rabelais saw as an allusion to the crucified Good Shepherd, the Christian All.

Views on theology

This work, too, was designed to please a Gallican court and is proudly and aggressively dedicated to the liberal Cardinal Odet de Châtillon. Despite its royal *privilège*, it, too, was condemned by the Sorbonne and banned by Parlement on March 1, 1552; the Gallican crisis had passed. On January 9, 1553, Rabelais resigned his benefices. Rumour had it—probably without foundation—that he was in prison. He died in 1553 (probably on April 9) and was buried in Saint-Paul-des-Champs, Paris.

In 1562 appeared in Lyons the *Isle sonante,* allegedly by Rabelais. It was expanded in 1564 into the so-called ***Cinquiesme et dernier livre.*** The work is partly satirical, partly an allegory; the Dive Bouteille—the ostensible quest of the ***Quart Livre—is*** consulted, and the heroes receive the oraculous advice: "drink" (symbolizing wisdom?). The work cannot be by Rabelais as it stands. Some scholars believe it to be based on his (lost) drafts; others deny it any authenticity whatsoever.

Influence and reputation. Rabelais was connected with some of the great scholars of his age (Budé, Tiraqueau) and with great and powerful political figures (Margaret of Navarre, Guillaume du Bellay) and was consistently protected—though not always effectively—by high-ranking liberal ecclesiastics (Bishop Geoffroy d'Estissac, Cardinal Jean du Bellay, Cardinal Odet de Châtillon). His works are those of a man who had studied several subjects in depth, including scholastic and biblical theology, medicine, and, above all, the law. His reputation for profound Humanist learning was a secure one in his lifetime. Despite his religious irregularities, the Franciscans considered him as a writer of their order. Rabelais's works were placed on the "Index of Forbidden Books" by the Council of Trent and tended to be interpreted in an increasingly Protestant sense when published—as for a long time they could only be—outside France. They have had an enormous influence not only on later French writers, such as Voltaire, Balzac, and Chateaubriand, but also on writers as diverse as Sterne, Swift, Trollope, and Charles Kingsley.

MAJOR WORKS

Gargantua and Pantagruel comprises: ***Pantagruel*** (full title, ***Pantagruel. Les horribles et espoventables faictz et prouesses du très renommd Pantagruel, roy des Dipsodes, filz du grand gdant Gargantua, composez nouvellement par Maistre Alcofrybas Nasier;*** published 1532, forms book ii); ***Gargantua*** (full title, ***La vie inestimable du grand Gargantua, pbre de Pantagruel, jadis compose'e par I'abstracteur de quinte essence;*** 1535, forms book i); ***Tiers Livre*** (full title, ***Tiers Livre des faictz et dictz héroïques du noble Pantagruel, composez par M. Franç. Rabelais, docteur en medicine et calloïer des isles Hières***; 1546); ***Quart Livre*** (full title, ***Le Quart Livre des faictz et dictz hdroiques du noble Pantagruel;*** complete, 1552); ***Le Cinquiesme et dernier livre*** (full title, ***Le Cinquiesme et dernier livre des faicts et dicts héroïques du bon Pantagruel, compose'e par M. François Rabelais, docteur en medicine;*** 1564). Also ***Pantagrueline Prognostication*** (1532).

BIBLIOGRAPHY. P.P. PLAN, ***Bibliographie rabelaisienne: les éditions de Rabelais de 1532 à 1711*** (1904, reprinted 1965); A. TCHEMERZINE, ***Les dditions anciennes de Rabelais, 1532–1742*** (1933); J. PORCHER, ***Exposition organisée à l'occasion du quatribme centenaire de la publicatiorz de Pantagruel*** (1933); J. BOULENGER, ***Oeuvres completes de Rabelais,*** pp. 997–1018 (1955). For further bibliography of studies, see the "Textes littéraires français" editions cited below as well as the following principal periodicals devoted to Rabelaisian scholarship: ***Revue des Etudes rabelaisiennes*** (1903–12); ***Revue du seizibme siècle*** (1913–33); ***Bibliothèque d'Humanisme et Renaissance*** (1934–); ***Etudes rabelaisiennes*** (1956–); and the ***Bulletin des amis de Rabelais et de la Devinière*** (1960–).

Editions: The standard edition is ***Les Oeuvres de Rabelais,*** ed. by ABEL LEFRANC ***et al.*** (1913– ; up to part of ***Le Quart Livre*** published). This edition, however, needs correcting and supplementing by editions of individual works (especially those in the "Textes littéraires français" Series: ***Pantagruel,***

ed. by VERDUN L. SAULNIER, 2nd ed. *(1965); Gargantua,* ed. by RUTH CALDER and M.A. SCREECH *(1970); Le Tiers livre,* ed. by M.A. SCREECH *(1964)*; *Le Quart livre,* ed. by ROBERT MARICHAL *(1947).* Also, for lesser works, *Oeuvres de Rabelais,* ed. by CHARLES MARTY-LAVAUX, *6* vol. *(1870–1903);* and *Oeuvres complètes,* ed. by PIERRE JOURDA, *2* vol. *(1962).*

English translations: The most colourful, though not the most reliable, translation is that of SIR THOMAS URQUHART and PIERRE MOTTEUX, available in "Everyman Classics," ed. by D.B. WYNDHAM LEWIS *(1929,* reprinted *1966).* Very useful are the incomplete translation of W.F. SMITH *(1893,* reprinted *1934);* and J.M. COHEN'S translation *(1955,* reprinted *1970).* See also SAMUEL PUTNAM, *The Portable Rabelais (1946)* and *The Essential Rabelais (1968).*

Biography and criticism: Especially useful for his life are JEAN PLATTARD, *Vie de François Rabelais (1928)* and *François Rabelais (1932);* and SAULNIER'S "Chronologie" in his introduction to *Pantagruel Club du meilleur livre (1962). (Sources)*: JEAN PLATTARD, *L'Invention et la composition dans l'oeuvre de Rabelais* (1909), needs supplementing with aid of more modern editions of "Textes littéraires français." *(Intellectual context)*: M.A. SCREECH, *The Rabelaisian Marriage: Aspects of Rabelais's Religion, Ethics and Comic Philosophy (1958)* and *L'Evangélisme de Rabelais (1959)*; A.J. KRAILSHEIMER, *Rabelais and the Franciscans (1963).* For a Marxist interpretation of Rabelais, see MIKHAIL BAKHTIN, *Rabelais and His World (1968;* orig. pub. in Russian, *1965).*

(M.A.S.)

Races of Mankind

The term race as applied to man has been variously used —by politicians, military leaders, philologists, human biologists, demographers, and historians. Some "races" constitute language groups, often of peoples whose only kinship is that they speak a common language. Such was the original meaning of the so-called Aryan race. Some "races" are simply hypothetical, invented to embrace present distributions of such genetic (hereditary) characteristics as stature or hair colour; *e.g.*, the Nordics. (The word Nordic also has been given a political meaning, referring, despite their differences in physical characteristics, to peoples in Northern Europe.) The term has been variously applied to national or cultural groupings, as in the days when English writers referred to an "Irish race" and to a "Scottish race." The word also has been applied to human groups inferred to have existed on the basis of archaeological discoveries; the "Etruscan race" is an example. Various religious groups who may or may not have common ancestry sometimes are called races— the "Jewish race." By extension of biblical thinking and in honour of Shem, son of Noah, a "Semitic race" was conceived in an effort to describe people who spoke Semitic tongues, some of whom may have learned their language more recently than others.

All of those uses of the term race are separate and distinct from its biological meaning in classification (taxonomy). Just as the term race is often too broadly applied to the entire species of man (as in the "human race"), particular race names invented to explain distributions of grossly observable physical characteristics of human populations are not biologically meaningful. Thus, the early meaning of "Negro race," once used to describe all people with dark skins whatever their ancestry, has been abandoned. The term Negro now is restricted to black people of African ancestry (*i.e.*, African Negroes); it excludes the "blacks" of Australia, the dark-skinned peoples of Southern India or Fiji, and some of the darkest skinned American Indians as well.

Biological basis of races

Today, races are defined neither as hypothetical umbrellas to cover people who share superficial similarities, nor as linguistic, political, or social groupings; nor are they defined from inferences concerning prebiblical human origins. Instead, they are biological races: breeding or mating populations of people (or groups of breeding populations). They are distinguished or classified in terms of genetically transmitted differences. They are studied (in terms of their hereditary origins and their biological relationships with other such races) for evidences of ongoing evolution and of continuing genetic change. Such studies help explain the origin and persistence of genetically determined diseases and serve to explore their long-term influence. For races that have been long resident in their present locations, research sheds light on the long-term genetic effects of such environmental factors as temperature and climate, and of food type, source, and availability. For races that have been formed in the recent and historical past, the investigations can reveal the proportion of different groups that entered into genetic admixture. For particular groups with common racial ancestry that have moved from their original locations, evidences of genetic change also interest researchers. As people moved from equatorial Africa into the temperate zones, there was an expectable decrease in the frequency of genetic adaptations to heat and moisture. Changes in genetic composition may well be expected among people who have moved from northern Europe to the warmer areas of the American Southwest, Hawaii, New Zealand, and Australia.

RACES: SUBGROUPS WITHIN A SPECIES

All living mankind constitutes a single biological species (Homo *sapiens)* within a larger grouping or genus (Homo). Within the human species, Homo *sapiens,* a large number of populations may be differentiated genetically through readily observable characteristics (*e.g.*, skin, hair, and face and body proportions) and through less obvious but more distinctive biological traits, such as blood type. These biological groupings within species are commonly called races, in man as well as in other living forms.

Some biological descriptions refer to human "stocks," one intention for this being to avoid political overtones. Other writers have favoured the word division in lieu of race, again apparently to escape what may be perceived as offensive connotations. Other references to these human groupings include: "strains" (without implying the equivalent of purebred strains of laboratory animals); "varieties" (although the specific botanical meaning does not apply to human races as ordinarily constituted); or "ethnic group," sometimes limited to cultural or political groupings (*e.g.*, Macedonians, Croats, Magyars, or Slovenes) and at other times used with exactly the same biological meaning as race.

Reproductive isolation

Geographical races. Naturally occurring (*i.e.*, produced by natural, usually geographical separation of human groups) races of the human species are by no means identical in number of members or degree of genetic differentiation. There are small groupings of a few hundred to a few thousand individuals, some slightly and only recently isolated in the reproductive sense from adjacent people. Other equally small groups may have been mating apart from the rest of mankind for centuries, or even for thousands of years. Members of some human races number in the hundreds of millions (as the peoples of modern Europe) or in the thousands of millions (as in Asia). The present populations of Tristan da Cunha and Pitcairn and Norfolk islands, long-isolated and only recently rediscovered, were formed by mutinous European sailors and "natives" of adjacent areas. Other natural populations are descended from slaves who were isolated reproductively from their ancestral group and from the peoples with whom they came to live. There are people of mixed (hybrid) recent origin (as in the U.S. coastal Carolinas), contrasted with such populations as the Aboriginal Australians who show evidence of a 25,000-year genetic separation from the rest of mankind.

It is both useful and meaningful to identify the very large human groupings that often correspond to continents or other major geographic areas as "geographical races," a term extensively used with other life forms. Such geographical races are numerically large, containing within them smaller groups of reproductive isolates (breeding populations). The reasons for the large groups' geographic delineation are usually clear. The Indians of the Americas were reproductively separated from the peoples of other geographic regions for many thousands of years. Thus, they have come to differ genetically from the rest of mankind, and even from those Asians from whom they stemmed. The Aboriginal peoples of Australia similarly constitute a geographically defined group of local races (see below) separated from the rest of the world until little more than a century ago, except for slight contact in the Cape York region, Though the Hawaiians are presumably of remote geographic origin, their historical spread over the broad range of the Pacific Ocean is consistently supported by their legends, by their genealogical records, and by genetic studies conducted over an expanse of thousands of miles of islands.

All in all, there are not many geographical races. In Africa (with due regard to scientific disagreement about the degree of historical separation between the Bushmen and Hottentot of South Africa and the remaining peoples south of the Equator), there seem to have been at most only two ancient African geographical races; some say only one. There is reason to argue for a continuity between pre-Columbian America and Asia, whence the earliest Americans manifestly came, yet most modern racial classifications separate American Indians from Asians on the basis of firm genetic evidence. There are some groups (such as the Ainu people of northern Japan) of puzzling status; they, and such groups as some of the "pygmies" of the Philippines. specific native groups of Thailand, and the Miao people of China, may be valid geographical races (though each is numerically small). As currently viewed, a listing of nine or ten geographical races encompasses about 99 percent of mankind.

Local races. Except for people including the Ainu and the Miao (and perhaps such groups as the Bushmen) whose classification is not entirely resolved, most geographical races include numerous local breeding populations. While such local races may have few members, others number hundreds of millions. Local populations (local races, usually with distinctive languages) dotted pre-Columbian America; there were hundreds in North America alone. One need only list the names of the Algonkin and the Cocopa, the Blackfoot and Cree, the Salish and Mandan, the Penobscot and the Seminole, the Zuni and the Navajo, and the Hopi, the Papago, and the Pima to realize how many local races of American Indians there weie and in many cases continue to be.

Sizes of local breeding populations

In other parts of the world, where high population densities have not built up, local races are also apparent, as in many parts of Africa. History provides a picture of ancient local races in much of the European continent, with numerous local groupings in the British Isles cataloged and categorized by Julius Caesar. In the course of the last two millennia, however, local races in Europe seem to have become less easily defined, except along such broad lines as Northwest European, Alpine, and Eastern Mediterranean. Marked average differences in many hereditary traits over the regions from northern to southern Europe may be plotted like lines on a weather map, interrupted by enclaves of genetically distinct peoples. The Basques of Spain and the Lapps of Scandinavia represent old enclaves. Gypsies, who stem from India, are spread across Europe; their language is still of Indian origin. There are also groups of people of Middle Eastern origin in modern Europe. Such local races are the ones most easily identified and studied; they constitute population units that display continuing evolutionary change.

It is the local population or local race that either expands or contracts its numbers relative to the geographical race as a whole that most readily provides material for the study of racial origin and contacts, and of genetic intermixture (see GENETICS, HUMAN).

Microraces and smaller units. For much of the world, local races exist as reproductively isolated, culturally distinct, and genetically differentiated breeding populations. Indeed, barriers to gene flow (interbreeding) are prerequisites for the continuing existence of these populations. Culture and geography together tend to maintain such hereditary differences over time; thus, social or caste prohibitions on interbreeding serve to maintain local races in India. Elsewhere, religious regulations have maintained many local races for millennia. Rivers were efficient barriers to interbreeding in older days, and oceans remain major barriers to intermarriage. As populations have grown, however, as in Europe, geographical distance itself has become less of a barrier. Where once there was a 40-mile (64-kilometre) distance between hamlets, there are now often continuous zones of human habitation; with roads, bridges, canals, and trains, distance alone does not have the isolating influence it once had.

Yet sheer density of population is itself a restriction on gene flow and on the distance over which human mating is likely to occur. A boy is less likely to venture far afield for marriage when there are thousands of "girls next door" within the same apartment block in which he lives. Under such circumstances, the human male has but the same effective mating range as the mosquito. Thus, even in heavily settled locations, genetic differentiation continues in its course. In a long-established city such as Tokyo genetic differences occur from district to district; this also holds for Rome and London and every other large city studied. Some of these differences clearly represent older residence patterns of breeding that support the continuance of localized groups of people (microraces) long genetically distinct (as in the Limehouse district of London or in certain of the ghetto districts of Rome). In moving from one coast of England to the other, there are found systematic genetic differences, some related to earlier settlements and cultural patterns, still others apparently reflecting later differential directions of selection. The distribution of blood group O in England alone provides a fascinating pattern of such differences, as does the frequency of another blood group (Diego) in related American Indian groups, or the occurrence of hereditary sickle-cell anemia in parts of Africa. Such local patterns draw attention to microgeographical differentiation among many living organisms; thus, the term microgeographical races has been applied to fruit flies and simplified as microraces when speaking of man.

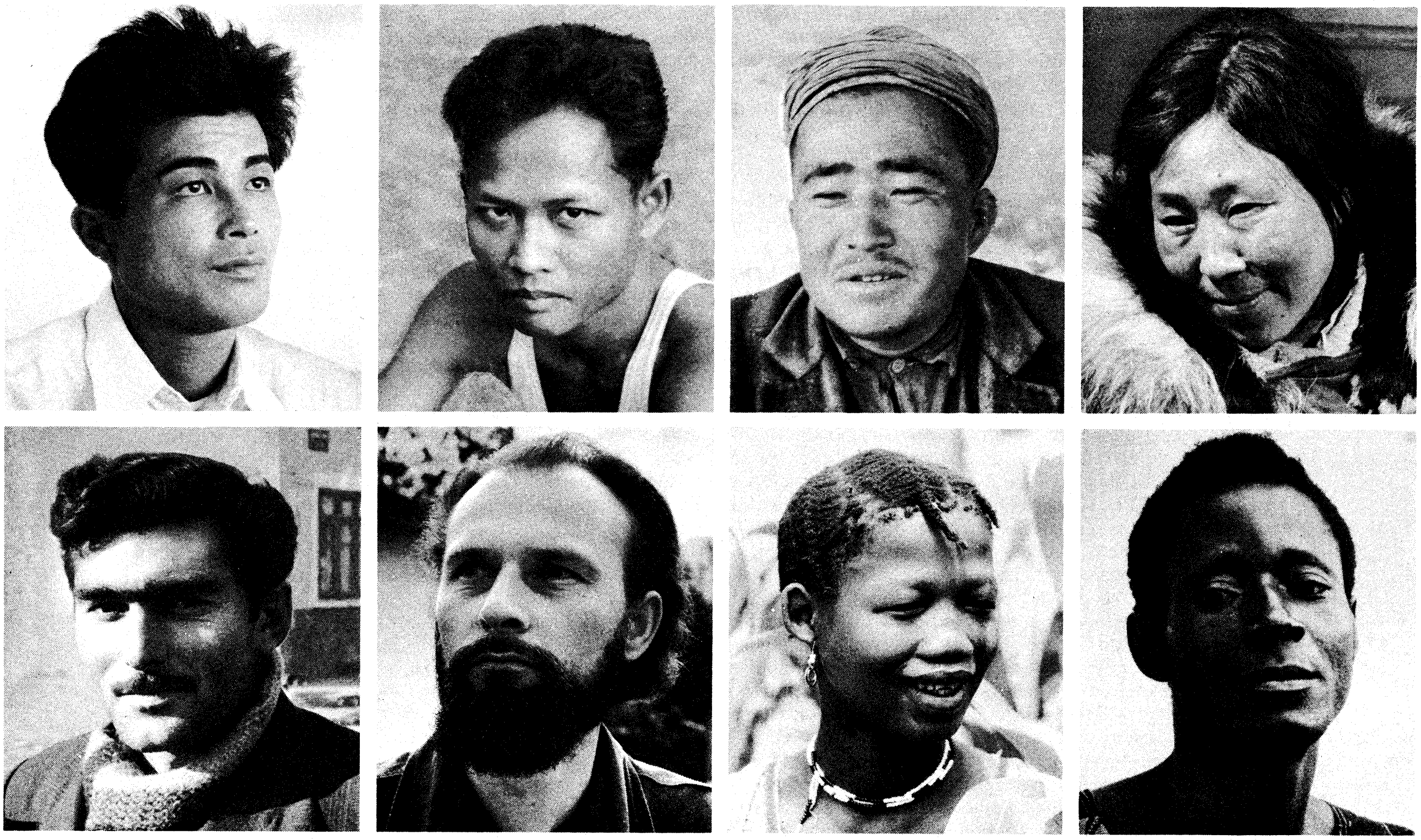
Causes of gent-flow restrictions

ASSOCIATIONS WITH RACE

Adaptations to environment. Local genetic adaptations may well be expected for races long occupying the same habitat, accustomed to a distinctive way of life, and used to a particular climate. Many studies demonstrate genetic heat adaptation in people living in the tropics and among those accustomed to the burning desert. Other studies reveal cold adaptation, as is found among the Eskimo, who have evolved a metabolic response for body heat maintenance under conditions of extreme cold. There is also evidence of selective genetic adaptation to moderate cold and night cold only, as among Australian Aboriginals and other desert groups. These people are exposed only occasionally during the day and commonly at night to temperatures not much below freezing. Since their food economy does not afford a caloric surplus for body heat maintenance, specific hereditary adaptations (*e.g.*, in body size) seem to have emerged by natural selection (see EVOLUTION).

One might expect more sweat glands per square unit of skin surface in peoples whose ancestors were long exposed to tropical heat, as an adaptation through genetic mechanisms that favour evaporative cooling of the body. By contrast, a minimum ratio of skin surface to body size is expected within breeding populations long exposed to Arctic cold; by minimizing surface (including the length of arms and legs), heat loss is reduced and more of the tissues are kept close to the waim body core. Evidence of just such genetic adaptations to cold, heat, humidity, and dryness have been found through systematic research. Careful study is required because, in addition to heredity, variable levels of nutrition may also affect body size or even fatness. Populations that are chronically exposed to specific diseases may evolve protective genetic adaptations; for example, the so-called sickle-cell trait is associated with resistance to malaria.

There is difficulty in sepaiating behavioral differences attributable to child-rearing practices and cultural expectations from those differences having a genetic basis, even within the same racial group (see HUMAN BEHAVIOUR, INNATE FACTORS IN: Genetically Mediated Group Differences). It is only when research turns from such behavioral attributes as intelligence or emotional stability to consider specific disorders of metabolism or neural development that clear differences between races on a genetic basis are found.

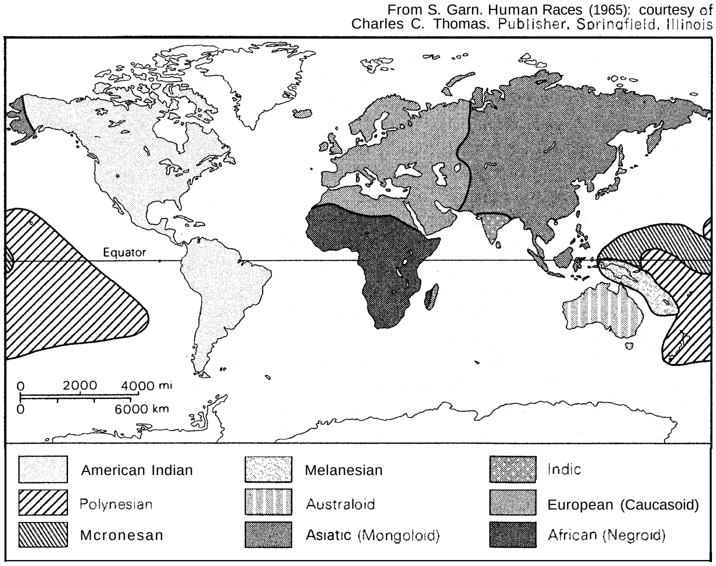


People of different geographicalraces.
(Top, left to right) Asiatics from Japan. Southeast Asia, and Mongolia; an Alaskan Eskimo. (Bottom, left to right) Examples of Mediterranean and of Northern European local races within the European geographical race; a Bushman-Hottentot; a West African, a local race of Africa.

(Top left: bottom, left and centre left) Photo Researchers, (top left and bottom left) George Daniell, (bottom. 'centre left) Fritz Henle. (too, centre left and centre right) Paul Almasy, (top right) Burt Glinn—Magnum, (bottom, centre right) Hubertus Kanus—Rapho Guillumette, (bottom right) Carl Frank

Language, culture, and behaviour. It used to be conjectured that languages were originally coloured by hereditary configurations of the mouth, teeth, and tongue of their earliest speakers. Research, however, has failed to relate the characteristics of any spoken language to the genetically transmitted facial configurations of its speakers. English, for instance, seems to be no better learned by children of English ancestry than by those of German, Italian, Egyptian, Ethiopian, or Pakistani derivation. Spoken language is learned rather than passed on through genes.

From S. Garn. Human Races (1965): courtesy of Charles C. Thomas. Publisher, Sprinqfield, Illinois



Historical locations of the major races of man.

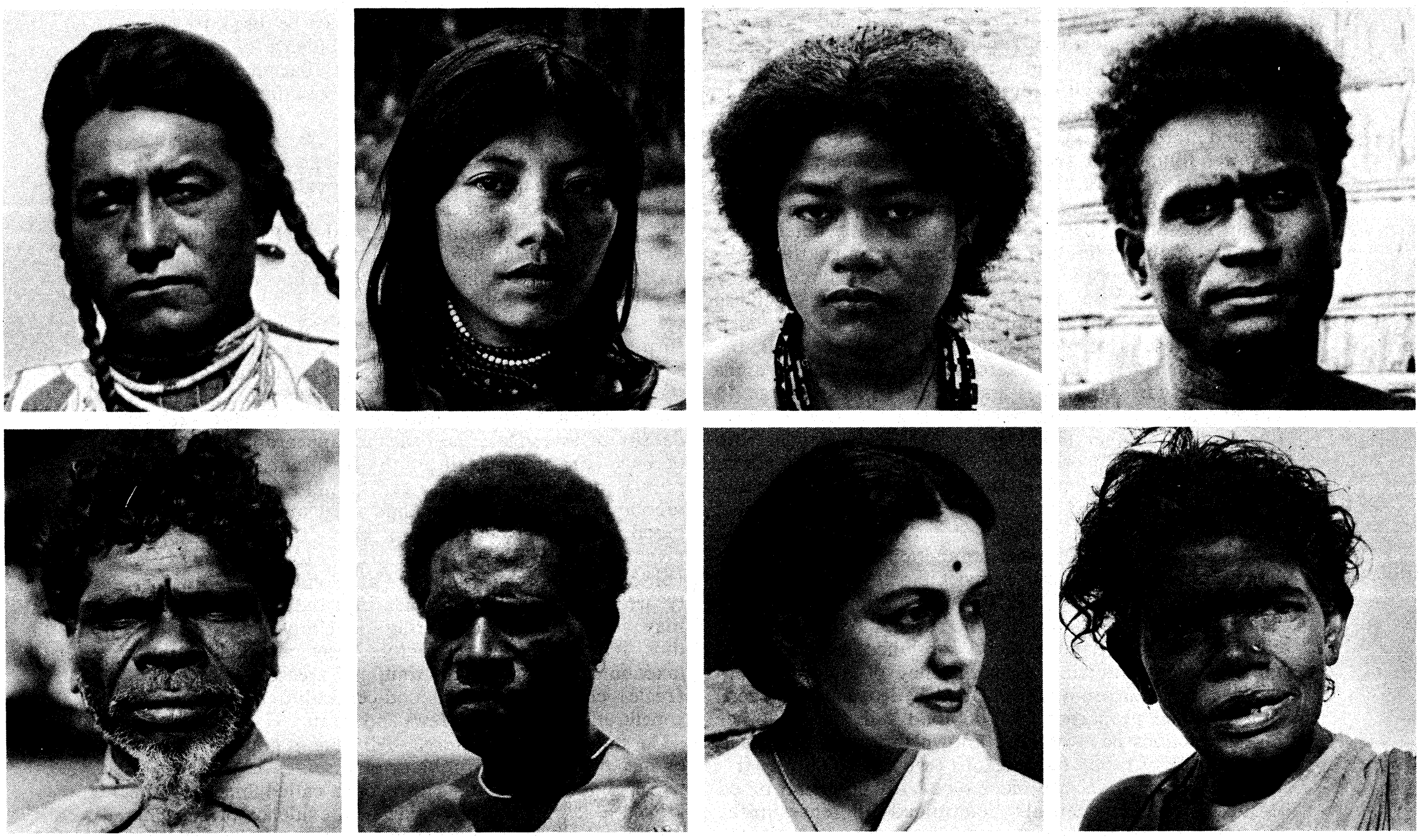
Gestures are typical of some groups, as may be seen in a Middle Eastern bazaar, but do not necessarily depend on a genetic basis. Postures are characteristic of some racial groups, and may have some basis in differences in bodily structure and functions, but they, too, are largely due to learning. Ways of expressing or suppressing emotion are learned in a cultural matrix, as American-born citizens of Japanese, or Italian origin discover when they visit the villages and cities from which their grandparents came.

Cultural versus genetic basis of behaviour

In theory, the distinctive characteristics of any culture could have been determined by the genetic makeup of its founders. Most people, however, are long removed from their ancestral cultures, which are themselves the products of social change. Once the most warlike people of Europe, the French expanded to northern and eastern Europe and as far south as Egypt. At one time the Spanish were widely feared for their military skills and threatened an invasion of England. The old Turkish Empire included part of Europe, the Middle East, and North Africa; and ancient Rome covered much of Europe, building still-famous baths in England (in the city of Bath) and farming North Africa. Yet, the imperial tendencies of all these cultures have now sharply waned.

In the face of such historical changes, it is difficult specifically to associate any of the races of man with enduring skills, abilities, gestures, or with warlike or pacifistic proclivities on any presumed genetic basis. Nor is it likely that cruelty, playfulness, mathematical or musical skill, engineering ability, or mercantile accomplishment have a genetic basis unique to any particular race of people.

Teeming and populous Tokyo now outdoes New York City, especially in subway rush hours and in the department stores on Sunday. Parisians now travel to London for less fettered play. In the United States Southwest a conservative voting record has replaced a once-adventurous pioneer spirit. Once relaxed Honolulu grows increas-



(Top, left to right) A North American Crow Indian of the Plains area; a Shipibo Indian from the Upper Amazon (both from the American Indian geographical race; a Polynesian of Samoa, Polynesian geographical race; and a Micronesian of the Solomon Islands, Micronesian geographical race. (Bottom, left to right) An Australian aboriginal, Australian geographical race; a Melanesian from New Guinea, Melanesian race: and two examples of local races within the Indian geographical race. Indic and Dravidian.

By courtesy of (top, left and centre right; bottom, left and centre left) the American Museum of Natural History, New York: photographs, (top, left centre and bottom right) Photo Researchers. (top, left centre) George Holton. (bottom right) Bradley Smith, (top right) Elliott Erwitt—Magnum, (bottom, centre right) Paul Almasy

ingly frenetic, and in Mexico City taxicab drivers simply do not behave as if they believe in the old ethos of *mañana.* With this kind of evidence it is difficult to support the notion that a characteristic way of life is simply an expression of the genetic makeup of the people.

OLD AND NEW "HALLMARKS" OF RACE

Old hallmarks of race. *Colour.* For centuries, geographical races and local races were identified by the most obvious physical differences, chief among them the colour of eyes, skin, and hair. From such observations came the simple notion of a few groupings based on the apparent colour of the skin alone: "black" men, "white" men, "yellow" men, and "red" men. Apparent skin colour is deceptive, however; it may be from dirt, grease, or mineral pigments (*e.g.*, ochre); it may be affected by exposure to sunlight and maintained with maximum tanning, as among the modern economic aristocracies of European origin on the beaches of Hawaii and North Africa. Moreover, all intrinsically dark skins do not necessarily have a common genetic origin; indeed, the people with the most skin pigment (melanin) in the unexposed areas of their skin come from quite unrelated parts of the world.

Suntanned "white" men

Hair form. In the 19th century, attempts were made to create racial classifications based on the form of the hair: straight, curly, woolly. To some extent these classifications have validity. Straight black hair of relatively wide diameter at the shaft is common among people of Asian origin, and the extreme of a spiral-tuft form of head hair is found in Africa south of the Equator, especially among Bushmen. Hair form, however, is affected by modes of hairdressing, and there is great individual variability in unmodified hair form within groups in many pads of the world. At any rate, no modern racial classifications are based on hair form done.

Anthropometry. The development of the technique of anthropometry (human bodily measurement) in the 19th century brought other approaches to the identification of races. Various groups of "pygmies" were identified by their extremely short stature. The proportions of leg and torso (trunk) was a distinguishing mark of short-legged groups such as the Eskimo and Southern Asians and long-legged peoples such as the Nilotes (a number of African tribes).

Head measurements (craniometry) showed distinctively round-headed (brachycephalic) populations like those of central Europe and long-headed (dolichocephalic) populations such as those typical of northern Europe. Cranial and facial bone measurements helped to distinguish differences in living people and in skeletons alike and clarified such features as the broad faces and jutting jaw angles of northern Asians and the broad-faced, narrower heads of some American Indian groups.

Body measurements do reflect a genetic component, but they only can be used with caution, since they are also affected by nutrition. Better nourished people tend to grow larger in many dimensions, because they have more fat and because proper diet enables them to build more muscle and bone. Most body dimensions are attributable both to the interaction of a number of genes and to the effects of nutrition and, thus, are used less than formerly in racial comparisons. In regions in which the nutrition of the population is relatively uniform, however, body measurement (anthropometry) can be useful in comparing genetically isolated subgroups.

Other features. Certain observations of eyelid and nose form, lip thickness, and even proportions of the fingers and toes are less subject to the objections on body dimensions. The form of the upper eyelid (the so-called mongoloid, or epicanthal, fold) often is diagnostic of Asian origins. Thickness and turning out (eversion) of the lips are notably great in some parts of Africa. Noses tend to be narrow in northern Europe and Tibet and wide in parts of Africa, New Guinea, and Fiji. But observations of characteristics with complex genetic bases have lost favour to simply inherited (single-gene) traits that can be studied more precisely.

The role of blood groups

New hallmarks of race. ***Blood traits.*** The classic human blood groups are examples of traits produced in the individual by the action of a single set of genes in the set of chromosomes (the majority of genetically determined features of the individual come about through the interaction of many gene sets). Because it is known how blood groups A, B, and O are inherited, the frequencies of their genes in any human population can be calculated from the results of blood tests. For blood group B, the gene frequency is close to zero among American Indians and Australian Aborigines and occurs in as many as 40 percent of the people in parts of Africa and Asia. Frequencies of another blood group, called M, range from zero to nearly 100 percent in different parts of the world; other blood groups have a tendency to show ranges nearly as wide.

Blood factors other than blood groups include abnormal red pigments, such as hemoglobin S, (the variant hemoglobin that is the cause of sickle-cell anemia), which is rare in most parts of the world but found in as many as half the local residents in some parts of Africa. Inherited blood factors include blood substances called haptoglobins and tranferrins and deficiency of a red-cell enzyme called glucose-6-phosphate dehydrogenase. Dozens of additional properties of red blood cells and of other blood fractions have become known through widespread medical use of blood transfusions (see BLOOD GROUPS).

Other genetic traits. Besides blood fractions and blood groups, there are genetically determined differences in the excretion of amino acids (protein components) via the urine; for example, there is a tendency to excrete high levels of β-amino-isobutyric acid among many people of Asian origin. There is a widespread, genetically determined deficiency in lactase (an enzyme that helps digest milk sugar), and there are simply inherited differences in earwax type: dry-flaky versus sticky-moist. With dozens of simply inherited traits available for study, far less reliance is being placed on once-popular body measurements, proportions, or descriptions of obvious physical differences.

METHODS OF RESEARCH

When a modern anthropologist or human geneticist studies a newly discovered subgroup of people or restudies one already known, he does not start from the very beginning. He is likely to have previously available statistical data that delineate the effective geographical and breeding limits of his group; he commonly knows their geographical race and, in most cases, their local race. He frequently has clues from earlier studies about what traits to study; for example, selected blood groups, abnormal hemoglobins, transferrins, haptoglobins, or β-amino-isobutyric acid, lactase deficiency, or earwax may already have been investigated.

Tracing evidence of admixture

In studying a group of people of largely African origin in Brazil, he may concentrate on a few blood groups and hemoglobin S. For a community of American Indians he may choose to study p-amino-isobutyric acid excretion and earwax. In New Guinea he may make dental casts (to study tooth size) and test for abnormal hemoglobins. Among so-called black people in the United States he may be interested in evidence of admixture with Europeans and American Indians; or he may want to relate an Indian tribe or "nation" to their origins in Mexico, the United States, or Canada. No group of human beings is now studied as if nothing were known about them.

Problems and approaches. In a group of related microraces or local races, the problem may be to investigate malarial selection, of particular social and economic concern in such places as the Middle East and Indonesia. The task may be to relate body size to climate or rainfall, to determine the frequency of resistance to smallpox infection, or to explain a high local occurrence of red–green colour blindness. Such research may be used to guide international relief organizations. Evidence of lactase deficiency, for instance, might reveal that members of a given microrace would have difficulty in digesting of milk products.

Some approaches to race research are purely classificatory (taxonomic), designed to place one group in relation to others; or the task may be historical, to trace origins. The goal may be to measure genetic change or to study disease susceptibility and other racial differences in terms of nutrition and climate. Each problem requires its own set of approaches, typically going well beyond the simple question of classification alone.

An anthropologist may be interested in the size and form of teeth; a geneticist may be concerned with the structure of different chromosomes; a human biologist may direct attention to the typical sequence of maturation within a group; a radiologist may measure skeletal density or investigate the order of appearance of centres of bone formation (ossification) in infants. Another approach may be to ascertain genetic influences on fat metabolism, perhaps to understand what makes fat-plugged blood vessels and deaths from heart disease so rare in a given microrace. The problem may be to study congenital (birth) defects such as cleft palate. While races ordinarily are studied with other ends in mind, the results still may add information bearing on simple classification; thus, research on blood factors casts new light on the taxonomic position of the Ainu, and measurements of Y-(male) chromosomes in blood-cell cultures clarified the genetic affinities between Jews and Arabs.

Identifying degree of admixture

The study of race mixture. All human populations are of mixed origin. The problem of research is to discover when and with whom. The people called black in the United States typically provide firm evidence of substantial admixture with European genes over the last two centuries; but information is lacking about their ancestral admixtures in Africa before American enslavement. Study of blood groups and other monogenic (single-gene) traits permits the estimation of the degree of admixture in any population of recent origin. People socially identified as blacks from Detroit, Oakland, Calif., or West Texas average close to 73 percent African genes and 27 percent European genes. For mixed populations of greater antiquity, such estimates are complicated by evolutionary changes that tend to obscure initial proportions.

People of English origin are provided with documented evidence of mixture with Vikings, Romans, occasional Spaniards, Irish, Scots, Welsh, and Bretons (among other Frenchmen). Among the "purest" (genetically isolated) North Americans, in a most restricted, recent sense, are those with ancestral Welsh surnames (Jones, Lewis, and Owens'), selected Pueblo Indian groups, and some more lately arrived Armenian communities.

Admixture with contiguous groups has been a human accomplishment through the ages—Israelites with Philistines, Egyptians with Sudanese, Aztecs with Toltecs, and Hindus with Dravidians. Mating with more remote groups was accelerated by such social movements as the romanization of North Africa, the Near East, and Europe; the spread of Bantu speakers southward in Africa; and the wide conquests of the Mongol troops of Genghis Khan. Through such avenues many Europeans came to incorporate African and Asian genetic elements; *e.g.*, the Mongol eye fold is discernible among Poles and Hungarians. With the development of sea transport, more far-flung intercontinental genetic mixing led to newer populations with immediate ancestors from even more geographical races; *e.g.*, the so-called triple hybrids (African, American Indian, and European) of the Carolinas in the United States.

It was once held that human "hybrids" or "half-breeds" such as Eurasians or Afroindians would be biologically inferior to the parental stocks. Contemporary studies of such groups indicate no systematic decreases in fertility or longevity; neither do they fully confirm contrary claims of "hybrid vigour." Admixture nevertheless does tend to dilute the pool of deleterious genes and favours the production of new genetic combinations upon which natural selection operates. The effect is to increase genetic variability and the rate of human evolutionary change.

MODERN MEASURES OF RACE

Little-used particular traits

Measures now used in the study of races are primarily those of factors that are genetically and simply determined, such as single-gene traits. While genetically complex differences as in the size and shape of the teeth can also be measured, their utility is limited by ignorance of their exact mode of inheritance. Thus, wisdom teeth never develop (third molar agenesis) among an exceptionally high proportion of Eskimo and in many Asians. While, by contrast, third molar agenesis is exceptionally rare in Australia, New Guinea, and in some parts of Africa. This genetically complex trait, however, is relatively little used in making racial comparisons. There are simply inherited skeletal traits, such as presence of a broad middle segment of the fifth, or little, finger (brachymesophalangia-5), that are common in some populations and extremely rare in others. Fusion of two of the wristbones occurs in up to 6 percent of individuals in some parts of Africa but in only the merest fraction of the populations of Europe or Asia. Such skeletal differences readily can be studied with X-rays but are not so easily detected in field studies through ordinary visual inspection. Significant racial differences in the size and thickness of the walls of longer bones also require X-rays for their measurement.

Single-gene traits. The greatest amount of information on simply inherited traits bearing on race has come from the study of the blood and from biochemical analysis of the urine. While the components of the blood are no more informative than are those of muscle or viscera, many millions of people throughout the world have had their blood typed, producing an accumulation of information on blood differences, their inheritance, and their geographic distribution. In similar fashion, modern medical interest in metabolic abnormalities ("biochemical lesions"), often diagnosed by measuring amounts of different amino acids in the urine, has revealed various biochemical polymorphisms (a variety of forms or types) that characterize families of close relatives, microraces, local races, and, in some cases, geographical races as well.

Some findings about race have been by-products of other kinds of study. It was discovered almost by accident that a chemical called phenylthiocarbamide tastes bitter to some people but seems tasteless to others. Only later did physical anthropologists and human geneticists discover that some racial groups are remarkably "taste-blind" to phenylthiocarbamide. Population differences in colour blindness were discovered in the course of routine, large-scale visual examinations for military purposes; ability to distinguish red from green can be urgently important to an infantryman, a submarine commander, and the driver of a military truck. Earwax polymorphism (dry-flaky and moist-sticky) was long known in Japan; this observation led European and African geneticists, who had assumed all earwax was sticky, to discover the polymorphism in their own lands.

Medical Disorders. Initial medical concern with abnormal hemoglobins led to the observation that sickle-cell anemia appears to be limited to people of African origin and that a disorder known as thalassemia (literally "sea blood") is apparently distinctive of people with Mediterranean ancestry. Discovery of genetically determined deficiency of the enzyme glucose-6-phosphate dehydrogenase (G6PD, noted above) came about when new antimalarial drugs were introduced after World War II and occasionally caused dangerous side effects in persons of African and Middle Eastern origin. Ancient Egyptians and Greeks knew about some people who got sick when they ate common broad beans (fava beans). The geographical limits of this genetically determined disorder (favism) seem clear, but the details of its inheritance have not yet been unravelled. It appears to be related to G6PD deficiency disease.

Disorders of genetic origin

Medical disorders of genetic origin are of importance in making racial comparisons only when they are common enough to permit comparison of one group with another. Geographical sleuthing has led to the realization that relatively high gene frequencies for the sickle-cell trait (because of hemoglobin S), various thalassemias, G6PD deficiency, and, apparently, favism, all tend to be found in places in which there is a particular kind of malaria (falciparum malaria). The abnormal hemoglobin S or the G6PD red-cell enzyme deficiency apparently makes for red blood cells that are relatively unsuitable for the reproduction of the malarial parasite. Thus, the individual can be said to gain protection against malaria at the cost of suffering another disorder. Microraces with high gene frequencies for disorders that are rare elsewhere may be indicative of local disease adaptation, perhaps fighting an infectious disease with one of genetic origin.

Blood groups and subtypes. Within the ABO blood-classification system, B is the least common of the three blood groups the world around. Absent among genetically isolated American Indians and lacking in isolated Australian Aborigines, B is relatively common in eastern Asia, India, and Africa; it is only a minor blood group in Europe, the observed frequency rarely exceeding 12 percent. Apparently, Australia became isolated reproductively from the rest of the world before B became frequent elsewhere. Most investigators assume either that the east Asian ancestors of American Indians crossed into the North American Arctic before blood group B had become established in Asia or that, having had it once, they lost blood group B in the process of evolution.

So-called Rh blood subtypes, first discovered in rhesus monkeys and later in people, leaped to medical importance in Europe when incompatibility between the Rh types of the mother and her unborn baby was found to damage or kill the fetus. The troublesome Rh-negative gene (r) is relatively common in Europe (especially among the Basques of Spain) but less frequent elsewhere; thus, its frequency can be used as one measure of European admixture in populations that once lacked it. Rare or absent in the rest of the world, R_o (another distinctive gene in the Rh series) is relatively common in Africa. Thus, the occurrence of R_o can provide a measure of the gene flow into the mixed population of Brazil or out from the "black" people of the port of Puerto Barrios in Guatemala. By its relative frequency, R_o serves as one measure of "white" admixture among American "black" groups.

M–N series in racial comparisons

Other blood groups often used in racial comparisons include the M–N series. Since the frequencies of M and N are about equal in most parts of Europe, they do little to characterize local differences there. Japanese and Britons both have M and N, so the frequencies also fail to distinguish adequately between the English village of Stoke Poges and the city of Kyoto, Japan. Since M is virtually missing in Aboriginal Australia and rare in the other lands of the Pacific Ocean, and since N is absent or nearly so in the Indians of the Americas, M and N have great value in comparing Pacific populations, in postulating their origins, and in assessing European or African admixture in various American groups.

Within the Diego blood group, the Diego positive gene (Di") is especially common among people of Asian origin. It also has been detected among people of Polish origin in the United States, perhaps as a residue of the Mongol invasions of Europe. Yet Di^a cannot be used to compare people within groups that lack the gene, nor is the question of origins easily resolved simply by comparing those who have it in the highest frequency. Not every genetically determined set of traits has equal value in comparing, differentiating, or searching for the origins of each race. Great differences within the same geographical race point to local selective factors and considerable blood-group diversity. Even within small and related villages, investigators do not use the same list of genetically determined traits for every set of people studied.

Abnormal hemoglobins. When present on only one of a person's chromosomes, the gene that generates abnormal hemoglobin S merely leads to sickle-shaped red blood cells (sickle trait) without making him ill. When the individual inherits a double dose (one gene on each of a pair of chromosomes), he rarely reaches reproductive age, dying first of severe sickle-cell disease (anemia).

As first detected, the sickling gene occurred in people from parts of Africa where methods of agriculture led to stagnant, standing water that gave malarial mosquitoes a chance to breed heavily. Thus, among people of the Western world the frequency of hemoglobin S indicates African origins, while the relative frequency of S in different African populations provides estimates of past malarial selection. The sickle-cell gene in modern, slave-owning Saudi Arabia presumably was introduced long ago by slaves of African origin but seems to have been maintained in oasis villages there by malarial selection. The gene is also known from Indonesia, where malaria is widespread; apparently, sickling is maintained and even increases wherever malarial selection is active.

The pattern for the thalassemias

Much the same holds for thalassemia minor (a single dose of the offending gene) and thalassemia major (a double dose). In the thalassemias the abnormal hemoglobin is a poor carrier of oxygen; the sufferer needs blood transfusions from time to time if he is to grow and live effectively. But the disorder is not without compensations; the oxygen-deficient blood cells provide the malaria parasite a relatively poor home, one that may be less hospitable than the sickle cells. Once thought to be restricted to the Middle East, thalassemia is found in a broad band aroucd the tropics, where malaria has been a problem for years. Absence of the abnormal thalassemia hemoglobins in some populations that were originally Mediterranean has been attributed to generations of genetic adaptation to nonmalarial regions. Jews of western Europe (Ashkenazim) seem to have lost whatever thalassemic genes their ancestors may have had 2,000 years ago. Spanish-Portuguese-North African Jews (Sephardim) in the United States appear to have lost the thalassemic gene a few hundred years after their forefathers came to North America. Thus, the abnormal hemoglobins provide evidence of racial origins that supports the record of history.

In similar fashion, the frequency of G6PD enzyme deficiency repeats the story. When one lacks this red-cell enzyme, the malarial parasite finds the blood cells poor places for development. People from Middle Eastern regions where malaria is prevalent have the highest frequency of G6PD deficiency; in some Kurdistani villages the gene frequency rises to a phenomenal 60 percent. In the United States the deficiency characterizes people who come from malarial areas of that country or who are relatively recent arrivals from the Middle East and North Africa. People whose ancestors left those Mediterranean regions thousands of years ago to live free of malaria no longer have the abnormal gene.

Since gene frequencies change with time, the presence or absence of a gene does not always confirm or disprove ancient relationships. If hemoglobin S can become frequent in Africa after the introduction of agricultural methods that favour mosquitoes, and if G6PD deficiency can disappear in people of Mediterranean origin, then such problems as posed by the absence of blood group B in American Indians seem less puzzling. It is likely that ancestral Indians came to the Americas with blood group B and then subsequent generations lost it, still retaining the Diego blood group, the high level of β-amino-isobutyric acid excretion, and many other traits found among modern Asians.

Metabolic factors. While some people excrete much β-amino-isobutyric acid (BAIB), the urine of most people shows very little of this substance even when all share the same diet. Genetically determined (BAIB) polymorphism and any metabolic advantages it may provide are incompletely understood. Nevertheless, people who do show high levels of BAIB excretion are Asians and American Indians. If the trait is assumed to have come from Asia, it is an unexpected finding that many Indian groups living nearest to Asia (theoretically of more recent arrival) seem to excrete less BAIB than those living in South America.

Lactase deficiency and its world distribution

Lactase deficiency is even more common in the Orient than is the BAIB trait. After drinking about eight ounces of cow's milk, people with lactase deficiency are likely to suffer abdominal cramps and diarrhea. This disorder of lactose metabolism is also very common among individuals of African origin and, indeed, in much of the world except Europe. The European geographical race is statistically unusual in tending to maintain high intestinal lactase levels into adulthood. That race has the highest proportion of people who can drink milk in great quantities without distress even when they are grown up.

Other diseases and congenital defects. Through continuing natural selection (favouring some genes and eliminating yet others), the frequency of genetically related diseases differs from race to race. A form of mental deficiency that arises from an inherited enzyme disorder (phenylketonuria) is far less common in Africa and Asia, for instance, than it is in Europe. Jews from near Vilnius, U.S.S.R., tend to a relatively high frequency of Tay-Sachs disease (progressive blindness and mental defect in infants); people from England are particularly prone to develop a bone disorder called Paget's disease; and congenital hip dislocations are especially common in American Indians from the Southwest.

Certain types of cleft palate are remarkably common among Japanese, while abnormal narrowing of the body's largest artery (the aorta) is far less frequent in Americans of African origin than in their fellow citizens of other origins. Minor wristbone (carpal) fusions are 15 times more common in people of African ancestry than in Mexican-Americans, who in turn show broad middle segments of the fifth finger 10 times more frequently than Afro-Americans. Women of African origin are comparatively unlikely to suffer the spontaneous fractures and low-back pains that come from adult bone losses (osteoporosis); yet men of African origin who are military pilots are unusually prone to suffer ejection-seat fractures of the lower lumbar vertebrae. Skin cancer is far more frequent among light-skinned, freckled Texans than it is in East Texans of African origin. Painfully impacted wisdom teeth (third molars) are most common among Asians and Europeans who are genetically delayed in third-molar development.

Symptoms of genetically related disorders often are manifested onlv under direct environmental influences. The headaches of what has come to be called the "Chinese-restaurant syndrome" are found to appear among people who have inherited a tendency to react deleteriously to monosodium glutamate, a flavour-enhancing product used by many Chinese cooks. An inherited inability to tolerate milk sugar usually is of little consequence unless the person drinks more than half a pint of milk at a time. Wheat-gluten sensitivity may produce no distress among people who subsist on rice or barley, but in Italy it may cause major discomfort and debilitation because the diet there includes pasta products (*e.g.*, macaroni), which are made of high-gluten wheat. Even a genetically related disease such as diabetes may diminish as a medical problem among populations that eat little carbohydrate (sugar, starch) and take most of their calories from protein foods such as meat and fish. Clearly, many genetically determined differences between races are "diseases" only under specific circumstances.

THE EVER-CHANGING RACES

Ongoing evolution. Human races drift genetically, resembling the shifting sands of a desert. They all are in a continuing state of change; new genes are readily introduced from the outside, since all races of *Homo sapiens* are interfertile. Explorers, military men, and others contribute their genes to populations they visit; conquerors still bring home the new genes of their captives or war brides. Maya Indian genes were introduced into Spain and other American Indian genes to England; genes from Japan, Korea, Vietnam, and Germany have been mixing on a large scale with those from the United States since World War II.

The role of explorers and conquerors

In a relatively short time, chance events (*e.g.*, in the selection of mates) can serve to build up the frequency of rare genes in particular populations, as among Polish Jews from the vicinity of Vilnius or French Canadians in the St. Lawrence Seaway region. Apparently lost by Ashkenazic Jews of Europe, the gene for G6PD deficiency has gained at a phenomenal rate among Jews from Iran and Kurdistan. The genes for a disorder called familial Mediterranean fever are so restricted to the eastern Mediterranean that their changes document local evolutionary change. Essentially, evolution is a change in gene frequency.

The distribution of the sickle-cell gene attests to recent changes in gene frequency; for example, there is evidence of its gradual elimination in North America over the last 150 years. Studies of religious isolates such as the Mennonites (*e.g.*, Amish and Hutterites) reveal differences that reflect local ongoing evolution. American Indians of the Southwest evidence many genetically related disorders that must be comparatively recent in their present gene frequencies.

Within a small group of a thousand or so mating pairs, genetic differences may be the simple result of chance sampling effects on pairing (*i.e.*, genetic drift), the group perhaps even arising as a distinctive race on that basis. Some individual families do especially well reproductively and have hundreds of offspring in a matter of generations. Other families tend to die out; for example, compare the relatively small number of people named Holmes today with those called Adams. In some societies polygamy may raise rare genes to numerical importance; men who live on to father several generations of progeny affect small populations by the sheer numbers of descendants who bear their genes.

The ingredients of population evolution

Ongoing evolution at the population level thus is a product of chance, geography, and natural selection. It also is affected by religion and local attitudes toward reproduction; *e.g.*, depending on whether such attitudes support a norm of large or small families. Between populations, the rate of human evolution may be a question of technological advancement, of the availability of food, or of major medical advances that reduce the number of deaths. As countries develop improved food technology, more effective food distribution, immunization, and other medical care, their populations tend to increase, thereby changing the genetic makeup of the world's population.

The future of races. In the last 200 years, many small and some once-large racial groupings have disappeared as distinct populations. The once-numerous Tasmanians are gone but not quite forgotten; the last officially recorded "full-blooded" Tasmanian died in 1876. Aleuts, natives of the Aleutian Islands, have declined to less than 8 percent of their former numbers. The Carib Indians who met Columbus are gone entirely (as a distinct racial group) from Caribbean lands. Indians of New York state are memorialized more by place names than by living people. Where now are the Picts and the Celts; or the Germanic tribes that Caesar cataloged; or the people of the Greek city-states; and where are the traditional ten tribes of Israel?

Many of these groups live on only in the sense that their distinctive genes were added to and survive among the larger groups that eventually encompassed them. Penobscot Indian genes are scattered throughout Maine, and those of ancient Norse invaders are spread over the map of England. In parts of New Hampshire, Hessian genes from the German mercenaries of the American Revolution live on; and in Rhode Island, some French genes still survive from Count de Rochambeau's troops who stayed there for about a year before they helped defeat the British at Yorktown. Other microraces and local races exist today as part of newly formed hybrid groups.

New races emerge, as in Hawaii and in South Africa; in South America the populations include varying proportions of genes from that continent, from southern and northern Europe, from Africa, and (in some regions) from India, too. Gene flow in the United States once was largely from the European population into that of African origin, up to 25 percent in American "blacks" in the North and West. More recently, gene flow in the other direction has also taken place, as individuals have "crossed over," or as African genes have entered the nominally European populations in the United States and Canada from Cuba, Puerto Rico, and Jamaica. Gene flow from the United States into Japan since 1946 has been countered by gene flow from Japan; thus, in the 1970s there were nearly a quarter of a million "hybrid" children in the U.S. whose mothers came from Japan, China, Korea, Taiwan, and Vietnam. Groups without some infusion of genes from half a world away are few indeed.

Continuous gene flow

Some gene pools have been quite eradicated from the human species, while other gene pools live on in newly named populations. The genetic makeup in all developed nations is changing as the result of the breaking up of isolates, of incorporation of new genes, and of new directions in natural selection. Genes from Pakistan and India are now spreading across the English countryside. Some long-isolated gene pools from North Africa and the Yemen are now becoming part of the Israeli genetic complex.

New races are forming as the results of history, geography, and political alliances, and as the results of bridges and oil companies and mining camps and military activities. The genetic makeup of existing races is changing because of improved medical care, ditch drainage, and the elimination of insects that carry disease. The same processes that created the Pitcairn Islanders in the 1790s continue at work, as jet transports erase some barriers to gene flow and as political units build still others. The races of the next century will include some that do not exist now, while some of the local races and microraces of the 1970s will be missing.

RACE AND SOCIETY

Race, intelligence, and behaviour. There is some small evidence that infants from different races do differ in reactivity, rates of skeletal maturation, and optimum birthweight. Beyond this, the possible existence of true racial differences in behaviour and intelligence becomes difficult to establish. Peoples long characterized as "stolid," "unimaginative," "childish," or "cruel" tend to become very different when they emigrate, when their economic status changes, or when they live in rapidly changing environments. Once misjudged by some Westerners to be innate imitators of other cultures, the Japanese nation has shown technological and scientific progress to such a creative degree as to disprove this belief. Pacifistic European Jews have become tough warriors since they took root in modern Israel; the once-warlike Navajo Indians have been peaceful since the 1880s, and the sexually repressed England of the 1930s has become more and more uninhibited since 1950.

The problem of race and intelligence is extraordinarily complicated simply because "intelligence" as conventionally measured is not a stable trait in the way blood group B, blue eyes, or extra cusps on teeth are. Measured intelligence changes with emotional state, with parental encouragement, with motivation, and with disillusion. If an intelligence test is constructed for use with city dwellers, then rural children may tend to fare poorly with it. If Aboriginal boys and girls in primitive Australian surroundings are tested in ways that demand literacy, knowledge of science, and a large vocabulary, they fail to do well regardless of their talent.

Social deprivation and malnutrition

Childhood social deprivation and juvenile malnutrition may also affect intelligence test scores; the test taker is also penalized if he comes from a cultural tradition that takes a dim view of competition. For these and other reasons, the problem of race and intelligence remains unsolved. While so-called racial groups may differ markedly in measured test performance, it is not known how the differences reflect variations in prenatal care, social experience, family structure, and way of life (see INTELLIGENCE, DISTRIBUTION OF: *Group differences* in intelligence).

Racism. In the record of human history, races (however defined) generally have met in conflict, in the course of empire building, and in forays for captives or slaves. Victor and vanquished built and nourished mutual dislikes. In the United States, immigrants imported as cheap labour were initially disliked as economic rivals by earlier residents and, in turn, came to resent later migrants. Chinese and Japanese labourers and their children suffered especially in the U.S. West; on the East Coast newly arrived Irish families were perceived as a threat to the wage structure in mills and shops. Africans, first imported into the United States as slaves, early showed such spirit and talent that alarmed slaveholders sought theoretical justification for the continuance of slavery, establishing laws against slave education, even forbidding mere literacy to Africans. Italians moved to the United States as labourers, and Jews from central Europe and Russia went there to face old, transplanted European prejudices. After World War II, the arrival of Pakistanis and West Indian nationals in England recapitulated problems earlier encountered by many groups in the United States. By the 1970s restrictive laws against immigration had been passed in England.

Racism in its simplest form is dislike of any people lumped together as a microrace, local race, or geographical race. It results in attempts to limit economic opportunity, to preserve status, to deny equal protection under law, and to maintain cheap labour. Racism is most likely to arise between groups that show obvious physical or linguistic differences or who are distinguished by various levels of education. Reverse racism occurs when members of previously underprivileged races achieve political and economic strength and proceed to discriminate against others.

After World War II and the American occupation of Japan, negative sentiments in the United States abated to the extent that Japanese-Americans became highly respected citizens in the very regions where they were earlier most despied. Puerto Ricans became more generally accepted in that country over the last generation or so. Modem Israeli military successes have stimulated respect for Jews by other Europeans, but rising anti-Jewish feelings have been reported among Americans of African origin. Recently, some Eskimos have balked at hiring technical advisers of European ancestry, tribal Navajo have condemned "Anglo" advisers of European ancestry, and Latin Americans of mixed origin (Chicanos) have refused to deal with agricultural experts of so-called "white" cultural background.

Racism tends to persist most readily when there are obvious physical differences among groups; *e.g.*, "white"–"black" differences in Australia and South Africa and the distinctions in bodily traits between Hindus and other residents in Guinea and the Caribbean region. Racism based on differences in language and behaviour tends to disappear more readily as cultural communality is attained; as later immigrations threaten older, established groups, however, they may provide new directions for fear and dislike.

When changing sentiments alter directions of dislike, those previously discriminated against may newly become racist if their social and economic gains are threatened, or if their group problems tempt them to seek easily identified "enemies" as a substitute for working toward social betterment.

BIBLIOGRAPHY. S.M. GARN, *Human Races,* 3rd ed. (1971); and (ed.), *Readings on Race,* 2nd ed. (1968), two standard texts on the phenomenon of race in man; *Culture and Direction of Human Evolution,* also ed. by Garn (1964), a set of readings derived from an American Association for the Advancement of Science symposium; PAN AMERICAN HEALTH ORGANIZATION, ADVISORY COMMITTEE ON MEDICAL RESEARCH, *Biomedical Challenges Presented by the American Indian* (1968), a discussion of genetic and environmental health problems of the American Indian; F.M. SALZANO (ed.), *The Ongoing Evolution of Latin American Populations* (1971), studies on evolution in contemporary man; R.A. GOLDSBY, *Race and Races* (1971), a recent book on genetically-different human populations by a professional biologist.

(S.M.G.)

Rachmaninoff, Sergey

Sergey Vasilyevich Rachmaninoff was the most popular Russian composer of the 20th century, a gifted conductor, and possibly the most formidable pianist of his time. He was the last great proponent of the tradition of Russian Romanticism.

Early life. Rachmaninoff was born on April 2, 1873, at Oneg, an estate belonging to his grandparents, near Lake Ilmen in the Novgorod district. His father was a retired army officer and his mother the daughter of a general. As a matter of family tradition, he was destined to become an army officer until his father lost the entire family fortune through risky financial ventures and deserted the family. Sensing his abilities, young Sergey's cousin Aleksandr Siloti, a well-known concert pianist and conductor, suggested sending him to the noted teacher and pianist Nikolay Zverev in Moscow for his piano studies. It is to Zverev's strict disciplinarian treatment of the boy that musical history owes one of the great piano virtuosos of this century. For his general education and theoretical subjects in music, Sergey became a pupil at the Moscow Conservatory.

Rachmaninoff.
Bassano and Vandyk, Elliott and Fry

At the age of 19 he was graduated from the conservatory, winning a gold medal for his one-act opera *Aleko* (after Pushkin's poem "The Gypsies"). But his fame and popularity, both as composer and concert pianist, were launched by two compositions: the *Prelude in C Sharp Minor,* played for the first time in public on Sept. 26, 1892, and his *Concerto No.* 2 *in C Minor,* which had its first performance in Moscow on Oct. 27, 1901. The former piece, although it first brought Rachmaninoff to public attention, was to haunt him throughout his life: the prelude was constantly requested by his concert audiences. The concerto, his first major success, revived his hopes after a trying period of inactivity.

Failure of First Symphony

In his youth, Rachmaninoff's passionate nature was not sustained by the will and equilibrium he later developed, and he was subject to emotional crises over the success or failure of his works as well as in his personal relationships. Self-doubt and uncertainty carried him into deep depressions, one of the most severe of which followed the failure, on its first performance (March 1897), of his *Symphony No. 1 in D Minor*. The symphony was poorly performed, the critics condemned it. (Ironically, this was the work that, following Rachmaninoff's death, was acclaimed by many musicologists as his greatest contribution to symphonic literature as well as his most original composition.) During this period, also brooding over an unhappy love affair, he was taken to a psychiatrist, Nikolay Dahl, who is often credited with having restored the young composer's self-confidence, thus enabling him to write the *Second Piano Concerto* (which is dedicated to Dahl).

The association with Dahl seemed to have an even greater influence on Rachmaninoff's personal life than on his music: about this time he married his cousin Natalie Satin. The writing of the *Second Piano Concerto* marked the resumption of his creative activity, and it was soon followed by a number of shorter works. The concerto is studded with such melodic themes that it has become his most popular longer work.

Major creative activity. At the time of the Revolution of 1905, Rachmaninoff was a conductor at the Bolshoi Theatre. Although more of an observer than a person politically involved in the revolution, he went with his family, in November 1906, to live in Dresden. There he wrote three of his major scores: the *Symphony No. 2 in E Minor* (1907), the symphonic poem *The Isle of the Dead* (1909), and the *Concerto No. 3 in D Minor* (1909). The last was composed especially for his first concert tour of the United States, highlighting his much-acclaimed pianistic debut on Nov. 28, 1909, with the New York Symphony under Walter Damrosch. Probably the composer's best unified longer work, the *Third Piano Concerto* requires great virtuosity from the pianist; its last movement is a bravura section as dazzling as any in the literature. In Philadelphia and Chicago he appeared with equal success in the role of conductor, interpreting his own newest symphonic compositions. Of these, the *Second Symphony* is the most significant: although it displays Rachmaninoff's usual propensity for lapsing into familiar romantic conventions, it is a work of deep emotion and haunting thematic material. While touring, he was invited to become permanent conductor of the Boston Symphony, but he declined the offer and returned to Russia in February 1910.

Place in
Russian
music

It was at that time and during musical feuds in Moscow between several divisions in the large family of Russian composers that Rachmaninoff's compositions were clearly classified and his place in Russian music defined. On the one hand there were the adherents to the St. Petersburg group of the "Mighty Five" (Balakirev, Borodin, Cui, Mussorgsky, and Rimsky-Korsakov), and on the other there were the more conservative followers of Tchaikovsky, Anton Rubinstein, and Taneyev. Another, smaller, group was composed of enthusiasts of Alexander Scriabin's music. Of these three factions, Rachmaninoff belonged unmistakably to the Tchaikovsky group. His lyricism, devoid of any particular innovation, is especially evident in the large number of songs he composed, even more than in his piano compositions.

Rachmaninoff's music, although mostly written in the 20th century, remains firmly entrenched in the 19th-century musical idiom. He was, in effect, the final expression of the tradition embodied by Tchaikovsky—a melodist of Romantic dimensions still writing in an era of explosive change and experimentation.

The one notable composition of Rachmaninoff's second period of residence in Moscow was his choral symphony *The Bells,* based on Konstantin Dmitriyevich Balmont's Russian translation of Poe's poem. Although it became a staple of the symphonic repertory, partly because of its scoring for chorus and soloists, the work displays considerable ingenuity in the coupling of choral and orchestral resources to produce striking imitative and textural effects.

Exile in
the U.S.

After the Revolution of 1917, Rachmaninoff went into his second self-imposed exile—this time taking his family to the U.S., where he made his home for the rest of his life. For the next 25 years he lived in an English-speaking country, yet never mastered its language nor thoroughly acclimatized himself. With his family and a small circle of friends, he lived a rather isolated life. He missed Russia and the Russian people—the sounding board for his music, as he said. And this alienation had a devastating effect on his formerly prolific creative ability. He produced little of real originality, but rewrote some of his earlier work. Indeed, he devoted himself almost entirely to concertizing the U.S. and Europe, a field in which he had few peers. His only substantial works from this period are the *Symphony No. 3 in A Minor* (1936), another expression of sombre, Slavic melancholy, and the *Rhapsody on a Theme by Paganini* for piano and orchestra (1934), a set of variations on a Paganini violin caprice. The *Rhapsody* has vied with the *Second Concerto* as the composer's most often-played work.

In the midst of one of his concert tours, Rachmaninoff was taken to his home in Beverly Hills, California, where he died on March 28,1943.

Of his compositions, the *Second* and *Third Piano Concerti,* as well as the *Rhapsody,* still remain in the concert repertory and may continue to do so. His symphonies, his vocal works, and his solo piano pieces have declined somewhat in their appeal, yet the general charm of his work and his unique performances retain for him an honourable position in the history of music.

MAJOR WORKS

Orchestral works

SYMPHONIES: *No. 1 in D Minor,* op. 13, 1895; *No. 2 in E Minor,* op. 27, 1907; *No. 3 in A Minor,* op. 44, 1936.

PIANO CONCERTOS: *No. 1 in F Sharp Minor,* op. 1, 1890–91, rev. 1917; *No. 2 in C Minor,* op. 18, 1901; *No. 3 in D Minor,* op. 30, 1909; *No. 4 in G Minor,* op. 40, 1927.

MISCELLANEOUS: *The Rock,* op. 7, 1893 (for orchestra); *Capriccio on Gypsy Themes,* op. 12, 1894 (for orchestra); *The Island of Death,* op. 29, 1909 (symphonic poem based on picture of Böcklin); *Rhapsody on a Theme of Paganini,* op. 43, 1934 (for piano and orchestra); *Symphonic Dances,* op. 45, 1941 (for full orchestra).

Chamber music

Trio Elegiaque in D Minor, op. 9, 1893 (for piano, violin, and cello); *Sonata for Cello and Piano in C Minor,* op. 19, 1901.

Piano music

SOLO PIANO: *Five Pieces for Piano,* op. 3, 1892 (including the *Prelude in C-Sharp Minor*), *Six Moments Musicaux,* op. 16, 1896; *Variations on a Theme of Chopin,* op. 22, 1903; *Nine Etudes-Tableaux,* op. 39, 1916–17; *Variations on a Theme of Corelli,* op. 42, 1932; preludes, études, and sonatas.

TWO PIANOS: *Suite No. 1,* op. 5, 1893; *Suite No. 2,* op. 17, 1901.

Vocal Music

OPERAS: *The Miser Knight,* op. 24, 1904; *Francesca da Rimini,* op. 25, 1904.

SONGS: Approximately 72 songs composed between 1893 and 1916.

MISCELLANEOUS: *Liturgy of St. John Chrysostomus,* op. 31, 1910, (for mixed choir); *The Bells,* op. 35, 1913 (choral symphony); *Vesper Mass,* op. 37, 1915; *Three Russian Folksongs,* op. 41, 1928.

BIBLIOGRAPHY. The central collection of Rachmaninoff papers, holograph scores, and autograph letters in the U.S. is the Rachmaninoff Archive at the Library of Congress. The Rachmaninoff Room at the State Central Museum of Musical Culture, Moscow, also holds autograph papers, letters, scores, and a complete catalogue of works. Rachmaninoff's letters have been collected and published in Russian, *Письма* ("Letters"), ed. by Z. APETIANZ (1955). This collection includes all previously published letters and some newly published ones. *Rachmaninoff's Recollections Told to Oskar von Riesemann,* trans. from the German by D. RUTHERFORD (1934), are reminiscences by the composer about his life and work. One of the few interviews with Rachmaninoff published in English, "Some Critical Moments in My Career," appears in the June 1, 1930 issue of *The Musical Times* (London). SERGEI BERTENSSON and JAY LEYDA, *Sergei Rachmaninoff: A Lifetime in Music* (1956), is a comprehensive biography whose preparation was assisted by Sophia Satina, the composer's cousin and sister-in-law, containing letters previously unpublished in English, a complete list of works in order of composition, and a discography. Other important biographical studies are VICTOR SEROFF, *Rachmaninoff* (1950), a sympathetic biography containing letters and a complete list of compositions; and WATSON LYLE, *Rachmaninoff:A Biography* (1939), which includes portraits, a list of works in order of composition, and a critical survey of recordings. JOHN CULSHAW, *Sergei Rachmaninov* (1949), contains a biographical outline, an extensive discussion and critical analysis of the music, a complete list of works, and a discography. A contemporary evaluation of Rachmaninoff may be found in M. MONTAGU-NATHAN, *Contemporary Russian Composers* (1917).

(V.I.S.)

Racine, Jean

Racine gave supreme form to what has come to be called the "classical" French tragedy of the 17th century; he was among the greatest French poets in any genre; and in modern eyes he was the only French playwright ever to catch the authentic note (however it is defined) of tragedy. He was also the first French writer born without rank or money to die a courtier, a noble, and a man of means and influence.

Early life. Born in 1639, of a provincial family of modest excise or legal officials at La Ferté-Milon, some 40 miles northeast of Paris, Racine was left motherless at the age of one, fatherless and penniless at three. When he was nine his paternal grandmother became a widow

Racine, oil painting by an unknown artist, 17th century. In the Musée National de Versailles et des Trianons.
Giraudon

and took him with her to the convent of Port-Royal des Champs in the Chevreuse Valley, near Paris. There she took vows, joining a daughter and a sister both of whom were nuns, and another sister, Mme Vitart, who lived with her family near the convent.

Influence of Jansenism on Racine

This was Racine's first stroke of fortune. Around this community lived a handful of distinguished scholars and enlightened teachers who had retired to live lives of austere seclusion under the influence of Jansenism, a theological tendency in Catholicism condemned by Rome in 1642 and 1656, of which they were the champions. Three of them, during a temporary dispersal, had found shelter at La Ferté with Mme Vitart. Their famous Petites Écoles were carefully taught groups of from 10 to 20 boys, never totalling more than 50 in all. Racine, admitted free in consideration of the Vitart connection, gained from them a knowledge of Latin, literary taste, and more Greek than any other pure man of letters of his day; he also learned something of their sombre theology of original sin and grace; and it was among aristocratic former pupils that he later found his first patrons. In 1653–54 he was in a college of similar spirit in Beauvais, after which he returned to Port-Royal for further study at the Petites Ecoles. In 1656 the schools were closed by royal command, but Racine stayed, for this was his only home. He is said to have continued his Greek reading in the woods; he also imitated slightly outdated models of descriptive verse in the immature but interesting Promenades *de* Port-Royal.

After a year of law studies in Paris (1658), the young man found employment with his grandmother's nephew Nicolas Vitart, steward of Charles Honoré d'Albert, duc de Luynes. In 1661 Racine's mother's family took an interest in him again; alarmed at his worldly associations in Paris, they sent him down to the cathedral of Uzès, in Languedoc, where an uncle was vicar general and might get him an ecclesiastical living. He spent about two years there, studying and noting the violent temperaments and feuds of this still-Huguenot region, but his uncle's ambitions for him bore no immediate fruit.

Back in Paris he followed his own bent and plunged into a career of authorship. His only literary associate of merit, as yet, was the future fabulist Jean de La Fontaine. Drama was at this time the best road to success, but this meant defying his Jansenist kinsfolk and teachers, who were ahead of general Catholic opinion in condemning the stage and all connected with it. He was told in 1663 to keep away from Port-Royal unless he reformed. In 1666, after his first dramatic success, he counterattacked bitterly with a brilliant anonymous open letter (Lettre *à l'auteur des 'Hérésies imaginaires'*) to his old master Pierre Nicole, who had recently written (with no reference to Racine) that a novelist or playwright was a public poisoner.

Mature life and works. Racine's career as an increasingly successful poet-playwright ran from 1664 to 1677, coinciding with the brilliant early years of Louis XIV's reign. Even before Uzès he had obtained a little notice by publishing an ode on the King's marriage (1660); another ode on a royal attack of measles (1663) brought him a small gratuity, the first of a great many. His letters of exile from the south speak of one rejected and one abandoned play (there may have been a third). On his return he somehow met Molière, the great comic poet and actor-manager, who accepted, perhaps even commissioned, his first tragedy, La *Thébaïde* ou *les frères ennemis*, produced with mediocre success in 1664. Alexandre, the year after, was as superficial and fashionable in its use of the themes of love and glory as La The'bai'de had been archaic in its gloom and violence. Alexandre did very well, but Racine, apparently dissatisfied with Molikre's relatively natural acting style, surreptitiously took the script to the oldest of the three permanent theatres then in Paris, the Hôtel de Bourgogne, which actually put it on while it was having its sixth performance with Molière. The latter took offense and dropped Racine, who gave all the rest of his commercial plays to the Hôtel de Bourgogne. Molikre's star actress, Du Parc, followed Racine and became his mistress. Hers was the title role in *Andromaque* (first performed 1667), in which Racine found what was to be his most appreciated (though not his only) theme: the tragic folly and blindness of passionate love. As poetry, the play was far superior to any competitor. In the neat efficiency of its dramatic mechanics it was a model—if anything, too clever. Its sometimes antiheroic moral realism disturbed some devotees of the idealistic tradition of romance, but in spite of critical cavils it was a sensational triumph.

Collaboration with Molière

Breach with Corneille

There followed Racine's only comedy, the three-act Les Plaideurs (1668; *The* Litigants) a slight but witty adaptation of The Wasps—far removed from the original, but no one else in the century would have thought of borrowing from Aristophanes at all. Then came the sombre *Britannicus* (1669) and the touching *Bérénice* (1670), both set in imperial Rome. The first was the occasion of an open breach with Pierre Corneille, the declining idol of an older generation of playgoers, who attended the first night and was openly hostile. *Bérénice* was based on the same historical episode as a play of Corneille produced almost simultaneously with it. (Such competitions were quite normal: Racine's *Thébaïde* seems to have been a counterblast to a tragedy put on by the Hôtel de Bourgogne, and *Iphigénie* and *Phèdre*, later on, both encountered rivals with the same titles and subjects.) One or both playwrights had some idea of what the other was writing, to judge by the texts. Corneille's play was not a failure, but Racine's was another triumph. He had shown up, by a deliberate contrast, his rival's weakness: the complexity of motives and interests that Corneille at this time was weaving into each of his plots. The point was pressed home in the prefaces to both Britannicus and *Bérénice*.

Not that this plea for "simplicity of action" was Racine's last word. Still trying new directions as he never ceased to do, he made his next subject, *Bajazet* (1672), from almost contemporary Turkish history, full of intrigue and danger. Mithridate (1673) confronts an aging Asiatic despot·with a Greek heroine. *Iphigénie,* first produced in 1674 in the outdoor setting of a great royal fête at Versailles, is an adaptation of *Iphigenia* at *Aulis* by Euripides, but with a love plot and a happy ending; *Phèdre* (1677) was to use the same Greek poet even more successfully.

Literary triumph

Iphigknie had a public run of extraordinary length and brought its author, at 35, to the apex of his first, literary, career. Racine was now head of his profession and powerfully patronized: official backing had won him membership in the Académie Française, the officially recognized body that arbitrated French literature and language, in 1672; in 1674 he obtained a sinecure conferring nobility, a treasurership of France. The man who had been in debt and without prospects in 1662 had been putting aside his savings in careful investments since 1669 and, for an author, was well off. The year 1676 saw a collective edition of all Racine's work to date, with the waspish polemical passages removed from the prefaces; the new or recast versions consciously built up the image of the scholar-poet.

Phèdre was the last, most profound, and most poetic of the tragedies for the Paris stage; the heroine suffers equally from her incestuous passion and her sense of degradation. Coming as it does before Racine's decisive break with the theatre, the play has often been thought to reflect the author's remorseful return to the principles of his youth.

Retirement from the theatre. Within eight months of the premiere of *Phèdre*, the great playwright had cut all links with the commercial stage, married a pious, unintellectual young woman of 24 or 25 (a connection by marriage of the Vitarts), and accepted—with his friend the critic and satirist Nicolas Boileau—the high honour of writing the official history of the reign.

Racine's motives

There will always be debate about Racine's motives for this sudden retirement. It was said that his present mistress, the actress Champmeslé, had played him false (but she was notoriously promiscuous, anyway). It is true that *Phèdre* had been challenged for a short time by a rival play, and Racine, with Boileau, had been implicated in an ugly episode as supposed authors of scurrilous sonnets against Philippe Mancini, duc de Nevers, supporter of the other playwright, Jacques Pradon—but these were passing clouds. Racine spoke later of having turned away from a life of sinful dissipation, and ttie son who wrote his life gives a radical conversion as the sole cause of the retirement. Some modern scholars suggest that, with *Phèdre*, Racine had said all it was in him to say; or that he had shocked himself by this powerful picture of a world where the gods could be evil.

Most attribute at least some part to ambition and the desire for security. It must be remembered that, although he had made sacrifices to embrace a career in letters, he had always courted patronage, and that art had far less dignity in his society than it has since acquired. The new post, proffered through the influence of the King's favourite, Mme de Montespan, was an unexpected honour and a surprisingly well-paid one; socially and financially it would have been hard to refuse; humanly, perhaps impossible. The stage, too, was more and more frowned upon in religious circles, and the author of *Phèdre* himself was to express the same views in his last years.

He remained a man of letters; as such he was already a classic (and as irritably jealous as ever of his reputation). He went on revising new editions of his plays (1687 and 1697), which in book form seemed less blameworthy than on the stage. "The scandals of his past life," as his deathbed testament calls them, had resulted from his playwright's life, and ceased with it. His only known love affairs were with his two leading actresses, whom he coached in their roles. The first, with Mlle du Parc, it is true, seems to have been a passionate one. The affair ended with her death, which may have been caused by an abortion; but accusations to this effect against Racine by a disreputable character, in the course of a criminal inquiry into poisonings and black masses in 1679–80, were never followed up. (One tradition says that she had earlier borne Racine a daughter, who died in childhood).

After 1677 his calling and way of life were irreproachably respectable. His wife presented him with two sons and five daughters, one of whom took the veil just before his death and one just after. He was a strict and solicitous father.

He renewed contact with Port-Royal and showed some courage in exerting his influence to help it. Not that he renounced the world: he had to be near the King, whose numerous campaigns he followed in order to report them; his pen was at the command of the King, Mme de Montespan, and later Mme de Maintenon, the King's morganatic wife. His social talents made him acceptable at court, so much that the envious accused him of hypocrisy. New honours arrived: he became a gentleman of the bedchamber in 1690, a second and more indubitable title of nobility than the treasurership, and one that was made hereditary by a new grant in 1693. All these had been—quite exceptionally—free gifts of the King; not so the secretaryship he purchased (for unknown reasons) in 1696. That post, and a tax levied on it in 1698, upset his hitherto flourishing finances and proved in the end that he could not afford the social eminence he had earned.

Last two plays

Racine's last two plays were commissioned by Mme de Maintenon for her school for daughters of impoverished nobles at Saint-Cyr. The preface of *Esther* (1689) explains that the plan for an improving dialogue with sung interludes was an unhoped-for chance to revive two features of Greek drama—the chorus and the religious inspiration. (Another model, not mentioned, was the opera.) *Esther* was a brilliant court occasion. The play's three acts, something of a nursery tale in comparison with the passion and subtlety of Racine's earlier dramatic actions, are redeemed by the poetry of the choruses. *Athalie* (1691), which for uncertain reasons never had a public performance in Racine's lifetime, has five acts with choruses, and, though different, equals most or all of the rest.

Racine died of an abscess (which may have been a cancer) on the liver on April 21, 1699, after 18 months of notably stricter austerity. He was buried at his own request in the Port-Royal graveyard. When Port-Royal was destroyed by royal decree (1710), his body was reinterred at Saint-Étienne-du-Mont.

MAJOR WORKS

PLAYS: *La Thébaïde ou les frbres ennemis* (1664); *Alexandre le Grand* (1666); *Andromaque* (1668); *Les Plaideurs* (1669); *Britannicus* (1670); *Bérénice* (1671); *Bajazet* (1672); *Mithridate* (1673); *Iphige'nie* (1675); *Phèdre* (1677); *Esther* (1689); *Athalie* (1691).

VERSE: "Cantiques spirituels" (1694).

PROSE: *Abrégé de l'histoire de Port-Royal* (1767).

TRANSLATIONS: The plays were rapidly translated into Dutch. Ten English stage versions apepared, 1675–1723, but differences of national taste made them grossly unfaithful: E. SMITH, *Phaedra and Hippolytus* (1707) and A. PHILIPS, *The Distrest Mother* (after *Andromaque*, 1712) had successes that now appear undeserved. SCHILLER's German *Phadra* (1805), is admired. The great obstacle in modern English is the absence of an accepted equivalent poetic style. KENNETH MUIR has translated *Five Plays* (1960); J. CAIRNCROSS, *Phaedra and Other plays* (1963) and *Andromache and Other plays* (1967); s. SOLOMON, the *Complete Plays,* 2 vol. (1967–68). See also R.C. KNIGHT, *Phèdre* (1971).

BIBLIOGRAPHY. The works are adequately cataloged in the standard editions. No manuscripts survive of the principal works. The correspondence is interesting for the early years, abundant for the married life; but almost every letter from the years of commercial playwriting has been destroyed. Most of the rest, with private papers and annotated books, was presented by Racine's son to what is now the Bibliothèque Nationale, Paris. See also E.E. WILLIAMS, *Racine depuis* 1885, *essai de bibliographie raisonnée* (1940).

The only critical edition complete with biographical material, still consulted though superseded in much of its detail, is the *Œuvres*, ed. by P. MESNARD, 8 vol. (1865–73). The "Pléiade" edition, by R PICARD, 2 vol. (1950–52), gives accurate text, good critical introductions, and many of Racine's notes. The Garnier edition (1960) and "l'Intégrale" (ed. by P. CLARAC, 1962—without variants), may be commended.

Racine's life has been treated exhaustively and definitively in its external and material aspects by R. PICARD, *La Carrière de Jean Racine* (1956). The *Me'moires* of L. Racine, who was not seven at his father's death (in Mesnard and Picard editions) are conscious whitewashing. See also G. BRERETON, *Jean Racine:* A *Critical Biography* (1951), balanced and sympathetic; FRANCOIS MAURIAC, *La Vie de Jean Racine* (1928), a religious interpretation; and R. JASINSKI, *Vers le vrai Racine,* 2 vol. (1958).

For historically-based academic criticism, see O. DE MOURGUES, *Racine, or, The Triumph of Relevance* (1967); J.C. LAPP, *Aspects of Racinian Tragedy* (1955); and R.C. KNIGHT (ed.), *Racine* (1969). Modern critical tendencies are represented, with moderation, by J.D. HUBERT, *Essai d'exégèse racinienne: les secrets témoins* (1956), a thematic study of images; and much more extremely by French *néo-critiques* —L. GOLDMANN, *Le Dieu cache': etude sur la vision tragique dans les Pense'es de Pascal et dans le théâtre de Racine* (1955; Eng. trans., *The Hidden God: A Study of Tragic Vision in the Pense'es of Pascal and the Tragedies of Racine,* 1964), structuralist and Marxist; R. BARTHES, *Sur Racine* (1963; Eng. trans., *On Racine,* 1964), structuralist and Freudian; and c. MAURON, *L'Inconscient dans l'œuvre et la vie de Racine* (1957), Freudian psychocriticism.

(R.C.K.)

Racism

Racism is the theory or idea that there is a causal link between inherited physical traits and certain traits of personality, intellect, or culture and, combined with it, the notion that some races are inherently superior to others. The term racism has no necessary relation to biological or anthropological definitions of race, a subdivision of a species (see RACES OF MANKIND). Racist ideas are often indiscriminately extended to apply to such nonbiological and nonracial groupings as religious sects, nations, linguistic groups, and ethnic or cultural groups. As one authority has noted, "Racism is quite different from a mere acceptance or scientific and objective study of the fact of race and the fact of the present inequality of human groups." (J. Comas, "Racial Myths," in The Race Question *in* Modern Science, UNESCO, 1956, pp. 52–3.)

(Racialism is an older term, more or less synonymous with racism, the latter now being more common. Some recent writers have attempted a division in their meanings —applying *racism* to the theory or doctrine and *racialism* to the practice of discrimination and prejudice.)

DEFINITION OF RACIAL TERMS

Racial and ethnic criteria. Race, in the sense relevant to racism, refers to a human group that defines itself or is defined by others as culturally different by virtue of innate and immutable physical characteristics. Thus, under racism a race is defined socially but on the basis of physical characteristics. Such physical characteristics have no inherent significance, but only such significance as is socially attributed to them in a given society. If a group in a given society is defined in terms of its skin colour, its hair texture, its facial features, its body build, and so on, then it is a "race," in racist terms. Presumably, if a group were socially defined in terms of sharing a common language, a common set of religious beliefs, or some other cultural characteristic—without physical considerations—then it would be an "ethnic group." In the case of both race and ethnic group, the defining criterion is relative to a given society, and the same person or group may be differently defined in different societies. Thus a black North American may be a "Negro" in his own country but a "Yankee" in Mexico. The first label is racial, the second ethnic or cultural.

In practice, the distinction between a "race" and an "ethnic group" is not always clear-cut, and many groups are socially defined in terms of both physical and cultural criteria. Jews, for instance. have frequently been regarded as both a racial and a cultural group. In such equivocal cases, it is necessary to try to determine which criterion is paramount. In Nazi Germany, Jews were primarily regarded as a race, whereas in the Middle East they are more an ethnic group. In any case, the phenomena of prejudice and discrimination that are commonly linked with differences between human groups are not markedly different whether one deals with a race or with an ethnic group. The one essential distinction is that culturally based groups generally tend to be more open and flexible in composition than physically based ones. One can convert to another faith or learn another language, whereas what one can do to one's physical appearance is decidedly limited. It should be remembered, however, that ancestry can mark a person in racist terms even though he may outwardly have no distinguishing physical trait. "Aryans" in Nazi Germany, for instance, were summarily reclassified as "Jews" when search into their ancestry discovered a Jewish forebear, and "whites" in South Africa have been similarly reclassified as "coloureds" after investigation had revealed a non-European ancestor.

Ethnocentrism and racism. Parallel to the problem of distinguishing between race and ethnic group is the problem of distinguishing between racism and ethnocentrism. Presumably, the distinction is that in ethnocentrism the alleged inferiority, disabilities, and negative traits of the outgroup are thought to be culturally determined (and only culturally determined), whereas in racism there is a belief that the disabilities are inborn. In practice, however, the cause of the alleged disability is not always made explicit, and it is not always easy to distinguish between racism and ethnocentrism. This is the case, for example, with anti-Semitism, which sometimes takes a cultural emphasis and sometimes a racial one. Ethnocentrism is in reality a much more widespread phenomenon than racism. Indeed, it may be said to be almost universal. Members of nearly all the world's cultures regard their own way of life as superior to that of even closely related neighbours. The exceptions to the universality of ethnocentrism are some of the societies, often small nonliterate ones, that have been conquered by more powerful ones and reduced to the status of a subject people. In some cases, after many years of oppression, some peoples have accepted their conqueror's derogatory views of themselves and flattering view of their masters. Under slavery, great numbers of American blacks, for instance, were psychologically conditioned to consider themselves inherently inferior. Unlike ethnocentrism, racism is far from universal. In fact, as a fully explicit theory of the biological causation of cultural behaviour, it is the exception rather than the rule. Ethnocentrism and racism are not mutually exclusive, of course: racist societies are almost invariably ethnocentric as well; people often feel superior to others on both racial and cultural grounds. Nevertheless, many highly ethnocentric societies are not racist.

Ethnocentric and racial aspects of anti-Semitism

DISCRIMINATION AND PREJUDICE

The term racial discrimination denotes all forms of differential behaviour based on race. The most notable form of racial discrimination is, of course, physical segregation by race, but there are many others, such as rules of etiquette defining forms of address between racial "superiors" and "inferiors," or choice of friends or spouses. Racial endogamy (that is, marrying within one's own racial group) is frequently required and almost always preferred in racially stratified societies. *Commensality* rules (rules determining with whom one may or may not eat) are also a very common manifestation of racial discrimination. The term racial discrimination, then, always refers to behaviour and indeed to social behaviour.

A second aspect of racism is racial prejudice. Prejudice is a psychological phenomenon; it is defined as an attitude, usually emotional, acquired without, or prior to, adequate evidence or experience. It can be favourable or unfavourable and can develop in a person through suggestion, belief, or emulation. Racial prejudice consists of negative attitudes directed in blanket fashion against socially defined races (not necessarily coincident with biologically defined races).

Although it is important to distinguish between the psychological and social aspects of racism—that is, between discrimination and prejudice—it is equally essential to understand that the two are inextricably related. Prejudice against a certain category of persons could not develop in the minds of individuals unless their society had already set groups apart and subjected some of them to discrimination. Conversely, it is difficult to conceive of a society in which a system of discrimination could exist in the absence of widespread prejudice among members of the discriminating group. Discrimination and prejudice are mutually reinforcing. Prejudice is a rationalization for discrimination, and discrimination often brings forth in the victims those behaviour patterns that seem to validate the prejudice. A white bigot, for example, can easily rationalize the existence of inferior schools for blacks if he believes that blacks are not capable of benefitting from equal schooling. The same bigot, when he compares the competence of graduates of the inferior black schools and white schools, will find confirmation for his belief that blacks are intellectually inferior.

Reinforcing roles of discrimination and prejudice

If there were always a one-to-one relationship between discrimination and prejudice, the distinction would serve no useful purpose. However, this is not the case. There are wide differences in individual prejudice, tolerance, and open-mindedness within a given society. One finds relatively open-minded or unprejudiced persons who live in virulently racist societies and pathological bigots who live in nonracist societies. It is in such cases that one can expect to find discrepancies between the level of prejudice and discrimination. The open-minded white person who lives in South Africa, for example, finds it practically impossible not to be racially discriminatory in his behaviour. The law compels him to be so; and even if he chooses to go to jail in protest, he will be sent to a superior white jail. The reciprocal is also true. The bigot who lives in a nonracist society will often hide his prejudice and refrain from open acts of discrimination. In both cases, the person conforms to social norms in his outward behaviour; that is, he behaves in a manner consistent with the behaviour of the majority but inconsistent with his own attitudes.

A number of studies have demonstrated, however, that such situations are transitory for most people because they are so difficult to live with. Open-minded persons tend to become prejudiced after moving to a racist environment, or at least to become more prejudiced than they would otherwise have been; conversely, prejudiced persons become more tolerant in a nonracist society. These changes in attitudes can be quite rapid. taking no more than a few days or weeks, and changes in behaviour can be nearly instantaneous. Generally, however, there is an appreciable time lag between change in behaviour and change in attitudes, and it is this lag that makes the imperfect correspondence between discrimination and prejudice.

THE HISTORY OF RACISM

In nearly all the world's societies, men have apparently developed pride in the cultural accomplishments of their own groups and a corresponding derogation of those of their neighbours. Notably, however, the idea that certain groups of people are superior to others because of their genetic makeup does not appear to have been widespread. Where it now exists, it is mostly an outgrowth of the rationalizations of slavery and colonial expansion in the vast territories dominated by European settlers.

Non-European racism. Some scholars have argued that the Hindu caste system originated in a physical difference between the conquering Aryans and the conquered Dravidians. The top three caste levels, or *varṇa*—Brahmin, Kṣatriya, and Vaiśya—were in all probability recruited from the ranks of the Aryans, whereas the fourth *varṇa*, the Śūdra—as well as the "untouchables" who fall outside the scope of the four *varṇa*—were perhaps originally composed mostly of Dravidians. The word *varṇa* means "colour," even though the colours traditionally associated with the *varṇas* show no relation to skin pigmentation. It is true that in Hindu culture, the colour dualism of white and black has associations with good and evil as it does in the Judeo-Christian tradition, but the link with skin pigmentation is not established. In modern India, there is a weak association between light skin and high caste, and perhaps a slight overrepresentation of lower castes among the darker skinned southern Indians, but skin colour in India is anything but a reliable index of social status. Kṛṣṇa, the most cherished character in the Hindu mythology, is often represented as a darkly handsome young man frolicking with a bevy of fair girls. Many Indians express a marked aesthetic preference for light skin, especially in women, but, on balance, the evidence that the Indian caste system is racial in origin and that India is or was a racist society is unconvincing. The basic caste dichotomy between "once-born" and "twice-born" was probably related to the *cultural* distinction between Aryan conquerors and Dravidian conquered. The latter were probably darker skinned than the former, but it is not established that this physical distinction was the socially significant one.

Skin colour and Hindu caste

Middle Eastern racial attituties. The Bible, one of the oldest sets of written documents in the Judeo-Christian tradition, contains no positive suggestion that the ancient Semites were racists. The ambiguous reference in the Song of Solomon ("I am dark, but comely, O daughters of Jerusalem, like the tents of Kedar, like the curtains of Solomon. Do not gaze at me because I am swarthy, because the sun has scorched me. My mother's sons were angry with me, they made me keeper of the vineyard;") might be interpreted as expressing a prejudice against blackness. Several other passages in the Bible, notably the curse of Noah on Canaan in Gen. 9:22–27, have been interpreted by later racists in a racial light, but in themselves they do not necessarily suggest anything of the kind, and, on balance, the evidence is negative.

The same is true of the Qur'ān and the Islamic tradition in which religious conversion has always been the test of membership in the ingroup. Although it is true that black Africa was the source of many of the slaves in Muslim countries and that, to this day, in some Arab countries a dark skin is a presumptive sign of slave descent and hence of low social status, none of the Muslim countries ever developed a racial caste system. Since the spread of Islam across the Sahara in the early 11th century, it has become by far the most important religion in black as well as in white Africa, and by now approximately one-fifth of the world's and one-half of Africa's Muslims are black. From medieval times, black African kings went on pilgrimage to Mecca with thousands of their subjects and were well received, and Arab writers on black Africa reveal no racism in their writings. They often express admiration for the black Muslim kingdoms that they encountered, and whatever negative value judgments they make are religious or moralistic in nature (such as the disapproval of the nakedness of women). Even the devastation brought about by the Arab slave trade in East Africa in the middle of the 19th century does not appear to have been rationalized on racial terms as European slavery was.

East Asian views of race. Chinese and Japanese cultures have traditionally been conscious of physical differences, especially from an aesthetic point of view, but neither society ever developed a racial caste structure. Even in pre-Tokugawa Japan, where there existed a fairly rigid "estate" system similar to that of medieval Europe, the distinctions were apparently not racial, and the Eta caste—an outcast, "untouchable" group that still exists today—is not physically distinguishable (even though one theory, popular in Japan, holds that the ancestors of the Eta were anciently Korean war prisoners and slaves). In any event, both Chinese and Japanese express a sometimes marked aesthetic dislike for physical types that are at variance with their own well-established canons of beauty. A 16th-century Chinese mandarin, for example, likened the hairiness of Portuguese sailors to that of monkeys and pronounced the Portuguese totally lacking in the social graces of civilized life, though surprisingly clever and capable of learning. Many Japanese find black skin and other Negroid characteristics unattractive and, since the recent wave of American influence, appear to have adopted some Western canons of beauty, as witnessed by the eyelid operations undergone by some women. The generally inferior status of Koreans in Japan seems to be based more on socioeconomic and cultural than on racial criteria; and the distinctive-looking Ainu minority on Hokkaido Island do not appear to be the target of racial prejudice. Thus, despite narcissistic canons of physical beauty and highly ethnocentric judgments of other cultures, East Asian civilizations do not exhibit what might properly be called racism.

Racism in nonliterate societies. The evidence from the world's many nonliterate societies is fragmentary on the topic of interethnic and interracial relations. There are, however, a few documented cases of indigenous systems of racism not attributable to contact with Western societies. One of the most notable cases is that of the traditional kingdoms of Rwanda and Burundi in the interlacustrine area of central Africa. Until the overthrow of the *ancien régime* shortly after independence in 1961, the population was stratified into three racial castes. The tall, semitic looking Tutsi constituted some 15 percent of the population and, as pastoralist conquerors, ruled over the Hutu horticulturalists. The Hutu, of average stature and distinctly Negroid in physical appearance, constituted a servile peasantry of approximately 80 to 85 percent of the population. At the bottom of the hierarchy was a small group (approximately 1 percent of the total) of pygmoid Twa. The close relation between physical stature and social status made for a truly racist system. The Tutsi claim to superiority was based to a considerable extent on their physical characteristics.

Development of racism in Europe. Far and away the most widespread, enduring, and virulent form of racism and the costliest in terms of human suffering has been that which developed in western Europe and its colonial extensions in Africa, Asia, Australia, and the Western Hemisphere. Western racism is of relatively recent origin. None of the main waves of influence on European civilization seems to have brought racism with it. In ancient Greece and Rome, the status criteria were cultural and not racial. The barbarian was the person who spoke another language. Slavery was a juridical and economic condition unrelated to racial or ethnic origin. The statuses of patrician or plebeian, patron or client, slave or freeman, citizen or noncitizen were all based on strictly nonracial criteria. Although Roman literature is often quite ethnocentric in its judgment of barbarians, differences in physical appearance are generally mentioned neutrally as interesting *exotica.* A number of black Africans must have reached Rome in the last centuries of the empire, and blacks appear on frescoes in a number of Roman sites in Europe and North Africa, but there is no evidence that they were regarded as inherently inferior.

In the Middle Ages, the religious criterion of membership in the ingroup became paramount. Anti-Semitism was clearly religious and not racial in nature and continued so through the Renaissance, the Reformation, and the Wars of Religion of the 16th and 17th centuries. The important thing was whether one was Christian, Moor, or Jew, Catholic or Protestant. Even at the height of religious persecution during the Spanish Inquisition and the Thirty Years' War, acceptance was typically granted in return for a simple public profession of faith. "Race," in any case, was irrelevant.

The anti-Semitic wave that swept Germany in the 1930s and ended in the crematoria of the death camps in the 1940s will hopefully remain the most heinous manifestation of racism in human history. Although Nazi anti-Semitism grew out of a long tradition of religious intolerance in Europe and more especially of victimization of Jews in eastern and central Europe, Hitler's theory of the master race gave it a hitherto unknown genocidal virulence.

Racism in European colonies. Even when Europeans first came into more extensive and direct contact with large numbers of dark-skinned peoples, as a result of their colonial expansion starting in the late 15th century, racism took some time to develop.

The Spanish record. The Spanish conquest of the New World was more than averagely brutal, and the economic exploitation of the Indians was thorough; but the Spanish crown quickly settled the issue of the Indians' humanity by declaring that they did have souls worthy of salvation and that they should not be enslaved. It is true that in the Spanish colonies of the Americas, there developed a caste system that was at least partly racial: a fundamental distinction was made between an Indian, a *mestizo* (or mixed-blood), and a Spaniard; but an equally important difference was made between a Spaniard born in Spain and a criollo, a Spaniard born in the colonies. Initially, the Spaniards expressed wonder and admiration for the accomplishments of the Indian civilizations that they destroyed. This did not prevent them from condemning what they regarded as barbarism and idolatry, but there is no suggestion that the Spaniards regarded the Indians as genetically inferior when they first met them. Only after the Indians had been reduced by epidemics, wars, tribute exactions, and countless acts of brutality to the status of an impoverished servile peasantry did negative stereotypes about them develop. The Spanish colonial caste system established in the 16th century was continuously undermined by a dual process of hispanization and miscegenation, which blurred both cultural and racial distinctions between groups. By the time that the criollos gained their independence from Spain in the second decade of the 19th century, the caste system had lost its social significance except in the countries that, like Guatemala, Peru, and Bolivia, were still predominantly Indian.

Today, racism, though not totally absent in Spanish America, is certainly much less prevalent than in other parts of the continent. Cultural criteria are far more important than physical ones in most of Spanish America, even in the heavily Indian countries. In some Latin American countries, the term *mestizo* no longer generally denotes a person of mixed Spanish-Indian descent but instead often designates a person who speaks Spanish as his mother tongue even though he may be of pure Indian ancestry. Many derogatory stereotypes about Indians do continue to exist, and in the heavily Indian countries, Indians continue to be subjected to various forms of social discrimination; but, by and large, an Indian ceases to be regarded as such when he becomes hispanicized in his language, customs, and religion. Racism, as such, is minimal.

Muted racism in Spanish America

Portugal in Africa and Brazil. The Portuguese case is somewhat different from the Spanish one. In Africa, the initial contacts between Portuguese and Africans were relatively free of racism and relatively peaceful and friendly, except in east Africa, where Portugal came in conflict with Arabs and Persians. There is, for example, a record of a friendly correspondence between the kings of Portugal and Kongo and of an exchange of ambassadors. Starting in the middle of the 16th century, these auspicious beginnings were increasingly compromised by Portuguese military incursions, the sacking of cities, and incessant demands for slaves for the Brazilian plantations. Brazil soon became, with the West Indies and the Southern English colonies of North America, one of the world's three major consumers of black slave labour, and it remained so until the middle of the 19th century. Generally, the Portuguese claim that its colonialism in Africa has been nonracial is correct, at least by comparison with the British, Belgian, and Dutch. This is not to say that the Portuguese regime in Africa has been any less oppressive and exploitative than the regimes of the other colonial powers, but whereas the latter have frequently applied racial tests of discrimination, the Portuguese have been ethnocentric rather than racist.

Complex race relations in Brazil

In Brazil, race relations are quite complex and vary greatly from one region to another. Brazil's reputation as a "racial paradise" appears to most observers to be undeserved; on the whole there is considerably more racism in Brazil than in Spanish America, though much less than in the United States. Consciousness of physical differences is highly developed, and Brazilians use a complex racial nomenclature to describe a score or more of combinations of facial features, hair texture, and skin pigmentation resulting from the intermixture of Afro-, Euro-, and Indo-Brazilians. Certain stereotypes are attached to these physical types as well as an order of aesthetic preference, but the very complexity of the nomenclature itself has made impossible the establishment of any system of racial castes, and members of the same family may frequently fall into different "racial" categories. Thus, Brazil might be described as a highly racially conscious country but without a rigid system of racial castes and without well-defined forms of racial discrimination. Such discrimination as exists is usually a subtle combination of racial, ethnic, and social-class factors, with race frequently not the most important one.

The French experience. The French, like the Portuguese and Spaniards, tended to be more ethnocentric than racist in their colonial policy of "assimilation," but in practice, like the Portuguese, they failed to assimilate more than a tiny minority of their African subjects. Consequently, except for a few black professionals and intellectuals who were totally integrated in French society, French colonial policy was not all that different in practice from that of more avowedly racist powers. It should also be noted that in Algeria, where there was a European settler population of over 1,000,000, the French exhibited considerable racism vis-à-vis the Arabs.

Dutch and English colonies. The Netherlands and Great Britain were responsible for the growth of the most racist colonial societies that the world has ever known—namely, South Africa, the United States, and Australia. In Australia, racism has taken the form of discrimination against the aboriginal population and the exclusion of immigrants of non-European stock. In the United States, virtual genocide against the Indian groups was accompanied by the institutionalization of a slave plantation system and, after 1865, of racial segregation and discrimination against the "emancipated" Afro-Americans. In addition, a number of states adopted racially discriminatory laws against other nonwhite groups, notably Asians, in the field of immigration, marriage, and political rights. As late as World War II, tens of thousands of American citizens were interned for several years in camps, solely on the basis of their Japanese ancestry.

The South African policy of apartheid has become a byword for racial discrimination and, next to the Nazi policy of genocide against Jews, represents the most extreme and systematic form of racism practiced in a modern society. The white government of South Africa is attempting to create four rigid colour-castes (Europeans, Asians, Africans, and Coloureds), to segregate them physically, and to perpetuate the economic and political privileges of the white minority at the expense of 80 percent of the population. In theory, apartheid aims to establish a "separate but equal" system, but in practice the indefinite maintenance of white supremacy is clearly the objective. Countless laws limit the nonwhite South Africans' rights to travel, own and occupy land, hold meetings, seek work, attend universities, enter public places, marry, vote, and indeed be present almost any place without the consent of the white authorities.

IDEAS ABOUT THE CAUSES OF RACIAL PREJUDICE AND DISCRIMINATION

"Scientific" support in 19th century

As a well-developed theory, racism is a fairly recent phenomenon, even in Western history. The 18th century was predominantly environmentalist in its outlook; the science of that day tended to attribute social behaviour either to climatic and geographical environment or to sociocultural factors. Racism as a widely accepted "scientific" theory of behaviour did not appear until the 19th century, which was the age of racism par excellence. Although Charles Darwin himself was not a racist, his theory of biological evolution was extended to social evolution, giving birth to the theory of social Darwinism. Mankind was regarded as having achieved various levels of evolution, culminating in the white-European civilization. These stages of evolution were thought to be related to the innate genetic capabilities of the various peoples of the world. By the second half of the 19th century, racism was accepted as fact by the vast majority of Western scientists, and various forms of it were popularized through the writings of Joseph-Arthur, comte de Gobineau, Houston Stuart Chamberlain, Rudyard Kipling, Alfred Rosenberg, and Adolf Hitler.

By the 1930s the intellectual climate had swung clearly away from racism, and racism had lost its apparent scientific respectability. The social sciences began to adhere to a strict theory of the social determination of human behaviour, to the nearly complete exclusion of biological or physical-environmental factors. Man was held to be almost entirely a product of his culture, and each culture was to be evaluated in its own terms. This cultural relativism was popularized by such anthropologists as Ruth Benedict and Margaret Mead, and the study of race relations became distinctly antiracist in orientation. Gordon Allport, Otto Klineberg, Roger Bastide, Gunnar Myrdal, and E. Franklin Frazier, to name but a few of the prominent scholars active in the 1930s and 1940s, all took the position that such "racial" differences as are found between human groups are attributable to the differential social environments in which they find themselves and not to any intrinsic physical properties.

Most recently the pendulum is once more swinging away from the extreme social environmentalism and cultural relativism that dominated the social sciences until the 1950s. Genetic and biological factors in human behaviour and aptitudes are once again becoming accepted as having at least some importance, although the crude racism, evolutionism, and social Darwinism of the 19th century seem to be permanently discredited.

As a rule, complex social phenomena like racism cannot be explained in terms of a single causal factor. Causation is not only multiple but also often reciprocal, in the sense that *A* generally is a cause of B and B in turn frequently is a cause of *A*. This is clearly what happens in the relation between discrimination and prejudice. Each is a cause of the other, and this vicious circle is often difficult to break. The reciprocal causation between discrimination and prejudice illustrates the interplay between psychological and sociocultural variables, and both approaches are essential to a comprehensive understanding of race relations.

"Their own fault" theory

Specious popular *theories*. Numerous theories on the causes of racism have been advanced, some patently false, others partly true; at the present stage of knowledge, only an eclectic but selective acceptance of several of the theories can give a comprehensive view of the phenomenon. Perhaps one of the commonest popular theories to account for prejudice and discrimination (both racial and ethnic) is that it is "their own fault." The theory advanced by most dominant groups in racially or ethnically stratified societies is that the allegedly reprehensible behaviour and qualities of the outgroup causes discrimination and prejudice in the ingroup. This theory, even when it is not made explicit, is often implied in the very definition of the situation as the "Negro problem," the "Jewish problem," the "Asian problem," and so on. It should be clear that such an approach is a *symptom* of racism or ethnocentrism and not an explanation thereof.

Another unviable theory of the cause of racism is that there is an innate or instinctive repulsion between groups of people who look different. Experiments on young infants have shown that infants could easily be conditioned to scream with fear when a white rabbit entered a room and, conversely, to pet snakes. There seem to be no innate dislikes of any animals in man, much less of relatively minor differences in appearance between members of the same species. If this is true, then there is no intrinsic relation between racism and the sheer presence of physical differences. Physical differences are one of the important conditions facilitating the development of racism once it is there, but they are not a cause of racism. Differences in physical appearance are simply visual stimuli to which prejudices may or may not become attached.

Psychological theories. Several psychological theories about the causes of racial and ethnic prejudice have linked the phenomenon with certain personality traits or with certain responses to social situations. The "frustration-aggression" theory holds that frustration frequently leads to aggression and that this aggression becomes "displaced" onto scapegoats that are quite unrelated to the source of the frustration. Outgroups are frequently blamed for one's frustrations and failures, and this displacement of aggression is often accompanied by "projection^{w}—that is, by the attribution to others of one's own undesirable or unavowed traits. Although a number of experiments have shown that experimentally induced frustration can lead to displaced hostility against outgroups, the theory does little to explain why certain persons are more apt to displace their hostility than others and why certain groups are chosen as scapegoats. It also fails to account for the fact that highly privileged groups can be as strongly prejudiced as groups that have suffered much frustration.

The "authoritarian personality" approach holds that persons who exhibit certain attitudes and personality traits such as respect for power, submission toward superiors, aggression toward subordinates, lack of self-insight, superstitiousness, and contempt for weakness are predisposed to be generally prejudiced against all ethnic and racial outgroups. Numerous studies have shown some relation between these factors, at least in Protestant. Anglo-Saxon cultures, but the theory does not easily account for great differences in levels of prejudice between groups of people who show substantially the same amount of "authoritarianism." Two matched groups of American whites, one Southern, one Northern, for example, showed nearly identical levels of authoritarianism, but widely discrepant levels of anti-Negro prejudice.

The relation between personality traits and prejudice seems to vary considerably depending on the social climate. Thus, in societies in which racial discrimination is the norm, as in South Africa, even persons who show little authoritarianism will typically be prejudiced. (This has been referred to as "conformity prejudice" to distinguish such attitudes from the more psychopathological forms of bigotry.) It may also be that, in societies that do not openly condone discrimination, the authoritarian personality is more apt than other personalities to be a bigot. Thus, although attempts to apply this psychological theory at the societal level—as, for example, by explaining the rise of Nazism through the authoritarian structure of the German family—are at best unconvincing, the theory seems able to account for individual differences in levels of prejudice within the same cultural group.

Economic influences. Social scientists have stressed a number of social variables that may be causally linked with prejudice and discrimination. Economic factors are of unquestioned importance. Marxian writers have interpreted racism as a rationalization for slavery and colonialism and also as a means of splitting the working class along colour lines and of deflecting attention from the central reality of class conflict to the ancillary problem of "race." It is certainly no accident that racism flourished at the time of the second great wave of European colonial expansion and the scramble for Africa, and the ideology of colonialism and the white man's burden was often expressed in racist terms. It is also true that racism provided ideological justification for slavery. But slavery antedates the development of racism in Western societies, and some slave societies, notably those of Latin America, have been much less racist than those of English or Dutch origin in North America, the Caribbean, and South Africa.

Non-Marxian writers have stressed the importance of other economic factors, such as competition for jobs, as problems aggravating interethnic or interracial conflicts. Racial segregation in urban ghettos and its resultant conflicts have been linked with the pattern of absentee landlordism in slums and the control of the housing market by realtors and investors. In blatantly discriminatory societies like South Africa the link between racist legislation and the economic interests and privileges of the white minority is obvious. A common form of economic determinism in racial and ethnic relations is the "middleman syndrome." A number of scattered minorities, often of alien origin, are subjected to much the same syndrome of prejudice. Indians in east and South Africa, Lebanese in west Africa, Chinese in Southeast Asia, and Jews in various countries, insofar as they have concentrated in retail trade and middle-level white-collar occupations, have often been the victims of an "anti-Semitic" type of prejudice, being accused of clannishness, rapaciousness, underhandedness, dishonesty, stinginess, and exploitation of the indigenous majority. These stereotypes are due in part to the envy felt toward an alien minority whose living standard is often above the average in the host country.

Minority
middleman
as bias
target

Relations of production—that is, the economic and class position of the constituent racial or ethnic groups within a society—are always of crucial importance in determining the type of relations that exist between these groups. Equally important and closely related to the relations of production are the relations of power between groups. For example, a multiracial society in which a minority exerts a clear domination over a majority, such as is characteristic of colonial regimes, makes for a type of race relations different from that of an ostensibly democratic society such as the United States, where a majority discriminates against a minority. An agrarian, patriarchal state such as characterized the slave regimes of the Western Hemisphere or the Boer republics in the 19th century makes for a type of race relations different from that of a highly urbanized and industrialized country like modern South Africa. The techniques of racial domination, for example, are radically different in a slave plantation and in an urban ghetto. Relations of power determine the legal structure of the society, and many studies have demonstrated the importance of legislation in changing a society either in a more egalitarian direction, as in the United States since the 1940s, or in a more racist one as in South Africa under apartheid.

Religion's role in racism. Religion has also been shown to be related to the amount of prejudice and discrimination. There is an undeniable difference between the more racially tolerant Catholic countries of Europe and their colonial extensions and the more racist Protestant countries. The Catholic Church has frequently taken a more universalistic position and rejected racism, whereas many Protestant denominations, especially the more fundamentalistic and puritanical ones, have often interpreted the Scriptures in a racist fashion. The role of the Dutch Reformed churches in South Africa in supporting apartheid as the will of God is well known, and Protestant Fundamentalism in the United States has sometimes also been deeply racist.

Demographic influences. Numerous racial studies have stressed the importance of demographic and ecological factors in intergroup relations. The demographic ratios between groups, the number of distinguishable groups, and the geographical concentration of these groups within a country all affect the system of intergroup relations. Thus situations in which the dominant group is a majority are different from those in which it is a small minority, and different again from those in which two groups of nearly equal size maintain a competitive balance of power. It also matters whether groups are evenly spread throughout a country or heavily concentrated in a given province or packed in urban ghettos. And the balance of power of course varies according to whether there are two, three, four, or many groups.

Social and cultural factors. The degree of cultural differences between groups is also a relevant factor. Ethnic or racial prejudice is often exacerbated by barriers of language or customs. Acute racial conflicts, however, can also exist between groups that are culturally nearly identical—as between white and black Americans.

Situations of rapid change are often marked by an intensification of racial and ethnic conflicts because one aspect of change is frequently an alteration in the traditional relations between groups. Such changes are commonly perceived as a collective threat, and this makes for outbursts of conflict. The rapid immigration of racially distinctive groups, as in post-World War II Britain, is a case in point. The rapid influx of a hitherto rural group into urban areas can have the same effect, as shown in the urbanization of Afro-Americans in the last few decades.

Conflict
from
sudden
social
change

International conflict can have domestic repercussions in intergroup relations, as witnessed by the wave of anti-Japanese racism in the United States during World War II or by the wave of anti-Algerian feelings in France during the Algerian war of independence.

The complex interplay of these numerous social and psychological factors on racial and ethnic relations makes predictions and generalizations hazardous. No single factor ever accounts for more than a fraction of the phenomenon to be explained, and the relative importance of various factors is often difficult to establish.

THE EFFECTS OF RACIAL DISCRIMINATION

The consequences of racial discrimination are as diverse as its causes. It is useful to distinguish here between psychological and social effects and between those effects on the group that discriminates, on the group that is discriminated against, and on the society at large.

Effects on victims. There is little doubt that psychologically racism is harmful to its victims. The most profound effect of racism associated with situations of extreme degradation (such as is found under slavery or in concentration camps or in racist states like South Africa) is the acceptance by the oppressed group of the dominant group's definition of the situation. This is the phenomenon of self-hatred found, for example, in cases of Jewish anti-Semitism or in the acceptance by blacks of white aesthetic criteria such as the desirability of having straight hair or a light skin. Self-hatred is often accompanied by neurotic symptoms of apathy, anxiety, and depression or by forms of self-destructive escapist reactions such as alcoholism or drug-addiction or, in extreme cases, by paranoid, schizophrenic, or manic-depressive psychoses. In such situations of extreme degradation then, the oppressed group frequently reacts in an "intropunitive" fashion; that is, it turns its frustrations inwardly against the self or the ingroup at large. At the social level, this intropunitiveness takes the form of predatory crimes by an organized underground against the oppressed group. Racial ghettos in the U.S. and South Africa, for example, have very high rates of crimes committed by blacks against other blacks, a phenomenon that is encouraged by the disinterest of the police in providing adequate protection.

Turning against the dominant group

When victimized groups do not accept their inferior status or when conditions improve to such a degree that they regain self-respect and conceive of the possibility of changing the status quo, frustration turns outward as hostility and aggression against the dominant group. Paradoxical as it seems, situations of open racial conflict, like other forms of revolution, are generally associated with periods of both relative and absolute improvement in the position of the subjugated groups. Hostility toward the dominant group may take many forms, ranging from nonviolent passive resistance to apolitical crime and politically inspired guerrilla warfare. The search for an identity independent of the dominant group's definition of the situation is often only a preliminary step to concerted political action, frequently violent, to try to overthrow the existing racial order. Group cohesiveness and political militancy replace apathy and self-hatred.

Effects on the dominant group. From the perspective of the dominant group, the effects of racism are more mixed. Psychologically, racism warps the personality of the oppressor as it does that of the oppressed, though probably in a less devastating way. Racially prejudiced persons living in highly racist societies can behave "normally" in situations not involving race; yet racism, by erecting an irrelevant and artificial barrier between people, strains relations and distorts social perception in the dominant group. Social consequences of racism for the dominant group can vary widely. Generally, in colonial-type situations in which the dominant group is a minority and has entrenched itself in an economically and politically privileged position, the benefits it derives from racism can be considerable. Thus the artificially high standard of living of the white South African population probably can only be maintained by the elaborate apparatus of racial laws limiting the freedom of movement, organization, and employment of the nonwhite majority. Some of those material benefits for the whites are diverted to the maintenance of the repressive apparatus; however, the economic balance remains positive for the whites, even though the total economic cost of segregation for the society as a whole is quite high.

In situations in which the dominant group is in large majority, as in the United States and Australia, the economic benefits of racial discrimination to the dominant group as a whole tend to be much more marginal and are often overshadowed by the costs in conflict, violence, and lost productivity. There are also situations in which the economic effects of racism may be opposite for different segments of the racially dominant group; in the antebellum South, for example, slavery benefitted the slave- and land-owning aristocracy but not the white yeomen and workers.

If one attempts to assess the overall effects of racism on an entire society, one must conclude that they are negative insofar as racism erects an artificial barrier to the full use of talents and often generates destructive conflicts.

THE REDUCTION OF RACIAL DISCRIMINATION AND FUTURE PROSPECTS

How discrimination is practiced. In practical terms. situations of racial discrimination and conflict may be reduced to three broad types:

Segregationist or apartheid societies. In societies like South Africa or Rhodesia, it is clearly in the collective interests of the racially dominant group (typically a minority) to maintain the status quo. The reduction of racial discrimination in such cases can come about through a drastic change in the power structure (as it did in the U.S. South as a result of the Civil War) or it can come about in evolutionary fashion (as it did in colonies ruled by Great Britain); but, in any case, the new ruling group no longer feels the need to defend its interests by maintaining a racial caste system.

Pluralistic societies. Those societies in which several racial groups, none of them clearly dominant, compete for power and economic resources (as in some parts of Africa and Asia) probably offer the widest range of alternatives for the reduction of racial tensions. Racial and cultural assimilation may reduce differences and blur old lines of cleavage. Or a more amicable modus vivendi may develop from changes in government policy, the spread of a broader nationalist ideology, or changes in the economic and political structure making for a more democratic society. When conflicts escalate to chronic violence and civil war, political partition and emigration are other possibilities for reducing tensions.

Societies publicly committed to ending racial discrimination. In some societies, as in Australia and the United States, official government policy is often against racial discrimination and has to contend both with the inertia of conservative opinion from the dominant group and the "revolution of rising expectations" from the hitherto subordinated group.

Ways of reducing racial bias and inequities. There seem to be two basic approaches to the alleviation of racial tensions in this third type of society, assuming an official desire to do so. One is to attack racial prejudice by educating the public. There is abundant evidence that people's attitudes can be changed by propaganda and that behavioral changes follow changes in attitudes; but there is also evidence that as a method to bring about rapid social change in a conservative population, this approach is relatively expensive and ineffective. In terms of practical consequences, it may be more important to reduce racial discrimination than racial prejudice, and, hence, it follows that the strategy of attacking discrimination is frequently more directly effective.

Dilemma of means and ends

Legislation is undoubtedly one of the main ways of destroying racial discrimination, provided it is followed by forceful implementation. This often entails a dilemma of means and ends: the achievement of a more democratic society may imply the use of force against a majority—a policy that democratically elected governments are naturally loath to adopt. Consequently, attempts to outlaw various aspects of racial discrimination (for example, in housing and employment) have frequently failed to bring about the rapid change that was intended because of a lack of determined implementation. The delays in school integration in the United States can be considered a case in point. By contrast, integration in the U.S. armed services has been more rapid and thorough, largely because it was done autocratically.

At the economic level, it is clear that the profit system of production helps to perpetuate disabilities for racial minorities. High returns on slum property, for example, are one of the factors making for the persistence of racial ghettos in the United States. Racial discrimination in housing creates an artificial scarcity of housing for blacks, who are thus forced to pay higher rents than their white counterparts. This makes for high returns on investments in the black areas that, because of high population density and municipal neglect, quickly become slums. Other economic systems, such as the "consumer credit," are also known to discriminate against the poor and thus to affect racial minorities disproportionately. Reforms in the economy might be effective in reducing racial discrimination, but the realistic prospect of such reforms in the United States is limited.

In the last analysis, the most effective way of reducing racial discrimination may be militancy on the part of the groups that are discriminated against. When such groups are large enough to affect the outcome of elections and when the political system is democratic enough to give the voters some real alternatives, political action can be effective within the constitutional framework. When such conditions are lacking, political action ranging from civil disobedience to guerrilla warfare has been known to bring about important changes.

BIBLIOGRAPHY. T.W. ADORNO *et al., The Authoritarian Personality* (1950), an influential study of the psychodynamics of prejudice from a psychoanalytical perspective; G.W. ALLPORT, *The Nature of Prejudice* (1954), a standard text by a social psychologist; B.N. COLBY and P.L. VAN DEN BERGHE, "Ethnic Relations in Southeastern Mexico," *American Anthropologist*, 63:772–792 (Aug. 1961), a description of a system of ethnic relations between Maya Indian groups and Spanish-speaking ladinos; O.C. COX, *Caste, Class and Race* (1948), a detailed critique of the field of race and ethnic relations by a Marxian sociologist; A.W. DAVIS, B.B. GARDNER, and M.R. GARDNER, *Deep South* (1941), a classic monograph on race relations in the Southern United States, with an influential theoretical introduction by W.L. Warner; C.W. DE KIEWIET, *A History of South Africa, Social and Economic* (1941), the standard social history of South Africa; J. DOLLARD, *Caste and Class in a Southern Town* (1937), a monograph on race and class in the Southern United States, and (with others), *Frustration and Aggression* (1939), the classic theoretical statement by psychologists that frustration leads to the displacement of aggression on ethnic and racial scapegoats; ST. CLAIR DRAKE and H.R. CAYTON, *Black Metropolis* (1945), the classic study of the urban black ghetto of Chicago; W.E.B. DU BOIS, *Black Reconstruction* (1935), a major work by one of the most influential early leaders of Afro-American freedom; E. FRANKLIN FRAZIER, *Black Bourgeoisie* (1957), a critical statement about the character of the black middle class in the U.S., and *Race and Culture Contacts in the Modern World* (1957), an account of race relations in various parts of the world; G. FREYRE, *The Masters and the Slaves*, rev. ed. (1964), a classic work on Brazilian slavery; R. HOFSTADTER, *Social Darwinism in American Thought*, rev. ed. (1959), a study of the development of racism and social Darwinism in the U.S., JOHANNESBURG, S.A. INSTITUTE OF RACE RELATIONS, *A Survey of Race Relations in South Africa* (annual), a useful compilation of developments in South African race relations; L. KUPER, *An African Bourgeoisie* (1965), an account of race and class in South Africa, with emphasis on the position of the black middle class; OSCAR LEWIS, *Life in a Mexican Village: Tepoztlán Restudied* (1951), a more recent study of a small Mexican town first described by ROBERT REDFIELD in *Tepoztlán: A Mexican Village* (1941); A.W. LIND (ed.), *Race Relations in World Perspective* (1956), a collection of essays dealing with racial situations in various parts of the world; A. MARCHANT, *From Barter to Slavery* (1942), the early history of racial and cultural contacts in Brazil; L. MARQUARD, *The Peoples and Policies of South Africa* (1952), the most concise introduction to racial problems in South Africa; GUNNAR MYRDAL, *An American Dilemma* (1944), a monumental study of black America in the 1940s; R.E. PARK, *Race and Culture* (1950), a collection of sociological essays on race and culture; T.F. PETTIGREW, *A Profile of the Negro American* (1964), a useful summary of studies on black Americans; D. PIERSON, *Negroes in Brazil: A Study of Race Contact at Bahia* (1942), a pioneer study of Afro-Brazilians in the North East; T. SHIBUTANI and K.M. KWAN, *Ethnic Stratification: A Comparative Approach* (1965), a text on comparative race and ethnic relations; G.E. SIMPSON and J.M. YINGER, *Racial and Cultural Minorities*, 3rd ed. (1965), a standard text stressing North America; K.M. STAMPP, *The Peculiar Institution: Slavery in the Ante-Bellum South* (1964), a good account of antebellum slavery in the U.S.; F. TANNENBAUM, *Slave and Citizen: The Negro in the Americas* (1947), a comparative study of slavery in the Americas; P.L. VAN DEN BERGHE, *Race and Racism* (1967), a text on comparative race relations in Mexico, Brazil, the U.S., and South Africa, and *South Africa: A Study in Conflict* (1965), a comprehensive sociological analysis of South Africa; C. WAGLEY, *Amazon Town: A Study of Man in the Tropics* (1964), a community study of a multiracial town in the Amazon basin, and (ed.), *Race and Class in Rural Brazil* (1952), a series of studies of race relations in various regions of Brazil; and R.M. WILLIAMS *et. al., Strangers Next Door: Ethnic Relations in American Communities* (1964), a compilation of research findings on race and ethnic relations in the U.S.

(P.L.v.d.B.)

Rackets

Rackets (racquets) and squash rackets are games played with a ball and a strung racket in an enclosed court, all four walls of which are used in play. Rackets, the older game, is played with a hard ball; whereas squash, or squash rackets, is played with a soft ball on a smaller court.

History. It was once a common notion that rackets originated in the debtors' section of Fleet Prison in England early in the 19th century. Charles Dickens in his novel *The Pickwick Papers* (1836–37) describes a court in which the inmates whiled away their time. The game might, however, and more accurately, be said to be an outgrowth of the ball games of the ancient Greeks and Romans. It has features in common with palone, pelota, and other Basque games. Historians of real tennis (court tennis) have traced the origin of rackets to the wall of a real tennis court, and Robert W. Henderson, in his researches into the antiquity of athletic games, found mention of rackets (in the singular form, racket, as then used) as early as 1529 in *The Complaint of Schir David Lindesay*. Most now place the origin of rackets in real tennis, quoting J.R. Atkins' opinion in *The Book of Racquets* (1872) that "both games (rackets and real tennis) have so much in common that it is impossible to separate them historically; for practical purposes we must regard them as identical."

In its beginnings, rackets was played in rather formless fashion without set rules. In Fleet Prison the game was well established by the middle of the 18th century, and in the new Fleet of 1782 it achieved such popularity that its fame spread to taverns and other public houses. Robert Mackey, an inmate of Fleet, is listed as the first "world" champion or at least as the first claimant of the title in 1820.

It was with its introduction into Harrow School in 1822 that rackets achieved respectability and was enclosed within four walls. The first roofed-in structure is believed to have been a court built at Woolwich by the Royal Artillery in the 1840s. The building of old Prince's Club in London in 1853 is regarded as marking the beginning of a new era in which rackets became the game of the clubs, military services, and universities.

Rackets flourished in the 1860s and 1870s. Earlier than this it had been introduced into Canada and the United States, and it spread to India, Malta, and Argentina. Queen's Club was opened in London in 1887 and became the headquarters of the game. The next year the Amateur Championships were started there and the Amateur Doubles began in 1890. The rules of the game were drawn up for the first time in 1890 by tennis historian Julian Marshall and rackets authority Major Spens. The Tennis, Rackets and Fives Association was formed in 1907 to govern the sport. During and following World War I, private courts closed and rackets play declined. The expense of building courts and playing the game and the rising popularity of squash rackets brought about a great reduction in the number of rackets players, except in the public schools. Nevertheless, the game continued to be played. In 1928 a British team travelled to the United States to inaugurate the International Racquets Cup matches, which still continue from time to time.

Rackets champions

The world rackets championship, which is decided by a challenge match, has been dominated by English players, although India and the United States have also produced outstanding players. Peter Latham, an English professional, is generally rated the greatest of rackets players. (Professionals, in rackets and squash rackets, are players who are paid to teach the games.) Latham was world champion from 1887 to 1902, when he resigned, and was also a great player of real tennis. The foremost English amateurs have included Sir William Hart-Dyke, the first amateur to hold the world championship (1862); and Geoffrey Atkins, world champion from 1954 to 1970, who excelled Latham's record of reigning for 15 years. Atkins is rated by some as the greatest of all amateurs.

Squash rackets. Squash rackets, often called squash, is believed to have originated around the middle of the 19th century at Harrow School. Students unable to get into the rackets court took their exercise hitting an india-rubber ball, which squashed when hit against a wall. Toward the end of the century, "almost every boarding school had its miniature racket court in which a game with a soft, india-rubber ball is played," wrote E.O. Pleydell-Bouverie in the Badminton Library, a British series on sports published in 1890. In the 1890s private courts were built and, after the turn of the century, club courts appeared at Bath, Queen's, and the Marylebone Cricket Club.

It was not until after World War I that squash rackets caught on, and in the 1920s the game spread rapidly, becoming more popular than its parent game rackets. Many courts were built in clubs, schools and colleges, and privately. Rules were formulated, the English national association was organized, and dimensions of the court were established, along with regulations in regard to the ball and racket. Many competitions were inaugurated: the Professional Championship in 1920, the Amateur Championship for men and women in 1922, and the Open Championship in 1930. International competition with the United States began with the sending of a team to America in 1924, and a member of the team, Captain Gerald Robarts, won the U.S. crown. Competition between the two countries, however, has been hampered by differences in the ball, the court, and methods of scoring. In 1933 the Wolfe-Noel Cup was put up for an international women's series with the United States.

In the United States the game played in the early years was actually squash tennis, using a lawn tennis ball and tennis racket. Around 1915 a soft, india-rubber ball and the racket with a small, circular head and long, slender handle were adopted instead. Only in Philadelphia was squash rackets the original game. Squash tennis has disappeared from most cities and continues to have a following only in New York.

From England the game spread throughout the Empire —to Canada, India, Australia, and South Africa. In the late 1920s King Prajahipek said to a newspaper correspondent in Bangkok, "Squash is nothing less than a Godsend to a country like Siam." Today squash is played in some 40 countries, including France, Germany, Denmark, Sweden, Belgium, The Netherlands, Egypt, Mexico, Pakistan, India, Thailand, Hong Kong, Kenya, and New Zealand. An International Squash Rackets Federation promotes the game and coordinates tours and championships between nations.

Outstanding squash players have included F.D. Amr Bey, an Egyptian amateur who won several British open titles in the 1930s; the Khans of Pakistan, a family of professionals who dominated open play in the 1950s and 1960s; and Janet Morgan, English women's champion from 1950 to 1959, winner of U.S. and Australian titles, and member of Britain's Wolfe-Noel team from 1949 to 1959.

Rackets. No dimensions are specified for the rackets themselves, which are made of ash and average 76 cm (27 in.) long and 255 g (9 oz) in weight. The head, strung with catgut, is usually 178–203 mm (7 or 8 in.) in diameter. The ball, which has a renewable covering of adhesive tape, is 2.54 cm (1 in.) in diameter and weighs 28.35 g (1 oz).

The court. Most courts are about 18 m (60 ft.) long by 9 m (30 ft.) wide and accommodate both the singles and doubles (four-handed) games. Courts have four walls. The roof, where skylights or other lighting is placed, is out-of-bounds for play; in India courts were left unroofed. The cement floor and walls must be perfectly smooth and very hard since the faster the ball travels the better the game. Front and side walls are about 9 m (30 ft.) high, the back wall being about half that height with a spectators' gallery and marker's, or scorer's, box above it. The court is entered by a door in the centre of and flush with the back wall. On the front wall is fixed a wooden board, the upper edge of which, 0.68 m (27 in.) from the floor, constitutes the play line; 2.93 m (9 ft 7½ in.) from the floor a second line called the cut line or service line is marked. On the floor, 10.92 m (35 ft 10 in.) from the front wall and parallel to it, the short line runs from wall to wall. From the centre of the short line to the centre of the back wall, the fault line divides the back court into two rectangular service courts. Against the side walls and separated from the service courts by the short line are the service boxes.

*The **game.*** Rackets may be played by two persons (singles) or four persons playing two against two (doubles). The players must return the ball either before it reaches the ground or on its first bound so that it strikes the front wall above the play line (or service line in the case of a serve) and returns into the court and continue to do so alternately (either player of each in doubles) until one player fails to make a valid return and loses the stroke. The ball must not go out of court (into the gallery or roof of the court) or touch the players' clothing or person. Hard, low hitting close along the side wall is the essence of the game, with cutting, volleying, half-volleying, drop shots, and angled shots also in the repertory. In the four-handed game (doubles) one of each set of partners takes the right-hand side of the court and his partner the left. The game consists of 15 points, called aces. Points can be scored only by the hand-in (the player, or side, having the service), and the hand-out (side receiving service) must therefore win a stroke or strokes to obtain service before he or they can score an ace. In doubles each of the partners serves in turn, and both must be ousted before their opponents obtain the service. In the first exchange of each game, however, only one partner of each side has service.

Serving the ball

The server, with at least one foot inside the service box, serves the ball as in tennis, but directly to the front wall above the service line so that it rebounds and hits the floor within the service court on the opposite side, permissibly striking the side wall, back wall, or both before or after touching the floor. The serve is a fault if the ball (1) strikes the front wall below the service line; (2) touches the floor on the first bounce in front of the short line; or (3) first touches the floor in the wrong court. If the receiving player chooses to take a faulty first serve, play proceeds as if the serve had been good; otherwise the server must serve again; if he serves a second fault he loses his service to his partner or opponent, as the case may be. A serve that makes the ball strike the board or the floor before reaching the front wall or that sends it out of court counts the same as two consecutive faults: it costs the server his innings. In the United States and Canada only one serve is permitted.

If the player receiving service succeeds in returning the serve, the rally proceeds. If he fails in the rally (or in receiving service), the server scores a point and the side that first scores 15 points wins the game. When, however, the scores reaches 13–all, the receiving side may, before the next serve is delivered, declare that he elects to set the game either to 5 or 3, making the game 18 or 16 points, whichever he prefers; and similarly when the score stands at 14–all, he may set the game to 3 (game 17).

"Lets"

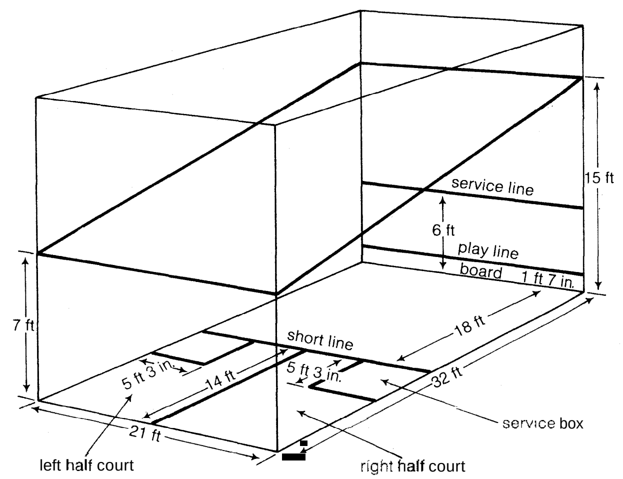
It is the player's first duty to give the opponent full room for his stroke, but it is not always easy and sometimes, especially in doubles, absolutely impossible not to obstruct him. The rules, therefore, carefully provide for "lets." When in matches a let is claimed by any one of the players and allowed by the referee, the service or rally counts for nothing and the server serves again from the same service box.

The server in possession at the end of a game continues to serve in the new game, subject as before to the rule limiting the first innings of a doubles game to a single hand. The usual number of games in matches is five for singles and seven for doubles. In matches where there is a referee, there is an appeal to him from the marker's decision but no appeal is allowed if a foot fault is called.

Squash rackets. Squash is played on the same principle as rackets. The rules are in most particulars similar, but the scoring is different. In England, hand-in only can win the point. A game consists of 9 points; if, however, the score becomes 8–all, hand-out has the option of a set to two—making the game 10 points. In the United States a game consists of 15 points, and a point is scored for each rally irrespective of which player is serving; at 13–all, hand-out may set to 5 points (game 18), to 3 points (game 16), or no set (game 15). In doubles games as played in the U.S. and Canada, a rubber consists of the best three out of five games.

The squash racket is similar to that used in the game of rackets, but the handle is shorter, and the U.S. racket is heavier than the British. The ball is of rubber or rubber and butyl composition. The ball used in the U.S. is larger and faster than that used in England.

The court. The standard British court has four walls, which are of wood or composition, and can either be covered or uncovered. The dimensions and markings of the British court are shown in the accompanying diagram.



British squash racquets court

The board, or telltale, is a strip of sheet metal or other resonant material that produces a clearly different sound when the part of the front wall "out-of-play" is hit. The standard U.S. court is considerably narrower than the English, being 5.6 m (18 ft 6 in.) wide, and some other dimensions are also slightly different. Doubles courts are 13.7 m (45 ft) by 7.6 m (25 ft).

Squash tennis. Squash tennis is played in the same court as squash rackets. There are minor differences in the line markings, such as the "out" line on the back wall being two feet lower than for squash rackets. Players use a lawn-tennis racket, an inch shorter in the handle, and a green, pressurized ball, similar to but slightly smaller than the lawn-tennis ball. Scoring is the same as in American squash rackets. Points were scored only by the server until a change in 1954 allowed hand-out to score as well.

Squash tennis makes fewer demands than squash rackets on the legs in pursuing the ball, but puts a greater premium on agility and quickness of foot and reflexes in turning and spinning. The ball caroms off the front, side, and back walls in dizzying action. It is so volatile that it may rebound from the front wall to the back wall and again to the front wall. Squash rackets is largely a game of wrist and touch, as well as of the arm. The squash tennis stroke is more comparable to the free-swinging lawn-tennis drive. The volley—returning the ball before it touches the floor—is important in both squash games.

BIBLIOGRAPHY

General: E.B. NOEL and C.N. BRUCE, *First Steps to Rackets* (1925), a brief history of the game, instruction in technique and tactics, advice on training, and a record of championships; A. DANZIG, *The Racquet Game* (1930), a history of rackets, squash rackets, squash tennis, and court tennis, with lists of championship winners.

Instructions: J. BARNABY, *Winning Squash Racquets* (1979), a comprehensive manual, includes points on teaching and brief sketches of great players; B. CONSTABLE, N. PECK and D. WHITE, *Squash Basics for Men and Women* (1979), instruction especially useful for beginners, with a glossary and official rules; A. MOLLOY, JR., *Winning Squash* (1978), discussion includes thorough explanation of doubles game; A.M. POTTER, *Squash Racquets*, 2nd ed. (1966), photos and diagrams of shots, with U.S. and British playing rules; J. SKILLMAN, *Squash Racquets*, 2nd ed. (1964, reissued 1980), on the origin of the game and court specifications; J. TRUBY, JR., *The Science and Strategy of Squash* (1975), thorough coverage of both theoretical and practical aspects of the game: M. VARNER and N.B. BRAMALL, *Squash Racquets* (1967), includes the language and lore of squash, and both official and unwritten rules.

See also *Official Rules of Sports and Games* (biennial); NORTH AMERICAN RACQUETS ASSOCIATION, *The Laws of Racquets* (n.d.); U.S. SQUASH RACQUETS ASSOCIATION, *Official Yearbook* (annual); and the U.S. WOMEN'S SQUASH RACQUETS ASSOCIATION, *Handbook* (quadrennial). These publications give tournament results, lists of champions, schedules, and rules.

(A.D.)

Radar

Radar is an electronic system that uses radio waves to detect objects invisible to the unaided eye because of distance, darkness, or cloud cover. Radar can, as well, determine the position of an object, its distance from the observing station, and, if the object is moving, its speed and direction of travel. The word radar is an acronym derived from the phrase radio detection and ranging. Although developed as a wartime device to detect enemy airplanes, radar has many nonmilitary applications. The equipment used for radar is called a radar set, radar system, or simply a radar.

BASIC PRINCIPALS

Echoes and targets. When radio waves radiated from a transmitting antenna are interrupted by any object such as a ship, airplane, or even a mountain, part of the energy is reflected back to the receiver. The reflection is called an echo, and the reflecting object is called a target. Echoes from desired targets are referred to as target signals or simply signals, whereas echoes from masses that make it difficult to detect the wanted signals are called clutter.

Target distance

The distance to a target is determined by the time required for the radio wave to travel from a transmitter to the target and back to a receiver. The calculation is possible because radio waves travel at a known speed, that of light, which for purposes of approximation may be taken as 1,000 feet (300 metres) per microsecond (millionth of a second).

Pulse modulation. If the transmitter were operating continuously, reflected signals, or echoes, would arrive continuously, and it would be impossible to associate a certain part of the echo with a specific part of the transmitted signal. The signal must therefore be labelled. One method of doing this is to emit the radar signal in short, high-power bursts or pulses; this method is called pulse modulation. The echoes then return as short pulses. If conditions are arranged so that the transmitter is off (not transmitting) during reception of a pulse, the received signal can be associated with a specific transmitted pulse. Radar using pulse modulation, referred to as pulse radar, was the first type used and is still the most common.

Since a pulse radar's transmitter and receiver are operating at different times, a single antenna system can be used for both. The antenna is connected to the transmitter during the short transmission pulse and is then switched to the receiver during the interval between pulses; this is accomplished with a device called a duplexer.

From Gershon J Wheeler, *Radar Fundamentals* (1967) Prentice-Hall

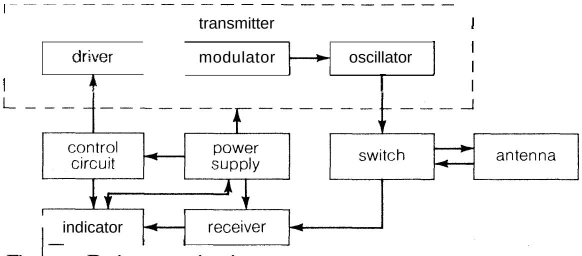


Figure 1: Basic radar circuit.

A block diagram of a basic radar circuit is shown in Figure 1. During transmission, a short pulse of radio-frequency energy is fed from the transmitter to the antenna. After the pulse has been transmitted, the antenna is switched to the receiver to detect echoes. The detected echo signal is fed to the indicator, which is usually some form of visual display, such as a cathode-ray (television) tube, that enables the operator to measure the time intervals that occur between the transmitted pulse and the reception of its echo.

Since the output tube is actually operating only a fraction of the time, pulsing the transmitter permits it to provide higher peak power than would be possible if it were operating continuously. Figure 2 shows two successive transmitted pulses, each lasting a time designated by the Greek letter tau (τ). The height of the pulse (P_P) indicates the power level.

Power output and duty cycle. Operating continuously with the same total energy, a transmitter would radiate a much lower peak power signal. The power level would be the average of the energy in the pulse taken over the entire interval up to the following pulse. This is called the average power (P_A), indicated by the dotted line in the figure. The average power is limited by the ability of the circuits to dissipate the heat produced.

If the pulse repetition frequency (or repetition rate) is designated f_r (in pulses per second), then the interval in seconds between pulses is the reciprocal of this, or $1/f_r$, as shown in Figure 2. For example, if the repetition rate is 2,000 per second, the pulse interval is 1/2,000 of a second, or 500 microseconds. The product of pulse width times the repetition frequency is called the duty cycle (τ f,) and is defined as the proportion of the total time that a device operates. Thus, if a radio transmitter has a 10 percent duty cycle, it transmits 10 percent of the time and is silent 90 percent of the time. The average power is equal to the product of the peak power and the duty cycle.

From Gershon J. Wheeler, *Radar Fundamentals* (1967); Prentice-Hall

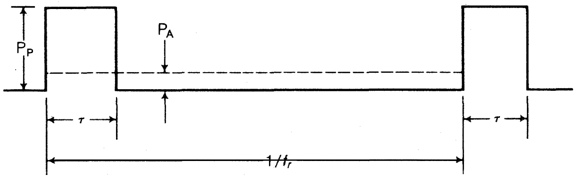


Figure 2: Pulse envelope (see text).

With a repetition rate of 2,000 pulses per second, the pulses are 500 microseconds apart. If echoes are received 100 microseconds after each pulse, they could be caused by a target 100 microseconds away or by one 600 microseconds away. Assuming approximately 12 microseconds per mile, this means that the target could be about eight miles (13 kilometres) or 50 miles (80 kilometres) away. To overcome this ambiguity, the repetition frequency must be so chosen that echoes from all detectable targets for one pulse will appear before the next pulse is emitted. Thus, if it is known that no target exists beyond 25 miles (40 kilometres) or if receiver sensitivity does not permit detection beyond this point, then a pulse interval of 300 microseconds is acceptable. If on the other hand targets as far as 50 miles away may be detected, then the pulse interval must exceed 600 microseconds.

Choosing the repetition frequency

Radar frequencies. A radar transmitter emits radio waves of a given frequency, measured in hertz, or cycles per second; this is called the carrier frequency of the transmitter. Choice of the carrier frequency is influenced by available transmitting equipment, power and range desired, and spreading (propagation) characteristics of radio waves at different frequencies. Most radars operate in the microu ave region of the electromagnetic spectrum; *i.e.*, between 1,000 and 35,000 megahertz (1,000,000,000 and 35,000,000,000 cycles per second). During World War II, for security reasons, radar frequencies were given code designations. Some of these codes are still in use today. The frequencies near 10,000 megahertz are called X-band; near 3,000 megahertz, S-band; near 1,000 megahertz, L-band; and near 6,000 megahertz, C-band. The band limits are not clearly defined, so that 2,000 megahertz. for example, may be called high L or low S band.

Doppler shift. When a target is moving toward or away from the transmitter, the frequency of the received echo differs from that of the transmitted carrier. This change in frequency is called the Doppler shift and is positive (an increase in frequency) for a target approaching and negative (decrease in frequency) for a receding target. The amount of frequency change depends upon the speed at which the target is approaching or receding from the transmitter. By careful measurement of the change in frequency, it is possible to determine both the speed and direction of a target.

THE DEVELOPMENT OF RADAR

Radar cannot be attributed to a single inventor or even to an identifiable group of inventors. Its basic concepts have been understood as long as those of electromagnetic waves.

Early history. The first patent for a radar-like system was granted in several countries in 1904 to a German engineer named Christian Hiilsmeyer. Evolving from the search for a means of detecting radio waves from ships, his system worked, was demonstrated to the German Navy, but was never accepted. Technical limitations made his system ineffective at ranges exceeding one mile.

The principles used in Hulsmeyer's system were actually known much earlier through the experimental work of the English physicist Michael Faraday and the mathematical investigations of the Scottish physicist James Clerk Maxwell, who predicted the existence of radio waves and formulated the electromagnetic theory of light. The German physicist Heinrich Hertz tested Maxwell's theories experimentally and in 1886 succeeded in proving the existence of radio waves. He also proved that radio waves were similar to light waves and could be reflected from solid objects.

Contributions of Faraday and Maxwell

The use of a radio echo for detection was frequently mentioned in sceintific literature after Hertz's demonstration of radio-wave reflections; but the idea was not seriously considered until 1922, when the Italian engineer Gugiieimo Marconi presented a paper on radio detection, and the United States Naval Research Laboratory tested his idea experimentally. Using a five-metre continuous-wave (nonpulsed) radar with a separate receiver and transmitter, the researchers were able to detect a wooden ship passing between the receiver and transmitter. This type of radar is now called a bistatic continuous-wave radar, differentiated from monostatic continuous-wave radar, which positions both transmitter and receiver at the same site.

Pulse modulation as a means of measuring distance or range was first developed in the U.S. in 1925. Using a pulse technique to measure the height of the ionized layer of air high above the earth, called the ionosphere, this pulse-ranging method became the standard for ionospheric investigations all over the world. It was not, however, applied to radar for several years.

Developments before World War II. Radar research and development was conducted during the 1930s in Great Britain, France, Germany, and the U.S. An early development of radar for practical use took place in the United States in 1930, when a researcher working on direction-finding equipment noticed that the received signal increased whenever an airplane passed between the transmitter and receiver of his experimental apparatus. The Naval Research Laboratory followed up this lead and by 1932 could detect aircraft as far as 50 miles from the transmitter.

Bistatic continuous-wave radar could detect the presence of a target but not its location. Since pulse radar promised a solution to this problem, scientists began development work early in 1934 and, after many failures, produced a workable radar in April 1936. The range of this device was 2½ miles (4 kilometres); it used a five-microsecond pulse at a frequency of 28.3 megahertz. By the end of 1936 a range of 7 miles (11 kilometres) had been achieved, and by 1938 an anti-aircraft fire-control radar had been placed in operation.

By 1939, the United States had lost what lead it might have had to European countries more imminently threatened with war.

Though work in France originally had been aimed at nonmilitary applications, military development came first in Britain, Germany, and, as has been clear, the United States. The first French radar was an iceberg detector for ships, used aboard the ocean liner "Normandie." As war became imminent and radar work in France shifted toward aircraft detection, a bistatic continuous-wave system was developed at the National Radio Laboratories, and in 1939 the French began work on pulse radar.

In Germany, development of a ship detection system began early in the 1930s; work on aircraft detection was soon added; and by 1939, a system called Freya, for early warning of approaching aircraft, was in production. A ship detection system followed. By mid-1940, the Germans were using a 600-megahertz radar system (called Wiirzburg) that provided position information with sufficient accuracy to direct effective anti-aircraft fire, outperforming the radars in use in other countries at that time.

Early British radar

Radar development did not begin in Great Britain until the mid-1930s, after which, with strong financial support from the government, it proceeded at a rapid pace. By September 1935, ranges greater than 40 miles (64 kilometres) were achieved, and by 1938 a chain of radar stations was operational.

Though these ground-based chain radars could detect enemy aircraft, they were not sufficiently accurate to guide British fighter planes to a successful interception. Consequently, the British worked to develop a successful airborne aircraft-intercept radar, to be installed in fighter planes and capable of detecting surface vessels and even submarines. Special radars, designated as air-to-surface-vessel radars, were later developed with an improved capability for detecting surface vessels.

Developments during World War II. When a radar transmitter scans the sky, its radio signal can be compared to the light beam of a searchlight. Obviously, if a radar beam is too wide, it will not locate objects very precisely. Development work during this period was thus aimed at narrowing the radar beam.

For several reasons, investigators concluded that the most logical approach to narrowing radar beams was to employ high transmitter frequencies, preferably those in the microwave range; but these were not available, since no one knew how to build them. In 1939 the klystron, a type of microwave signal generator, became available for use in microwave receivers. Its power output was too low, however, for employment in a transmitter. In 1940, the highest power output achievable was about one watt.

In England, a breakthrough occurred in 1939 with the development of the multicavity magnetron, able to produce about 20,000 watts of power at 3,000 megahertz, a tremendous advance over the klystron. Modern radar was born with this tube, because it made microwave radar practical for the first time. The development of the magnetron and the fact that high-power microwave did indeed exist were both closely guarded secrets in England for more than a year.

In the U.S., important advances were also taking place. Early in November 1940, a group of U.S. scientists began work on military radar. Their first experiments were set up on a roof, with antennas pointed toward buildings on the Boston skyline; and on January 4, 1941, the first aircraft interception radar, using two parabolic (dish-shaped) antennas, detected the buildings and a month later detected and tracked an airplane. In March 1941, an experimental aircraft intercept radar installed in an aircraft flying over water detected a ship clearly and unmistakably; and the design of that radar became the basis of World War II systems for detecting surface vessels from the air, aircraft from a ship, and ground-based harbour defense. By May 1941, a radar set had been developed that tracked planes automatically; this was the ancestor of highly successful gunlaying radars of World War II.

Another program, aircraft navigation, also began early in 1941. Basic techniques were worked out before the end of the year. The name Loran (an acronym for long range navigation) was adopted for this system (see NAVIGATION).

Early in the war, the British had developed an airborne S-band (3,000 megahertz) radar for bombing, called the H2S. Later, U.S. and British laboratories cooperated in developing the H2X, an X-band (10,000 megahertz) bombing radar, whose shorter wavelength (three centimetres for the H2X versus ten centimetres for the H2S) permitted sharper beams and consequently more accurate bombing. After further cooperation with the U.S., the British developed gunlaying radars that surpassed the German Wiirzburg radars; these were placed in operation in 1944.

German radar

Although Germany had entered the war with superior long-wave radars such as the Freya and the Wiirzburg systems, the German high command, placing heavy emphasis on rocketry, reduced expenditures for radar development. In response to Allied air attacks, a hastily developed aircraft-interception set, called Lichtenstein, was developed. Thousands of Freyas, Wiirzburgs, and Lichtensteins were deployed and helped maintain German radar superiority until 1943. When the Allies began using microwave radars, they quickly gained supremacy. In February 1943, the Germans downed a British plane carrying an H2S radar and learned for the first time that the Allies were using microwave radar. Germany immediately set up a laboratory to develop magnetrons and microwave radar. In December 1943, a U.S. bomber carrying an H2X crashed in The Netherlands and Germany shifted its efforts to make an X-band radar similar to the H2X recovered from the crashed bomber. Before the end of the war, the Germans did succeed in building microwave radars, but these came too late to affect the war's outcome.

Japan, Italy, and Russia also worked on radar during the war, but in general their efforts lagged behind those of the U.S., Great Britain, and Germany. Work in Italy and Russia was minimal, with no original development. Japan began working on pulse radar in 1940, but the radars they deployed during the war were versions of the Wiirzburg system that had been shipped to them from Germany or versions of U.S. sets captured in battle.

Applications in World War II. The original purpose of radar was aircraft warning. Large ground-based installations detected approaching enemy bombers in sufficient time to enable fighter planes to engage them. Britain's chain radar played a major role in the Battle of Britain, detecting German planes even as they massed for flight on the continent. British fighter planes could thus be strategically placed in the air by the time the bombers arrived. Even though greatly outnumbered, the Royal Air Force repelled the German planes.

During daylight bombing raids, the British frequently observed their fighter plane and the enemy bomber simultaneously on the radar indicator. They developed the technique of directing their fighter plane from the ground to a position behind and close to the bomber; this enabled the fighter pilot to see and intercept the enemy. Called ground control of interception, the technique worked well during daytime raids but was almost impossible at night. When the Germans realized this, they began nightly bombing runs at the end of 1940. The British responded with aircraft-interceptor radar, which they placed in their fighter planes. Now a fighter pilot was directed by ground control to a point about a mile behind the enemy bomber, where he could see the enemy on the screen of his radar. Final interception was accomplished by use of radar, rather than by visual sighting. The first British aircraft-interceptor sets operated at 200 megahertz, and though not highly accurate, they helped keep the nighttime raids in check. By early 1943, microwave aircraft-interceptor radars were operational, and the Nazis stopped nighttime bombing raids completelv.

Like the Americans, the British modified their aircraft-interceptor radars to locate surface vessels and direct fire from ships against enemy vessels. Airborne surface-vessel detectors worked effectively, since a ship on the open sea cannot manoeuvre out of range of the bomber once it has been located. As the state of the art progressed, the airborne surface-vessel detection sets were reduced in size and weight, an important consideration in the design of an airborne package, and were able to pinpoint the location of a target more accurately. Though the improved accuracy was not necessary for airborne surface-vessel detection, it did make the radar practical for many other applications. The first sets designed specifically for airborne surface-vessel detection were flown early in 1942.

First U.S. Army radar

The first radar sets used by the U.S. Army were gun-layers that determined the direction and range of enemy planes at night, so that searchlights and then anti-aircraft guns could be directed at them. Fire-control radars, an outgrowth of the gunlayer sets, made it possible to direct guns, rather than searchlights, by synchronizing their direction with that of the radar antenna. Gunlayer sets were not accurate for fire control, however, because the radar beam was too wide, and not until the advent of microwave radar did fire-control radars become practical. The most effective fire-control set, the SCR-584, measured the angular position of a target plane with an accuracy of less than three minutes of arc and the range within 0.1 percent. It was also the first radar with automatic tracking; once pointed at a target, it could be locked on that target and made to follow it automatically.

The SCR-584 was also used for close control of aircraft from the ground. Since a friendly aircraft's position could be accurately determined by the device, radar operators were able to direct the pilots on missions, other than that of aircraft interception. Small bombers, for example, could be directed to an exact position over a target and instructed when to release their bombs, even under conditions of darkness or overcast. A special application of close control involves leading a pilot to a safe landing when weather conditions make the landing field invisible. An accurate short-range set, called ground-controlled approach radar, was developed for this purpose and first placed in operation in Europe early in 1945.

Since the trace on a radar screen does not differentiate between friendly and enemy planes, an additional means of identification is required. The British devised a system for use by the Allies throughout the war. All friendly planes carried a device called a transponder, which was simply a radar transmitter that responded only when triggered by a signal from the inquiring radar set. The signal from the transponder was coded to keep enemy planes from transmitting signals to foil identification. If an airplane showed a response to the inquiry, it was known to be friendly. But pilots occasionally forgot to turn on their transponders and the transponders sometimes failed to operate. In addition, when many planes were in the air, signals returning to the inquiring radar could not be associated with specific aircraft. Consequently, this identification system failed in air operations, though it was used successfully to identify friendly surface vessels. Rapid identification of aircraft is still an unsolved problem, especially when several are in view at one time.

Beacon devices

The transponder used in the identification system is one of a class of devices called beacons. A beacon is a transmitter triggered by a signal arriving from another transmitter. The main radar transmits a signal that is received by the beacon; the signal that returns is not a simple echo but a much stronger signal originating from the transmitter in the beacon. Thus, the beacon signal can be easily detected and separated from weaker echoes. Beacon bombing systems were developed by Great Britain and the United States to pinpoint the position of a bomber with great accuracy. In the first British system, called Oboe, the bomber carried the beacon. Two separate ground stations queried the beacon, and the exact position of the bomber was determined by triangulation, a method of determining position that involves the use of trigonometry (see NAVIGATION). Radio signals transmitted to the bomber from the two stations directed the pilot on his course and indicated when to release bombs. This system could handle only one bomber at a time and had a usable range of about 250 miles (400 kilometres).

A later British system, called H, used many ground-based beacons and an inquiring radar on each bomber. It could handle many bombers simultaneously. The U.S. developed a similar system called Shoran.

Though the extreme accuracy of Oboe, H, or Shoran was not needed for guiding a plane between airfields on a friendly mission, some sort of navigational device was desirable. Long-range systems using pulsed radar were eventually developed in both England and the United States, designated, respectively, as Gee and loran.

Until 1943 radar was used primarily for defense. With the introduction of the H2S radar, the Allies were armed with an offensive tool. The Royal Air Force used it in combination with Oboe to direct nightly bombing raids against Germany, while the U.S. Air Force used it in combination with Shoran for daylight raids. When the more accurate H2X was developed, it was installed in 12 planes for tests. Beginning in November 1943, the 12 planes made daily bombing raids against Germany for five months, permitting the Allied ground forces to advance without interruption.

In 1944 a microwave early-warning radar was placed in operation to monitor the invasion of Normandy. During the landing, the microwave early-warning system directed the fighting against the Nazi defenses, keeping track of landing ships and friendly bomber planes. Equipped with H2X radars, the bombers wiped out the enemy defenses. In 1944, the Germans began firing long-range V-1 rockets against London. Detected by microwave early warning, many were shot down. In November 1944, the Germans switched to the much faster V-2 rockets; these could be reliably followed only with the microwave early-warning radars, which tracked them in flight and determined the launching sites from their trajectories.

From the standpoint of offense, the three most important radars in the European conflict were the H2X, airborne bombing systems, microwave early warning, and the SCR-584 fire-control system. Improved versions of these were used in the Pacific theatre in what were almost routine bombings. The Japanese Navy was destroyed in battles in which pilots directed their bombs entirely by ıadar.

Developments after World War II. When World War II ended in 1945, military radar research and development work was drastically reduced in all countries. With the lifticg of military security after the war, however, scientists began experimenting with radar as a research tool. In 1946, the U.S. Army Signal Corps successfully detected radar echoes from the moon with a modified radar set. Though the experiment had little practical application at the time, it marked the beginning of the study of radar astronomy. Subsequently, radar signals were reflected off Venus (1958) and the Sun (1959) and their echoes detected. The techniques developed in radar astronomy proved valuable in tracking man-made satellites in space a few years later.

Non-military applications

Civilian, or nonmilitary, applications of radar increased as soon as the security ban was removed. Radar as a navigational aid and for collision avoidance became a standard accessory on ships and planes. Harbour surveillance sets now guide ships when visibility is poor.

Radar on planes detect storms and help pilots to avoid bad weather and a phenomenon known as clear-air turbulence. A familiar police application is radar that determines the speed of a moving automobile. Operating on the Doppler principle, this radar determines the speed of the car by the shift in frequency of the reflected signal. In general, radars for civilian use involve new applications rather than new developments or techniques.

When the U.S.S.R. tested an atomic bomb in 1949, the U.S. began to consider air defense an urgent problem again. Radar was reactivated, and contracts for radar development, design, and manufacture were given to companies with radar capabilities. During the Korean War, the U.S. found that its old radars were not effective against the new high-speed planes; new equipment and techniques were required. In the elapsed time between viewing and relaying information by radio, planes could move great distances at the new high speeds, and the relative positions would be changed.

It was obvious that radar and computer techniques had to be combined to permit automatic interception. Development of the airborne interceptor radar subsequently advanced to a point at which a computer could control the aiming and firing of guns aboard the plane.

In the United States and Canada, ground radars of advanced design were stationed across the country to detect planes approaching from any direction. Signals from these radars were transmitted to control computers that had been given information about the flight plans of all friendly aircraft. The computers determined which friendly planes should be assigned to particular missions and performed the ground-controlled intercept function. In addition, aircraft batteries and other ground installations received directions from the system, and ground-based missiles were brought under automatic control.

Early warning has always been a most important consideration, and with the higher speed of jets, the warning has to be even earlier to be useful.

To this end the United States has constructed distant early-warning stations across the northern portion of North America; this system is called the DEW line. These radars presumably would detect any supersonic plane coming toward North America on a polar route. In addition, man-made islands supporting early-warning radars were constructed on the Atlantic continental shelf of the United States.

DEW stations

When the intercontinental ballistic missile became a threat, it was recognized that the DEW line could not supply information on approaching missiles soon enough, because missiles fly at speeds several times faster than the fastest jet plane. To counter this threat, a ballistic missile early-warning system was developed, consisting of three powerful radar installations—in England, Alaska, and Greenland. The first such radar, in Thule Greenland, was the largest and most powerful set in existence at that time. It had four antennas, each more than 300 feet (90 metres) wide and transmitters that generated millions of watts of peak power. It could detect a missile at 3,000 miles (4,800 kilometres), and its computers could almost instantaneously determine the missile's trajectory, target, and time of arrival. Since missile speeds approach 200 miles (320 kilometres) a minute, the radar could provide at least 15 minutes' warning of an approaching intercontinental ballistic missile.

At the other end of the scale, the military found many applications for small radars. Sets that could be carried on a man's back were developed to detect moving tanks or trucks at short distances up to three miles (five kilometres). Working on the Doppler principle, these radars translate the movement of the vehicle into a sound signal. Smaller sets capable of detecting a man crawling at distances up to 100 yards (90 metres) were built into soldiers' helmets.

An unusual airborne radar with military and civilian applications involves a radar-equipped plane flying a fixed course with a small antenna directed downward. The echoes received are detected and recorded on film, which is developed and fed to a computer. In effect, the computer assembles all the echoes and analyzes them as if they had been received by a huge antenna the length of the plane's flight. The result is a picture of the terrain with almost photographic clarity, even though the area may have been hidden by clouds or darkness during the flight.

The advances in radar since 1950 were made possible by a number of new developments. Although the Doppler principle was well known before World War II, radars making use of it were not developed until after the war. By detecting the frequency shift caused by a moving target, the target's echo could be separated from echoes arising from fixed ground masses (clutter). This resulted in moving-target-indicator radars in which a moving plane or other object could be detected in the presence of ground clutter. Later sets, both continuous-wave and pulsed, were developed to extract the Doppler shift as useful information so that the speed of the target could be determined. Moving-target-indicator radar was made possible by the invention of the high-powered klystron amplifier, which was more stable than the magnetron as a source of microwave power.

The development of more sensitive receivers also advanced the capabilities of radar. Low-noise-amplifier devices extended the range of the radar system by permitting the detection of weaker signals. Though these amplifiers were not originally invented for radar, they were quickly adopted by radar-system designers.

The capabilities of radar were further enhanced by advanced methods of information processing. While a radar antenna is pointing at a target, the transmitter emits a series of pulses in succession (pulse train), rather than a single pulse. A succession of echoes is received, one for each pulse in the train. If the energy in an echo is below the noise level of the receiver, it will not be detected by an ordinary receiver. Through a process called signal integration, however, these weak signals can be added together by a computer and "heard" by the receiver. Signal integration thus greatly increases the range of a radar set.

The pulse train

Early radar frequencies extended into the ultrahigh frequency (300–3,000 megahertz) range. Microwave radar was desirable but seemingly unattainable because of the lack of a suitable transmitter. When the magnetron became available, radar moved into the microwave region (3,000–300,000 megahertz) and remained there throughout World War II. With the advent of the ballistic missile threat, high power became an overriding consideration for early-warning radars. Since higher power was technically feasible at lower frequencies, radar began to move back to the ultrahigh frequency range. Ultrahigh frequency antennas were much larger than microwave antennas, however; had to be fixed in place; and could not be rotated. To fit the situation in which scanning was required, the antenna system was made to consist of an array of several hundred or even thousands of small antennas. The signals at the several antennas were varied electronically, thus controlling the direction of the transmitted beam. This type of antenna system is called a phased array or an electronically steerable array.

RADAR PERFORMANCE CONSIDERATIONS

Range. The range of a radar—that is, the maximum distance at which a target can be detected—depends on the sensitivity of the radar receiver, the peak power transmitted, the size of the antenna, and the size of the target. Power must be multiplied by 16 in order to double the range. Antenna area, on the other hand, need only be multiplied by four to double the range. Increasing the frequency by a factor of four will also double the range, other things being equal.

Radar cross section. The radar cross section of a target is equivalent to an area that would intercept the transmitted signal and reflect uniformly in all directions an amount which produces the returned echo. The actual area of the target is usually greater than its radar cross section, because some power is absorbed. In fact, in designing countermeasures against radar, one approach involves the possible use of absorbent materials to reduce the radar cross section. In special cases, however, a target may be designed to have a radar cross section larger than its physical area.

The radar cross section of a target is not constant with frequency. In general, there are three ranges of interest. In the first, the target dimensions are small compared to the radar wavelength. This is called the Rayleigh region, after Lord Rayleigh, the British mathematician and physicist who first studied scattering of electromagnetic radiation from small objects (Rayleigh scattering). In the next region, the target dimensions are approximately equal to the wavelength. This is called the resonance region. The third, or optical, region is that in which the target dimensions are much greater than the radar wavelength.

Principal ranges of the cross section

These scattering principles are important when it is necessary to choose a frequency to pick out or see through a specific target. For example, if it is desired to detect planes flying during rainstorms, a frequency should be chosen such that the dimensions of the raindrops are very small compared to the radar wavelength. On the other hand, for weather radars it is important to get strong reflections from rain. In that event the frequency should be chosen so that the dimensions of the drops lie in the resonant region.

Resolution. Range is measured by the time interval between the transmitted pulse and the received echo.

If two targets are in the same direction but at different distances, two separate echoes will be received. These will arrive at different times; each arrival will indicate the range of a specific target. In order to separate the two returns, however, it is necessary for the second echo to arrive after the first one has ended. A wide transmitted pulse will produce a wide return pulse. Hence, the narrower the transmitted pulse, the closer together the two targets can be and still be distinguished from each other (resolved). The smallest distance that can be discriminated is called the range resolution of the radar. Targets closer together than this minimum distance appear as one large target.

The direction of a target is usually specified in terms of two angles — azimuth and elevation. Azimuth is the angle in the horizontal plane between a fixed reference direction and the direction of the target. The reference direction may be a compass direction such as north, or it may be the direction in which the search radar is moving (for example, on a ship or airplane). The angle of elevation is the angle above the horizontal.

Angles of direction

As a radar antenna scans, its beam moves across a volume of space. When a target is intercepted, the reflection returns in microseconds, during which time the antenna movement has been infinitesimal. Consequently, when a target signal is detected, it may be assumed to lie in the direction in which the antenna is pointing. Measuring the angle of arrival of a target signal consists essentially of noting the direction in which the antenna is pointing when the target is acquired.

If two targets are at about the same range but at slightly different angles, they can be separated if they are more than a beam width (the angular width of the transmitted radar beam) apart. If they are closer than this, echoes from both will return simultaneously from a single pulse and will appear in the receiver as one signal. The angular resolution of the radar, then, is the beam width.

Doppler. The Doppler shift or Doppler frequency, is the change of frequency in the returned signal from a target having a component of motion toward or away from the radar.

If the returned signal is at a different frequency from the transmitted signal, the target must be moving radially. The amount of this shift (called the Doppler frequency) can be measured, and the velocity of the target can thus be determined. The radial velocity of the target is the rate at which the range of the target is changing, and it is usually referred to as the range rate of the target.

The measurement of Doppler frequency as well as range and angle makes it possible to pick out a moving target from larger, stationary masses. For example, the echo of a plane flying close to a mountain may be completely obscured by the echo from the mountain, but the plane's echo will be changed in frequency. A radar using Doppler can detect the plane easily.

Losses. The achievable range of a practical radar system is always much smaller than the theoretical value because of losses (attenuation) in the system. These include losses that result from passing a signal through the transmission lines and components in the system, which may be equivalent to reducing transmitter power by as much as 50 percent.

Operator loss is a measure of the efficiency of the human operator, who may miss a weak signal when tired or make errors when rushed. Propagation loss occurs between the radar and the target and is caused by weather, reflections from the ground, refraction (atmospheric bending), and other kinds of interference.

Propagation effects. Radar signals have many properties similar to those of light waves. In general, signals travel in straight lines but may be reflected or bent by refraction. Signals are also affected by moisture in the atmosphere and other meteorological phenomena.

If either the radar or the target is near the earth, the earth could be within the antenna beam width when the radar antenna is pointing directly at the target. Part of the transmitted signal hitting the earth may be reflected toward the target. The target is thus illuminated by two signals, the direct beam and the reflected beam. If the two beams arrive thus at the target out of phase (out of step) they cancel each other, and no radiation reaches the target. The target then remains undetected even if within the normal range of the radar. On the other hand, if the two beams reach the target in phase (in step), the target will be illuminated by a stronger signal and will then reflect a stronger echo toward the radar. This permits detection of targets normally beyond the maximum range of the radar.

Atmospheric effects

Although radar waves travel in straight lines in free space, close to the earth they may be bent or refracted by variations in the density of the atmosphere. Since a target will then appear to be displaced from its true position, it is necessary to correct observations to take into account the effects of refraction. Because of this bending of radar signals, it is also possible to see objects below the horizon.

It is sometimes possible for a layer of warm air to occur above a layer of cool air in the atmosphere, an abnormal condition. This temperature inversion changes the index of refraction (amount of bending) enough that signals may follow the curvature of the earth for great distances, far beyond the radar horizon. This phenomenon is called superrefraction or anomalous propagation.

All forms of precipitation adversely affect radar signals. The increased moisture causes loss by absorption. In addition, unwanted echoes from the precipitation, called weather clutter, may mask signals from a wanted target. By careful selection of frequency, it is possible to minimize the effects of weather clutter, where unwanted, or to maximize the reflections, where desired, to detect the presence of storms.

Displays. In practically all radars used through World War II and in many present-day radars, the presence of a target was indicated by some sort of visual display on the face of a cathode-ray tube, similar to the picture tube in a television set. The cathode-ray tube and its associated circuitry are simply called a scope an abbreviated form of oscilloscope.

There are many different types of visual displays possible on the face of the cathode-ray tube. The simplest, called an A-scope presentation, is shown in Figure 3A.

From Gershon J. Wheeler, *Radar Fundamentals* (1967); Prentice-Hall

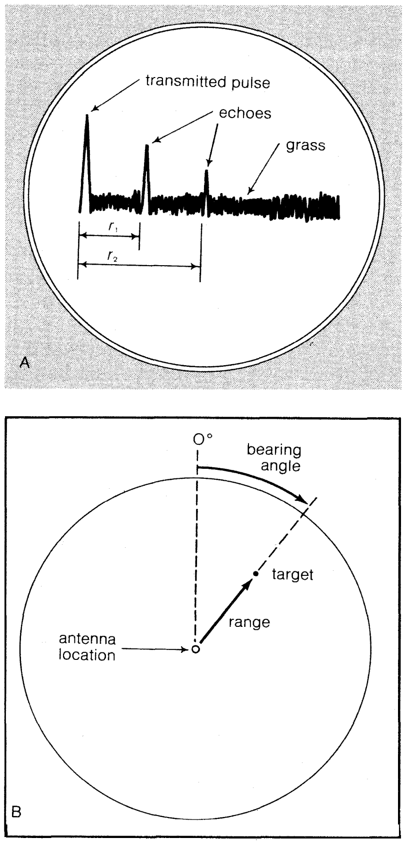


Figure 3: (A) A-scope presentation. (B) **Plan position indicator (PPI) presentation (see text).**

The horizontal distance across the face is calibrated in range, and the vertical displacement is the amplitude of the received echo. Each sweep begins when a pulse is transmitted and is indicated on the cathode-ray tube by a large inverted V or spot of light, called a pip, caused by a small part of the transmitted signal leading into the receiver. Echoes appear as pips along the horizontal line. The range, or distance, of a target producing the echo can be determined directly from the horizontal calibration. In Figure 3A two echoes are shown, one at a range r_1 and a weaker one at a range r_2. The horizontal sweep is not a sharp line but a constantly shifting pattern of short vertical lines caused by random noise in the receiver. This is called grass, possibly because in most of the early cathode-ray tubes the indications on the tube face were green.

There is no indication of direction of target on an A-scope. The operator knows that a target is in the direction the antenna is pointing. He can get this information by looking at the antenna or at some meter which indicates the antenna direction. Other types of displays were tried to present both range and direction simultaneously. On a B-scope presentation, azimuth is plotted horizontally and range vertically. The C-scope shows elevation vertically and azimuth horizontally. The B- and C-scopes were rarely used and were soon superseded by the plan position indicator display.

The plan position indicator, shown in Figure 3B, depicts the range and angular position of all targets seen by the radar as its antenna scans through a complete circle. A radial line sweeping from the centre to the perimeter is the equivalent of the base line in the A-scope. A target appears as a dot, and its range is indicated by the distance of the dot from the centre of the cathode-ray tube. As the antenna rotates, the radial line rotates on the face of the scope in synchronism, so that any targets that appear are always indicated in the proper relative directions. A point on the circumference is chosen as a reference. The face of the scope is specially treated so that images persist for about one revolution of the antenna. The result is a virtual map of the area being scanned. Moving targets can be detected by noting displacements of their echoes on successive sweeps.

Cathode-ray tube displays are not used in all applications. For example, in a small radar built inside a soldier's helmet, the echo is displayed as a sound in the earphones connected to the radar. If the soldier turns his head until the sound is maximum, he will then be facing approximately in the direction of the target. The pitch of the sound is related to the Doppler signal and thus gives some indication of the radial speed of the target.

The plan position indicator was used successfully during World War II, but its drawbacks became apparent toward the end of the war. In operation, the radar operator observed the plan position indicator and directed positioning of planes or guns and even firing times. With increasing target speeds, however, this method proved too slow for most practical operations, and computers had to be added.

Although computers are important components of most modern radars, there are still some applications in which the plan position indicator is used. At airports, for example, many planes may approach simultaneously at comparatively slow speeds. Nothing surpasses the plan position indicator for presenting a picture of all the planes in the area and their relative positions.

Countermeasures. During World War II both sides recognized the importance of developing countermeasures to render enemy radars ineffectual. As new techniques for jamming or deluding enemy radars were discovered, counter-countermeasures were developed, and radar warfare became a battle of wits.

Jamming
and
deluding

Jamming was a moderately successful, early technique of radar countermeasure that consisted of transmitting signals at the frequency of the enemy radar, so that his radar receiver was saturated or blocked and could not detect target signals. The flaw in jamming was that the enemy knew his radar was being jammed and could take evasive action by changing frequency. The object then was to discover the new frequency and jam that. In the early days of the war, radar frequencies could not be changed easily, and thus jamming was a successful countermeasure until better radars were developed.

One of the simplest, and yet most effective, methods of deluding radars consisted of dropping strips of aluminum foil from planes during raids to produce false echoes. Maximum return from the foil strips occurs when they are cut to a length of a half-wavelength. The Allies dropped long strips cut to a half-wavelength at the low frequency of the German radars while the Germans dropped short strips to yield false echoes for the microwave radars used by the Allies. On a full-scale bombing raid, the Royal Air Force dispensed enough aluminum foil to equal the weight of three bombers. The aluminum foil dispensed was called window by the British, chaff by the Americans, and *Dueppel* by the Germans.

More sophisticated countermeasures were developed after World War II, some of which are still secret. Absorbent materials were developed to coat friendly vehicles and thus absorb enemy radar signals and minimize reflections. Coatings of this material are used on military missiles to reduce their radar reflectivity and make the missiles more difficult to detect. In another technique the enemy radar beam is detected by the friendly vehicle, which, after a short delay, transmits a signal at the same frequency. The enemy radar then receives a signal much later than it should have from the particular target, and thus the range seems greater than it really is.

In all countermeasure systems it is important to know the frequency of the enemy radar. If a radar is to be jammed, signals must be set at the proper frequency. Aluminum strips must be cut to the proper length. Delayed signals must be transmitted at the received frequency. A large part of the radar countermeasure effort is spent in determining these frequencies.

RADAR EQUIPMENT

Continuous-wave radar. If a radar transmits continuously, the continuous echo cannot be associated with a specific part of the transmitted wave. Thus it is impossible to derive range information from a simple continuous-wave radar. This equipment can, however, derive speed information from the Doppler shift by measuring the shift in frequency of the reflected signal.

A simple continuous-wave radar is shown in Figure 4.

From Gershon J. Wheeler, *Radar Fundamentals* (1967); Prentce-Hall

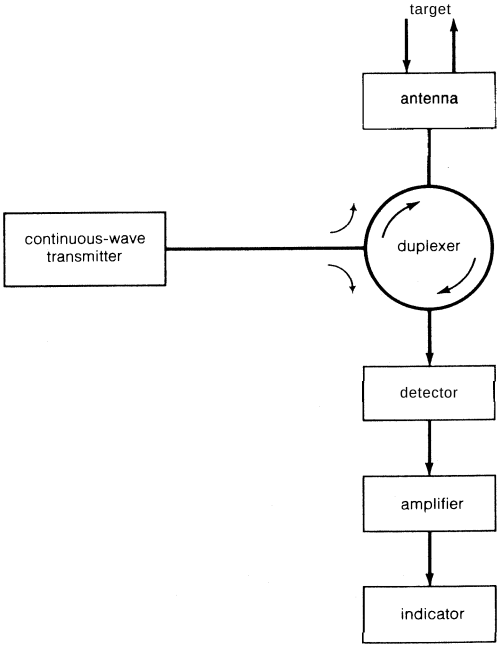


Figure 4: Continuous-wave radar (see text).

A signal transmitted at a given frequency is coupled to the antenna through the duplexer (a device which permits use of the same antenna for transmitting and receiving) and radiated into space. When the transmitted radiation is interrupted by a target moving radially, the reflected signal will be changed in frequency by the Doppler shift.

Though simple continuous-wave radar cannot furnish range information, it can be made to do so if each portion of the transmitted radiation is tagged so that it can be recognized on reception. One method of tagging the signal is to change the frequency continuously. When an echo is received, its frequency will be different from that of the signal leaving the transmitter at that time; if the rate of change of frequency is known, the difference in frequency will be an indication of the range; this is frequency-modulated radar.

Pulsed-Doppler radar. When early pulse radars were used with a plan position indicator for display, it was not necessary to use the Doppler shift to know that a target was moving. With each revolution of the antenna, a moving target would appear slightly displaced from its last position, and an experienced operator could learn to pick out the moving targets on the plan position indicator and ignore the fixed targets. When the moving target was very close to a large land mass, however, the large echo from the land mass completely masked the echo from the relatively small moving target. Experienced pilots soon learned this and tried to fly as close as possible to hills to avoid detection by enemy radar.

Radially moving targets produce Doppler shifts; fixed targets do not. Since the phase difference (the amount the two waves are out of step) between a reference signal and the echo is constant for a fixed target but is continuously changing for a radially moving target, there are differences in echoes from fixed or moving targets; and a radar has been developed to exploit these differences so that moving targets can be detected even in the presence of large land masses. Called moving-target-indication radar> this equipment can detect a moving target in the presence of clutter hundreds of times as strong.

Sometimes a target is moving at such a speed that its echoes are indistinguishable from those from fixed targets. Such a speed is called a blind speed, and various approaches have been made toward solving the problem. A possible solution is to increase the pulse-repetition frequency so that the blind speed is greater than the velocities of the targets of interest. Another approach is a staggered pulse-repetition frequency, in which each alternate pulse is delayed by a small amount, alternately producing two different interpulse intervals. A speed that is blind for one of the intervals will usually not be blind for the second.

Transmitter components. The heart of the transmitter, the high-powered device that produces the electromagnetic radiation, is one of the most expensive parts of a radar system. There are two basic types. One is the self-excited oscillator, which converts direct-current power directly into radio frequency energy. The magnetron is a typical example. The other radiation device employs a stable low-power oscillator followed by one or more amplifiers. Typically, the final amplifier is a high-power klystron. The oscillator–amplifier chain is inherently more stable than the self-excited oscillator and thus is more dependable for Doppler radars, in which the amount of frequency change is important.

Use of the high-power klystron

Today klystrons produce powers in the megawatt (10^6 watts) range; many recently developed tubes have relegated the magnetron to a less important role. Even though the klystron amplifier is not as efficient as the magnetron and is relatively larger than most microwave tubes, its stability makes it a preferred source for moving-target-indicator and other Doppler radars.

The travelling-wave tube is a microwave amplifier with an extremely wide bandwidth. Bandwidths of an octave or more (frequency ratio of two to one) are achievable in low-power, travelling-wave tubes, and bandwidths equal to 20 percent of the operating frequency have been obtained with high-power tubes for radar applications. A wide bandwidth is desirable to enable the operator to shift frequency over a large band, to avoid jamming.

Like the klystron, the travelling-wave tube has an electron gun at one end and a collector at the other. An external magnet along the outside of the tube keeps the electron beam from spreading. A radio-frequency signal is fed into the tube near the electron gun and is removed near the collector end. A structure built into the tube slows down this radio-frequency signal so that its velocity is approximately the same as that of the electrons. The electrons give up some of their energy to the signal, greatly amplifying it. The structure for slowing down the wave is usually a long helical coil.

Despite the demands for bigger tubes and more power, there is a growing need for compact, low-power radars for both civilian and military applications. **A** miniature continuous-wave Doppler radar using integrated circuitry is shown in Figure 5. This is a complete transmitter, receiver, and duplexer in a package smaller than a man's hand. With the addition of a small battery, an antenna, and a display, it is a complete radar systcrn. The transmitter produces 50 milliwatts (0.050 watt) of power at X-band (10,000 megahertz).

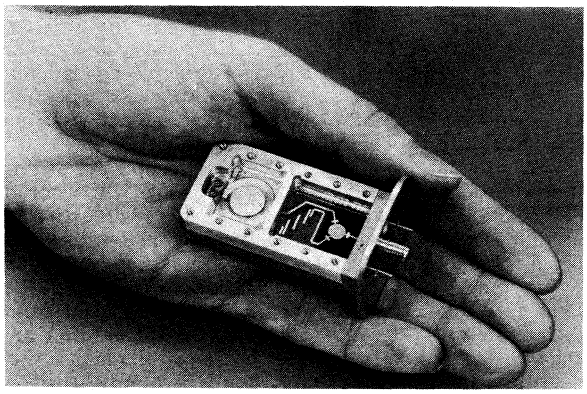


Figure 5: Miniature continuous-wave transmitter-receiver.
By courtesy of Hewlett Packard, Palo Alto, Calif.

This type of low-power radar is aimed at the growing civilian market. Since it can determine the velocity of a moving object, it can be used as an automatic collision-avoidance system in an automobile. It can also be used as an intrusion alarm, a traffic monitor, and in other situations in which detection of motion is needed.

Receivers. Two different types of receivers are used in radar systems—the superheterodyne and the video receiver. More complicated than the video receiver, the superheterodyne has far greater sensitivity (see RADIO). In radars for applications in which ranges are short and echoes are strong—such as traffic-monitoring radars and intrusion detectors—a video receiver is usually sufficiently sensitive and is chosen because of its simplicity. For most radar applications, however, the superheterodyne is preferred.

Electrons in motion generate electric currents that have the appearance of signals. In any practical receiver, the constant motion of electrons in tubes, crystals, resistors, and other components produces a measurable voltage at the output. These unwanted "signals," random in frequency and amplitude, are called noise in radio receivers. The main source of noise is the motion of electrons in conductors and resistors due to heat. It is called thermal noise.

Effects of noise

Noise is the limiting factor in a receiver. Without it, any incoming signal, no matter how weak, could be detected. In a practical receiver, however, noise is always generated at the input of the first stage, and any incoming signal must be strong enough to override it.

The sensitivity of a receiver can be expressed in terms of the weakest signal that can be detected. Since this would depend to some extent on the operator, it would not be a reliable measure of the receiver's capability. A more consistent measure is a value of signal related to the noise level at the input of the receiver. The standard measure of sensitivity is when the signal level just equals the noise level.

The noise in a receiver fluctuates continuously, and some of the noise peaks exceed the signal levels of target echoes. These peaks can be mistaken for signals or can mask signals when they occur at the same time. If the receiver is turned off part of the time, some of the noise peaks will be missed; the statistical probability of a noise peak being mistaken for a signal will be reduced. The use of so-called range gates reduces noise peaks in this way. A range gate is simply a switch that permits the receiver to operate only when targets at a desired range are sought. For example, if targets are to be located between 20 and 30 miles (32 and 48 kilometres) from the observing station, an echo from the nearest will arrive at the receiver about 240 microseconds after a pulse is transmitted and an echo from the farthest about 360 microseconds after the transmission (assuming about 12 microseconds per mile). Thus, the range gate might be set to turn the receiver on about 220 microseconds after transmission of each pulse and shut it off about 160 microseconds later. This would reduce the number of noise peaks appearing as signals and in addition would eliminate echoes from unwanted targets outside the desired range.

When two targets are in the same direction but at different distances from the radar, both echoes will occur at one position of the antenna, but the echoes will arrive at different times. If the second echo arrives before the echo pulse from the first is ended, however, it may be hidden by the first pulse. If, for example, two targets are in the same direction but about 500 feet apart, the echo from the second would begin about one microsecond after the echo from the first. If the pulse width of the radar is two microseconds, the first echo will only be half done when the second begins, and the two echoes would not be separable. Range resolution depends on the pulse width and could be improved if the pulse could be compressed on reception. Various techniques have been developed for accomplishing this and thus improving range resolution.

Duplexers

A duplexer is a device that permits use of the same antenna for both receiving and transmitting. It is indicated as a switch in Figure 1, and in most high-power pulsed radars the duplexer *is* a switch, connecting the antenna to the transmitter during transmission and to the receiver when the transmitter is off.

Antennas. A radar antenna is a coupling device between free space and the radar set. During transmission, energy from the transmitter is coupled to the antenna and caused to radiate. On reception, the antenna intercepts signals and couples them to the receiver. If a duplexer is used, the antenna is switched from one function to the other. Alternatively, the receiver and transmitter may have separate antennas that are permanently connected.

The main purpose of the antenna is to shape the transmitted beam so that the radiated energy is aimed in the desired direction in space. In general, the larger the aperture (opening that permits radiation) of the antenna (in terms of wavelengths), the narrower will be the beam. A narrow beam indicates that the energy is concentrated in one direction, which means the antenna has a relatively high gain in that direction. Antenna gain is the electrical output per unit of input, or the electrical efficiency of the antenna. Antennas are reciprocal, exhibiting the same gain and beam widths for reception as for transmission.

When an antenna is used for reception, the amount of power it intercepts is proportional to its area. Since not all parts of the antenna are coupled uniformly to the transmission line, however, the amount of power reaching the receiver seems to come from a somewhat smaller antenna. This "reduced" size is called the effective area of the antenna. The gain of the antenna is proportional to the effective area.

One form of antenna is a horn, formed by flaring out the walls of a wave guide to make a large aperture, which is in effect a large wave guide. It directs the radar beam in much the same way as a megaphone directs a voice.

It is possible to have fan-shaped antenna beams, wide in one direction and narrow in the other. This is accomplished by having an aperture that is wide in one dimension (to produce the narrow beam in that direction) and narrow in the other. It may be a specially shaped horn or a segment of a parabola.

Parabolic dish-shaped antennas are often used as reflectors in microwave systems; they are capable of focussing energy in one direction because of two special geometric properties. First, all rays from a fixed point, called the focal point, to the parabola are reflected as parallel rays. Second, in any plane perpendicular to the axis of the parabola, the reflected rays are all in phase (in step). When installed, the parabola may have holes cut in it or may be made of wire mesh instead of a continuous metal sheet, to reduce weight and decrease the effect of wind blowing on the dish. These openings in the surface have negligible effect as long as they are small compared to a wavelength.

A simple pencil-shaped radar beam cannot determine target position accurately, since, when a target is first detected, it may be at the top of the beam or off to one side. Even if the beam is swept through the target, the top of the beam may be too broad to fix the position accurately enough for gunlaying or mapping. Additional information is needed to find the target's exact position and, if the target moves, to follow it with the beam.

Conical scanning and sequential lobing

One method of supplying this information is called conical scanning. The beam is made to rotate spirally about an axis, so that a target in the "dead-ahead" position would be illuminated by slightly less power than the top of the beam. If the target is slightly off the axis, it will receive more power in one position of the beam as the beam rotates. In fact, the echo will be amplitude modulated (the height of the wave will be changed) at the conical-scanning rate (the rate of rotation). In operation, when a target is detected the antenna is moved so that the modulation is decreased, and when the modulation reaches zero the target lies on the axis of rotation.

Sequential lobing is another method of determining position accurately. Instead of rotating a beam, two or more fixed beams are transmitted sequentially. These beams occupy positions that would occur at equal intervals in a conical-scanning antenna and thus supply the same sort of modulation information.

The word scanning is used in two different senses in radar applications. After a target is acquired, the radar scans it to follow it as it moves. In this sense scanning refers to the comparison of target returns from two or more beam positions to determine the direction accurately. Conical scanning is an example. Scanning also refers to searching for targets. The radar beam is made to sweep over a volume that might contain targets of interest. This is usually done in a prescribed pattern.

In helical scan, the antenna beam rotates through 360° in azimuth (the arc of the horizon) while it is slowly raised in elevation. In spiral scan, the beam rotates in an ever-widening spiral. Helical scanning covers a hemisphere, while spiral scanning is limited to a conical volume. These are illustrated in Figure 6. Rectangular shapes can be covered by moving the antenna beam back and forth horizontally, raising it slightly at the end of each sweep.

From *Radiation Laboratory Series*, McGraw-Hill *Book* Co., Inc

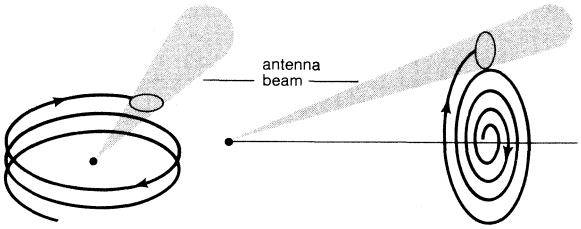


Figure 6: (Left) Helical scanning pattern and (right) spiral scan.

APPLICATIONS

Search and tracking. Radar either searches for targets or follows them. A radar that seeks a target is called a search or acquisition radar. During search, the operator may or may not know that a target exists in the covered space. Thus, a weather radar on an airplane searches for storms, which may or may not be in the plane's path. On the other hand, a radar may be used to check a satellite following a known orbit.

A radar that follows a moving target is called a tracking radar. To do this, the radar must sense the direction of motion and must move to keep the transmitted beam always pointed at the target. A scanning radar is sometimes used to sample the position of one or more targets as the beam sweeps through them. Such a radar is called track-while-scan radar.

Tracking implies following the target in azimuth and elevation. This is called tracking in angle. If the tracker also supplies continuous information about range or Doppler shift, it is said to track in range or in Doppler as well as in angle.

If a search radar employed a narrow beam to search a

Fan-shaped and pencil beams

large volume of space, it would take too long to scan the space systematically. For this reason, a search radar frequently uses two fan-shaped beams—one horizontal and one vertical. The horizontal beam moves up and down to determine the elevation of a target, while the vertical beam moves horizontally to determine the azimuth. The search radar can acquire many targets at different ranges simultaneously. For small volumes of space, the search radar can use a pencil beam. Then when a target is acquired, the same radar can be used for tracking.

A tracking radar uses a pencil beam that is pointed at the known position of a target. If the target moves off the centre of the beam, an error signal is produced that indicates the direction of motion and operates a mechanism to move the tracker back on target.

Altimeter. A radar altimeter is a simple form of search radar with the earth as a target. The radar produces a continuous-wave signal that is frequency-modulated. This is directed toward the earth, and the echo is detected in the receiver and compared in frequency with the signal being transmitted at the time of reception. Since the rate of change of frequency (that is, the frequency-modulation rate) is known, the difference in frequency indicates the elapsed time and hence the distance to the earth. The radar altimeter needs less than one watt output and a small antenna because the earth is a large target.

Doppler navigator. A Doppler navigator is a simple continuous-wave Doppler system used to measure a plane's ground speed. The radar carried on the plane has an antenna that directs a beam forward and down at a 45° angle to the direction of flight. In this case the earth is also the target. The true velocity can be determined from the relative radial motion.

Gunlaying. The error signal produced by a tracking radar can be used to point a gun at a target as well as to keep the radar beam pointing in the right direction. The radar then controls the direction of firing and is called fire-control radar. The gun is said to be slaved to the radar. During World War II, anti-aircraft guns were slaved to tracking radar to point in the same direction as the antenna beam. With the faster planes in use today, however, it is necessary to have the guns point ahead of the beam. During the time the shells would reach the plane's altitude, the plane would have moved too far ahead to be hit. In a modern fire-control radar, information from the tracker on range, angle, and rate of change of angle are fed to a computer, which then combines this with the velocity of the shells to calculate an intercept point. The guns are slaved to the computer.

Early warning. This is a long-range search radar that looks for enemy action at a distance in order to provide warning in time for effective counteraction. Typically, an early-warning radar may be used to detect an intercontinental ballistic missile at a distance greater than 2,000 miles. These radars are characterized by very high power, large antennas, and frequencies in the uhf or vhf regions where atmospheric attenuation is low.

Search radars can also be used to keep track of friendly planes in both military and civilian applications. An air-traffic-control radar at an airport monitors movements of planes arriving and departing in order to prevent collisions. Targets are usually displayed on a plan position indicator. When a plane is acquired, the radar operates as a track-while-scan radar for that plane while searching for others.

In a large airport, radars also monitor the movements of ground traffic. These radars must have a very narrow beam width in order to get the required resolution in azimuth and short pulses to separate objects close together on the ground. Plan position indicator displays are used with these radars. Similar radars are used at ports to monitor movement of ships.

Mapping. Radars for mapping transmit very short pulses in order to get sufficient range resolution. Ordinarily, distances to the area illuminated are comparatively short so that high repetition rates can be used without fear of ambiguity.

A weather-avoidance radar is used in a plane to detect storms so that the plane can avoid them. The frequency is chosen so that raindrops in the storm will reflect as much energy as possible without too much atmospheric attenuation of the signal.

Beacons. A beacon is not really a radar, in that it furnishes no information as to range. Beacons, however, are used in conjunction with radars and are usually described as a type of radar. A beacon transmits a signal only when it receives a specific coded signal from a radar. The radar in effect interrogates and the beacon responds. Beacons were originally designed for identification, friend or foe, radars but have many other applications. Weather satellites carry beacons. As these satellites revolve about the earth, they gather information about weather conditions around the world. When a satellite is overhead, its beacon is interrogated and responds with the data it has accumulated. Beacons may also be used in satellites to simplify tracking.

Each time a satellite is successfully launched by any country, it is usually announced publicly. Nevertheless, for political reasons it is necessary to have a capability to detect every satellite in space. The existence of this capability may be the reason why nations normally announce their launchings. Satellite detection in the U.S. requires a network of radars forming a radar fence, so that a satellite must pass through a radar beam to pass over any part of the country.

FUTURE DEVELOPMENTS

Missiles. In the future, if an enemy intercontinental ballistic missile lifts off its launch pad, it will be detected by a long-range early-warning radar; and data on the missile's flight will be fed to a computer, which will calculate the trajectory and estimated time of arrival and then control the firing of an antimissile missile, while the ballistic missile is still 3,000 miles (4,800 kilometres) away. Speeding toward the intercontinental ballistic missile, the antimissile missile will carry an atomic warhead to destroy the ballistic missile out of range of its target.

Each missile will carry its own radar and computer. As the antimissile missile approaches, it will be detected and tracked by the radar on the ballistic missile, which will change course to avoid its pursuer and, after the latter passes, continue toward its target. The radar aboard the antimissile missile will detect the evasive action, however, and the antimissile missile's computer will direct it to follow.

As the antimissile missile approaches, the Doppler shift will indicate the closing speed, and if it does not intercept its target but passes close to it, the Doppler will rapidly drop to zero at the closest point and will then reverse. The computer in the antimissile missile will fire the atomic warhead when the Doppler changes to zero—that is, at the closest point to the ballistic missile.

Computers. Computers play an increasing role in radar technology, eliminating human decision-making in many applications, and the outcome of future wars may depend on the sophistication of the computers that are connected to the radars.

Computers will not only use the information supplied by the radar but will also operate the radar or repair it in case of fault. The computer will check the operation of the radar continuously, and in case of fault or degradation, the cause will be located and a new part substituted automatically.

Radars today can provide accurate information about the direction, range, and velocity of a target. To some extent they can provide approximate indications of size and shape. It is probable that future radar systems will operate at higher frequencies, probably approaching optical frequencies. Improved resolution is possible at these higher frequencies.

Components. Although development of higher power sources continues, it is probable that greater improvement in radar performance will be achieved by advances in receiving techniques. Here again the computer will play a large role—storing information, separating signals from noise, and in general increasing the sensitivity of the receiver.

It is probable that newer tubes will replace the magnetron and klystron. New forms of amplifiers show promise of supplying stable high-power signals in less space than the klystron. For low-power radars, solid-state sources and circuits will eliminate all tubes.

Radar was developed in wartime as a military tool. Civilian applications may, however, come to predominate. Radar has been used for accurate mapping, intrusion detection, weather avoidance, and a number of other nonmilitary applications. Radars can be used to land planes during zero visibility. There are many applications of radar for automobiles under development. A typical collision-avoidance system uses a radar in the front of an automobile to measure the distance to the vehicle ahead and the closing speed. If the rear vehicle is overtaking the forward one too rapidly, a computer circuit connected to the radar takes over to slow the rear vehicle, either by braking or by reducing the flow of gasoline to the motor.

BIBLIOGRAPHY

Historical: H.E. GUERLAC, OSRD, ***Long History,*** vol. 5, division 14, Office of Technical Services, U.S. Dept. of Commerce (1946); A.P. ROWE, ***One Story of Radar*** (1948), a history of the Telecommunications Research Establishment; SIR ROBERT WATSON-WATT, ***The Pulse of Radar*** (1959), an autobiography; RICHARD MORENUS, ***Dew Line: The Miracle of America's First Line of Defence*** (1957); MASSACHUSETTS INSTITUTE OF TECHNOLOGY, ***Five Years*** (1946), a pictorial history of the Radiation Laboratory.

Technical: L.N. RIDENOUR (ed.), "M.I.T. Radiation Laboratory Series," 28 vol. (1942–53), vol. 1 provides a comprehensive description of all kinds of radars developed during World War II; each of the other volumes covers one facet of radar, such as antennas, receivers, and transmission lines; G.J. WHEELER, ***Radar Fundamentals*** (1967), a nontechnical, elementary presentation; J.F. REINTJES and G.T. COATE, ***Principles of Radar,*** 3rd ed. (1952), a definitive work on pulse radar; M.I. SKOLNIK, ***Introduction to Radar Systems,*** 2nd ed. (1980), a complete description of all forms of radar, and (ed.), ***Radar Handbook*** (1970), a comprehensive reference work; E. BROOKNER (ed.), ***Radar Technology*** (1977), a collection of papers covering a wide range of topics; J.V. DIFRANCO and W.L. RUBIN, ***Radar Detection*** (1968, reprinted 1980); D.K. BARTON, ***Radar System Analysis*** (1964, reprinted 1976), provides both a theoretical and practical analysis of various radar systems and in addition contains an excellent bibliography, and (comp. and ed.), "Radars" series, 7 vol. (1974–78), is a collection of reprints of important articles on all aspects of radar.

Technical treatment of special subjects: G.J. WHEELER, ***Introduction to Microwaves*** (1963); L. YOUNG (ed.), ***Advances in Microwaves,*** 2 vol. (1966–67); R.C. HANSEN (ed.), ***Microwave Scanning Antennas,*** 3 vol. (1964–66); J.V. EVANS and TOR HAGFORS (eds.), ***Radar Astronomy*** (1968); D.R. RHODES, ***Introduction to Monopulse*** (1959, reprinted 1980); R.J. SCHLESINGER *et al.*, ***Principles of Electronic Warfare*** (1961, reprinted 1979); ARTHUR ROBERTS (ed.), ***Radar Beacons*** (1947); L.J. BATTAN, ***Radar Observation of the Atmosphere,*** rev. ed. (1973).

(G.J.Wh.)

Radiation, Biological Effects of

All life is constantly being bombarded with various kinds of radiation—visible light, infrared, ultraviolet, radio waves, X-rays (from cosmic and terrestrial sources), and particulate cosmic rays—all of which are manifestations of energy transfer from one place to another. In a wide sense, any consequence of the transfer of radiation energy to a living organism is a biological effect of radiation. This definition includes both the normal effects on many life processes (*e.g.*, photosynthesis in plants and vision in animals) and the abnormal or injurious effects resulting from the exposure of life to unusual types of radiation or to increased amounts of the radiations commonly encountered in nature. This article, devoted chiefly to the discussion of the deleterious effects of radiation, is divided into the following sections:

I. Definitions and concepts
  Types and measurement of radiation
    Electromagnetic waves
    Particulate or corpuscular radiations
    Measurement of radiation
  Mode of action
    Mechanisms of energy transfer
    Ionization and penetrating power
    Mechanisms of biological action
II. General biological effects
  Genetic effects
    Gene mutations
    Chromosome mutations
  Bodily (somatic) effects
    Long-term effects
    Short-term effects
III. Radiation and human health
  Man's radiation burden
    Natural sources
    Artificial sources
  Injury from ionizing radiations
    Historical background
    Radiation effects on cells
    Radiation effects on tissues and organ systems
    Acute lethal effects
    Prenatal exposure and exposure of children to radiation
    Protection against external radiation
    Long-term effects, somatic and genetic
    Radioisotopes and fallout
  Accumulation of isotopes in critical organs
    Radiation and cancer
    Chromosome aberrations and radiation
  Effects of Hertzian waves, infrared rays
    Hertzian waves
    Infrared rays
  Effects of visible and ultraviolet light
    Intrinsic action
    Photodynamic action
    Effects on development and biological rhythms
    Effects on the eyes
  Effects of lasers

I. Definitions and concepts

For a better understanding of the later sections of this article, a review of the two main types of radiation—electromagnetic waves and moving atomic particles—and of their measurement, sources, and mechanisms is in order.

TYPES AND MEASUREMENT OF RADIATION

Electromagnetic waves. Energy can be transferred through matter or through space by means of oscillatory variations in electric and magnetic fields originating at various sources. Electromagnetic-wave radiation is classified into several types, depending on frequency and wavelength. The various types (in order of decreasing wavelength and increasing frequency) and their sources are summarized in Table 1.

Table **1: Electromagnetic Waves**

wave type	sources
Hertzian (radio waves and microwaves)	radio transmitters
Infrared	hot bodies
Visible	hot bodies; excited molecules
Ultraviolet	hot bodies; excited gases
X-rays and gamma rays*	atoms struck by high-energy particles; radioactive materials; cosmic sources

*Similar to X-rays but emitted from the nuclei of some radioactive atoms.

Particulate or corpuscular radiations. Simply speaking, an atom consists of a nucleus, composed chiefly of neutrons and protons, surrounded by a cloud of electrons; in addition, there are other special particles. When separated, either by natural radioactive disintegration or by artificial means—as in a cyclotron, for example—these particles, charged or uncharged, are capable of transferring their energy, wholly or in part, to any substance through which they pass. Streams of such particles constitute particulate radiations. The more important types of particles, the magnitudes and signs of their charges (relative to that of an electron, defined as the unit negative charge), their masses (relative to the mass of a hydrogen nucleus, defined as the unit mass), and their main sources are summarized in Table 2. The rate of transfer of energy by a particulate radiation depends on the mass of the particle, its velocity, and, if charged, the magnitude of its charge.

Measurement of radiation. Radiation is detected and measured by various methods that rely on the property of the radiation to cause ionization in gases or emission

Radiation detection

Table 2: Particulate Radiations

particle type	charge	mass	sources
Elecron* and positron	−1	1/1,843	accelerating machines, radioactive materials
Proton†	1	1	accelerating machines and cosmic rays
Alpha (a) particle‡	2	4	radioactive materials and cosmic rays
Neutron	0	1	nuclear reactors, accelerating machines, and cosmic rays
Heavy nuclei	§	§	

*Electrons emitted by radioactive nuclei are called beta (β) particles. †Protons are the nuclei of hydrogen atoms. ‡Alpha particles are the nuclei of helium atoms. §Various.

of light from sensitive materials or blackening of a photographic emulsion. Instruments used are the ionization chamber, Geiger counter, and scintillation counter.

Roentgen unit. The original biological unit of X- (or gamma-) ray exposure, the roentgen unit (R), is defined as that quantity of radiation that produces a given number of charged ions in a given quantity of air under standard conditions.

Rad. A more general and more accurately defined measure of local exposure to all types of radiation is the rad, the radiation dose absorbed in tissue itself. It is equal to 100 ergs per gram.

Rem. The same dose in rads from different radiations may produce different degrees of biological effect. The dose in Rem (roentgen equivalent man) is the product of the dose in rads and a factor called the quality factor, which depends on the relative biological effectiveness (RBE) of the radiation used. The Rem is taken to be that dose from any radiation that produces biological effects in man equivalent to one rad of X-rays.

For more technical definitions, see **RADIOLOGY**; see also **RADIOACTIVITY**.

MODE OF ACTION

Mechanisms of energy transfer. There are several mechanisms by which energy is transferred from radiations to biological materials. Wave radiations yield up their energy through scattering and quantum absorption. The energy of infrared and visible radiation is absorbed mainly by whole molecules and atoms. The energy of ultraviolet radiation is absorbed partly by the planetary electrons of atoms; an electron jumps to a higher energy orbit, thereby bringing the atom into a state of excitation, in which it is chemically reactive. The energy of X-rays is also absorbed mainly by the planetary electrons of atoms, but the amount absorbed is usually great enough to produce ionization—*i.e.*, to induce an electron (or electrons) to escape altogether from the atom, thereby ionizing the atom; an ionized atom is highly reactive, even more so than one in a state of excitation.

Particulate radiations yield up their energy through collision with the planetary electrons or atomic nuclei of the material through which they pass. Charged particles usually react with the planetary electrons and therefore leave the atom in an excited or, more commonly, ionized state. Neutrons, being uncharged, are able to pass through the orbits of the planetary electrons and collide with the nucleus of an atom. The nucleus may then recoil, or it may capture the neutron. In biological material, neutrons yield up most of their energy through collision with hydrogen nuclei, which recoil as hydrogen ions (*i.e.*, protons). The neutron is eventually captured by some nucleus, usually followed by the ejection of one or more charged particles, often accompanied by gamma rays.

High-energy radiation

Ionization and penetrating power. Although all higher energy radiations (*i.e.*, X-rays and particulate radiations) produce ionization in biological material, the spatial distribution of the ionization depends on the type and penetrating Power of the radiation, the nature of the material irradiated, and the spatial distribution of the radiation source or sources. High-energy X-rays produce ionization that is rather sparsely but uniformly distributed along the track of the radiation; alpha particles and heavy ions cause intense ionization along their tracks; in between these come neutrons, protons, and beta particles. In general, the lower the speed of a particle, the denser the ionization along its track, so that all particles tend to be more densely ionizing toward the ends of their tracks. High-energy X-rays penetrate deeply (X-rays commonly used in medicine are less penetrating), but most beta particles penetrate only a few millimetres in biological tissues and alpha particles only a small fraction of a millimetre. Animal tissues, such as bone and teeth, that contain elements of fairly high atomic weight, such as calcium, absorb much more energy from a given radiation than do soft tissues, which are composed mainly of elements of low atomic weight. Ultraviolet radiation is of very low penetrating power.

Mechanisms of biological action. The biological effects of the higher-energy electromagnetic radiations and of all the particulate radiations are mediated through the ionization (and, to a lesser extent, the excitation) that they produce in biological tissue. Thus, all ionizing radiations have broadly similar biological effects; such differences as are found are the result of differences in the spatial distribution of the ionization. An explanation of its mode of action must be sought at the cellular level; effects on the whole organism are likely to be secondary. Two main types of intracellular action have been distinguished; first, direct action through ionization of biological structures along the ionized track; and, second, indirect action through the formation of reactive chemical fragments (free radicals) that diffuse away from the ionized track and undergo further reaction elsewhere.

Types of intra-cellular action

Direct action. Direct biological actions were studied in great detail in the period between **1927** and 1947. **A** detailed quantitative theory was elaborated, the target theory, or *Treffertheorie*, whereby a tissue undergoing irradiation was likened to a field traversed by the fire of a machine gun. It was supposed that, to produce a given effect, there must be one or more hits by an ionized track on a sensitive target, so that the probability of obtaining the effect was dependent on the probability of obtaining the requisite number of hits on the appropriate target. This theory was very successful in giving a quantitative treatment of many of the biological effects of radiations, particularly in the field of genetics.

Indirect action. In the field of radiation chemistry, where free radicals produced by radiation play a vital role as intermediates in chemical reactions, the target theory had little application, and, from about **1940,** the interest of radiation biologists tended to shift toward the indirect actions of radiations. This shift was given impetus by the discovery in **1947** that the induction of chromosome breakage, which had until then been viewed as a target-theory effect *par* excellence, was enhanced if the tension of oxygen was increased in the material irradiated. Similar effects are known in radiation chemistry, in which oxygenated solutions can lead to more extensive radiation damage to solute molecules than do oxygen-free solutions. Highly active molecular particles called free radicals are known to be mainly responsible for the changes produced in these cases. Substances were then sought that have a high reactivity toward free radicals in order to reduce the biological effects of radiation exposure. Two such protective substances are cysteamine and cysteine. Unfortunately, these compounds can themselves cause adverse biological effects.

II. General biological effects

The biological effects of radiations may conveniently be subdivided into somatic effects—short-term somatic effects (*i.e.*, short-term effects on the body of the individual) and long-term somatic effects—and genetic effects. Historically, they were discovered in that order. But, as the genetic effects are nearest to the cellular level and the short-term somatic effects farthest from it, it is more convenient to consider them in the reverse order.

GENETIC EFFECTS

The genetic effects of radiations arise through damage to those intracellular bodies in the germ cells that are the

Effects on sex cells

material basis of heredity (*q.v.*). These are the chromosomes, with their constituent genes (see GENE). Genetic effects therefore occur only if the radiation reaches the germ cells. Thus, in those plants and animals (man is one) in which the germ cells are fairly well covered by tissue, radiations of low penetrating power, such as ultraviolet light and alpha and beta particles, cannot produce genetic changes; however, radiations of high penetrating power, such as X-rays and gamma radiation, can induce genetic effects.

Radiation-induced genetic effects on animals and plants are probably not qualitatively different, if consideration is given to the physiological variations between these two kinds of life. Mutations induced by irradiation of seeds are of interest to plant breeders because of the possibility of producing new varieties. Differences have been found in the mutagenic effect of radiations on plants, which appear to depend upon the density of ionization produced. Fast neutrons and heavy particles were found to be up to 100 times more mutagenic than X-rays. Radioactive elements taken up by plants can also be strongly mutagenic. In choosing a suitable dose for the production of mutations, a compromise has to be reached between the mutagenic and damaging effect of the radiation on the plants. As the number of mutations increases, so also does the extent of damage to the plant. A critical dose is defined as that allowing 40 percent survival. In the irradiation of dry seeds by X-rays, doses of 10,000 to 20,000 roentgens are usually given. Mutants can be produced by radiation that affects such properties as early ripening, resistance to disease, and the like, and economically important varieties of various species have been produced.

In plants and animals, genetic changes of two sorts are induced, namely, gene mutations and chromosome mutations. They differ in important respects and will be considered separately.

Gene mutations. Radiation-induced gene mutations are all of the same types that occur spontaneously in nature. Most gene mutations, whether spontaneous or radiation-induced, are harmful. They reduce the fitness of the organism, in the sense that an individual carrying a mutated gene is usually less capable of surviving and leaving descendants than is an individual carrying the gene in its unmutated form. Favourable mutations produced by radiation and other mutagenic agents are believed to be responsible for continued evolution.

Although the mutations induced by radiations are all of types that occur spontaneously, the relative frequencies with which any two particular genes mutate under irradiation may not be the same as the frequencies with which they mutate spontaneously; however, if whole classes of genes are considered, the mutational idiosyncrasies of particular genes are lost, and the overall picture shows that exposure to radiation causes a proportionate increase in the mutation frequency of all classes of genes.

Increase of mutational frequency

Experiments with animals and plants have established that mutations imply alterations in the molecular structure of genes. The yield of mutation is usually proportional to the amount of energy absorbed in the germ cells; *i.e.*, to the radiation dose. This linear dependence of mutation on dose, which is to be expected on target-theory consideration if gene mutation is the result of a single hit by an ionized track, has been demonstrated experimentally for doses down to a few rads. It would be difficult to obtain experimental verification at doses much lower than this. The implication is that there is no threshold dose for the induction of gene mutation; *i.e.*, any dose, no matter how small, will induce some mutation.

Radiation intensity was originally thought to have no importance, the total amount of mutation induced being dependent only on the total accumulated dose. In 1958 it was shown, however, that this does not hold true for mutation induced in immature germ cells (spermatogonia and oocytes) of the mouse, the amount of mutation induced by a dose accumulated at a rate of about one roentgen per day being only about a third of that induced by the same dose accumulated at about 100 roentgens per minute. If this is true of mammalian spermatogonia and oocytes generally, it holds important implications for man, whose germ cells are present mainly in the form of spermatogonia in the male and oocytes in the female.

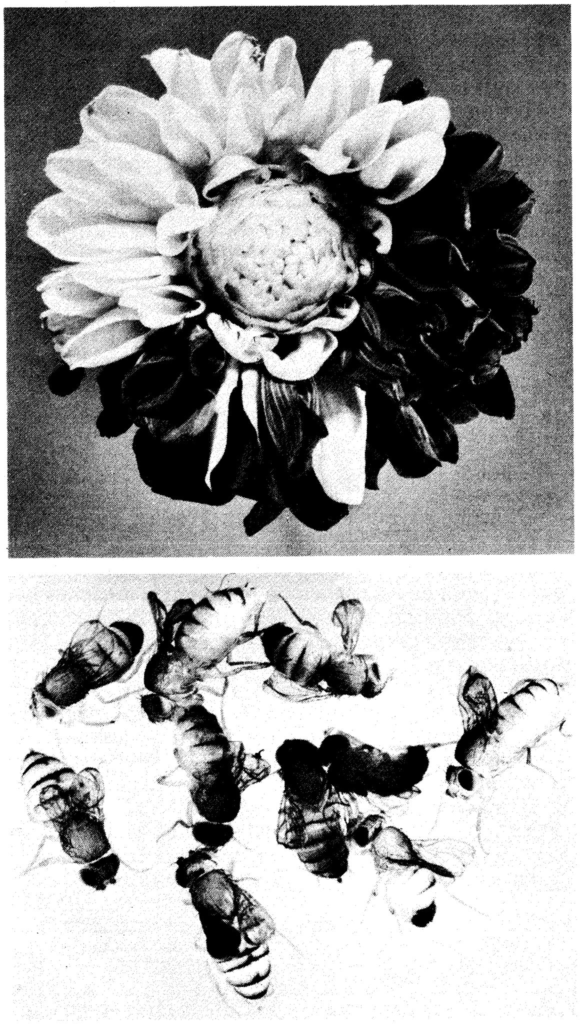
The amount of mutation induced by radiation is usually expressed in terms of the mutation-ratedoubling dose, which is the dose that induces as much mutation as already occurs spontaneously in each generation. This is merely a mathematically convenient way of relating the amount of induced mutation to that arising spontaneously; there is no special biological significance in a mutation rate twice the spontaneous rate. The more sensitive the genes to radiation, the lower the doubling dose. Doubling doses for high-intensity exposure of several different organisms have been found experimentally to lie between about 30 and 150 roentgens; for seven specific genes in spermatogonia of the mouse, the doubling dose is about 30 roentgens for high-intensity exposure and about 100 roentgens for low-intensity exposure. Very little is known about the doubling dose for human genes; most geneticists assume that it is about the same as the doubling dose for the mouse. Studies of the children of survivors of the atomic-bomb explosions at Hiroshima and Nagasaki have not so far yielded radiation-induced mutations in humans (see below).

The linear law

The existence of a linear law relating induced gene mutation to radiation dose holds an important implication for populations: it implies that very small doses of radiation given to very large numbers of individuals may introduce into the population as many mutant genes as would be introduced through large doses to small numbers of individuals. Furthermore, it is necessary to take into consideration the probability that an exposed individual will contribute to the next generation at some time after exposure. It follows that exposure of individuals below reproductive age (including unborn fetuses) is genetically of the greatest importance, and exposure of those above reproductive age is of no importance.

The effect on a population of a rise in its mutation rate depends on the role played by mutation in determining the characteristics of the population. Deleterious genes enter the population through mutation, but, because they reduce the fitness of their carriers, they tend to die out; thus a genetic equilibrium is set up at a point where the flow of deleterious genes into the population through mutation is counterbalanced by the loss through reduction in fitness. At equilibrium a constant fraction of the population rate would increase the gene-handicapped fraction proportionately. The full increase, however, would not be manifest immediately; this would occur only when genetic equilibrium had again been established, which would probably require several generations.

Chromosome mutations. Ionizing radiations not only cause genes to mutate but also may break the chromosomes in which the genes are carried. Chromosome breaks often heal spontaneously, in which case no damage becomes apparent. If a break fails to heal, a germ cell may be formed that lacks an essential part of the gene complement. This is called deletion. Such a germ cell may be capable of taking part in the fertilization process, but the ensuing zygote is usually incapable of full development and dies in an embryonic state. When two chromosomes in the same nucleus are broken, it sometimes happens that the broken ends join together, but in such a way that the order of the genes in the chromosomes is changed; for example, part of chromosome *A* may join onto part of chromosome *B* and vice versa in a process termed translocation. A germ cell carrying such a chromosome structural change may be capable of giving rise to a zygote that may develop into an adult individual, but the germ cells produced by the latter include many that lack the normal chromosome complement and so give rise to zygotes that are incapable of full development. An individual of this sort is called semisterile, and the number of his descendants is correspondingly lower than normal. Chromosome structural changes, therefore, usually die out of a population; for this reason they are of little importance in human populations. In some species, however, there are mechanisms that reduce the loss of fertility usually associated with chromosomal changes, and in them such changes may be present in the majority of



The mutational effects of radiation.
(Top) Somatic mutation in dahlia. White petals on a normally red flower caused by daily exposure to gamma radiation.
(Bottom) Genetic mutations in fruit flies include these flies with abnormal wings, offspring of males that received internal radiation from radioactive phosphorus.
By courtesy of Brookhaven National Laboratory

individuals. They include the evening primrose (*Oenothera)* and some species of the fruit fly (Drosophila). As would be expected from target-theory considerations, the induction of two-break chromosome changes is not dependent on radiation dose in a simple linear fashion. High doses and doses given at high intensities induce proportionately more changes of this sort.

BODILY (SOMATIC) EFFECTS

Somatic effects due to radiation include all the injuries, long-term and short-term, to al! the cells of the body of an organism—plant or animal.

Cell and tissue damage

Long-term effects. Long-term somatic effects include damage to cells that are continually proliferating, so that the injury is passed on to succeeding cell generations. Such cells are found in embryonic tissues and in those tissues of the adult in which cell division normally continues throughout life; in vertebrates these include the blood-forming tissues, the basal layer of the skin, the intestinal mucosa, and, in the male, the germ cells. Furthermore, the radiation may affect the rate of cell division: the cells may be killed or otherwise rendered incapable of further division; division may be slowed down; or, conversely, tissues in which cell division normally ceases in the adult may escape their normal biological control and cell division may continue. Mosaic sectors have been observed in many organisms exposed to radiations during embryonic development. They include humans who have been exposed during diagnostic examination of the mother; in some fetuses, there was a brown sector in the otherwise blue iris of an eye; in others, there was a black patch in an otherwise blond head of hair. Unilateral and nonheritable expression of a congenital defect of a type that is normally bilateral and heritable is also believed to arise sometimes through somatic mutation. Sometimes the mutant region includes some of the germ cells, and the genetically mutant nature of the mosaicism can be confirmed by a breeding test.

Effects on embryonic tissues

Slowing down or suppression of cell division, including radiation-induced cell death, may have important end results where embryonic tissues are exposed. This is because the rate of cell division is not uniform throughout an embryo; at any one stage of development, the cells in some parts will be dividing rapidly, while, in other parts, there will be relatively little cell division. Exposure to radiation will affect the rapidly dividing tissues more strongly than the others, with the result that the pattern of development will be distorted, and the individual, when adult, will have djsproportionate parts. This phenomenon, which is well-known in experimental plants and animals, is also known in man.

Radiation exposure of the blood-forming tissues leads to a reduction of cell division and therefore to a reduction in the blood-cell count, which may, if sufficiently great, amount to clinically recognizable anemia (for this reason regular blood-celi counts used to be carried out on all radiation workers). Radiation exposure of the germinal tissues likewise leads to **a** reduction of cell division, which may be sufficient, if the dose is large enough, to induce sterility; a dose of 200 roentgens to the gonads of the human male induces temporary sterility, lasting a few months; a gonad dose of about 500 roentgens to either sex induces permanent sterility. Opacity of the lens of the eye, which may develop into a full cataract, has been observed following exposure to ionizing radiation, especially neutrons.

Shortening of the life-span

A further long-term effect of radiation exposure is a general shortening of the life-span, not associated with death attributable to any specific cause. It is a phenomenon well established in laboratory animals, but evidence about its extent in humans is conflicting; the mechanism, whether through gene mutation in somatic cells, a reduction of cell division, or some other cause, has not been established.

Loss of biological control over cell division (*i.e.*, the induction of cancer in animals and distortion of parts in plants) is one of the more serious long-term somatic effects of excessive radiation exposure. It may occur in any tissue. Cancer of the skin was the first long-term effect to be observed; it may result from exposure to ultraviolet as well as to ionizing radiation.

The nature of the dependence of the long-term effects of radiation on radiation dose, type, and intensity is not yet established with certainty. Many authorities assume that the relationship is of a nonthreshold type, meaning that any additional exposure increases the probability of long-term effects occurring. It is a characteristic of the induction of cancer that there may be a long period of time between the causal exposure and the appearance of the effect: for some skin cancers in man, induction periods of 40 years have been recorded.

Short-term effects. A sufficiently large dose of radiation will kill any organism, but the dose required varies greatly from species to species; mammals are killed by less than 1,000 roentgens, but fruit flies may survive 100,000 roentgens, and many bacteria and viruses may survive even higher doses.

This lethal action of radiation can be used as an alternative to heat treatment for sterilizing materials such as surgical sutures. Its application in the sterilization of foodstuffs requires high doses and is a subject for research.

III. Radiation and human health

MAN'S RADIATION BURDEN

Natural sources. Throughout the ages man has been exposed to what is called natural background radiation. This includes radiation from cosmic as well as terrestrial sources. It has probably played the major role in the evolution of life.

Cosmic rays

Cosmic rays (*q.v.*) rain down incessantly on the Earth's atmosphere. The primary particles, protons, helium ions, heavier nuclei, and some electrons interact with atoms of the air and cause secondary radiations—chiefly electrons, gamma rays, and mesons (a class of atomic particles)—by the time they reach the Earth. There is considerable variation in the intensity of cosmic radiation, both geographically and temporally, since sunspots and solar flares influence the primary intensity. For about every 5,000 feet of altitude, the intensity approximately doubles (see Table *3)*. One of the hazards of space travel is a belt of "trapped" radiation that surrounds the Earth at a height of several thousand miles (see VAN ALLEN RADIATION BELTS).

Table 3: Cosmic-Radiation Exposure

location	mean dose in rads* per year
Sea level, temperate zone	0.020–0.040
5,000 ft	0.040–0.060
10,000 ft	0.080–0.120
40,000 ft	2.8
30–600 km	7–15
Interplanetary space	13–25
Van Allen radiation belt (protons)	<1,500
Single solar flare (protons and helium)	<1,000

*Rad is the unit of radiation dose; it corresponds to 100 ergs energy transferred to one gram tissue.

Variations of cosmic-ray dose have occurred during the geological history of the earth by the reversals of the geomagnetic field, which acts as a protective shield at present. Appearance of a supernova near the solar system could cause a sudden rise in cosmic-ray intensity several hundredfold.

Natural radioisotopes

Man also receives external and internal radiation from natural radioisotopes, particularly radium and its daughters, members of the thorium series, potassium-40, carbon-14, and hydrogen-3 (tritium). The last two are produced by cosmic rays, whereas the former are mainly the result of decay of radioisotopes believed to have been present when the Earth was formed. The greater part of external radiation comes from the radioactivity of minerals, whereas radioactive contamination of drinking water and food plays an integral part internally. Drinking waters vary by a factor of 10,000 in their radioactivity content. Among the foods, radium content of nuts and cereals is higher than of milk or meat. Internal doses in normal persons from natural radioactivity are listed in Table 4.

Table 4: Internal Dose Due to Natural Radioactivity

isotope	radioactivity (in curies*)	radiation	dose in rads (per year)	critical organ
Carbon-14	9×10^{-8} per g	beta rays	0.0016	gonads
Potassium-40	10.4×10^{-8} per g	beta rays	0.0165	gonads
Potassium-40	1.15×10^{-8} per g	gamma rays	0.0023	gonads
Radium and daughters	1×10^{-10} in body	alpha, beta, gamma rays	0.0380	bones

*Curie is the unit of radioactivity; one curie corresponds to 37,000,000,000 disintegrations per second.

There are certain localities with relatively high background due to the presence of radioactive minerals near the soil surface. Such areas occur in Brazil, Sweden, and France. The Kerala province of India is rich in monazite sands, with considerable natural radioactivity. External doses due to natural radioactivity from sources such as these are shown in Table 5.

Table 5: External Dose Due to Natural Radioactivity

source	dose in rads per year
Ordinary regions	0.025–0.160
Active regions	
Granite in France	0.180–0.350
Houses in Sweden (alum shale)	0.158–0.220
Monazite alluvial deposits in Brazil	mean0.500;max 1.0
Monazite sands, Kerala, India	0.37–2.8

Man-made radioactivity

Artificial sources. Since the discovery of radioactivity, man has substantially added to his natural-radiation load. Medical X-rays and radioisotopes often are necessary for the diagnosis or treatment of disease. In some countries almost the entire population is exposed to periodic diagnostic X-rays and a significant fraction to therapeutic doses. Therapeutic doses vary widely, depending on, among other things, the disease being treated. Routine diagnostic doses, though they may vary somewhat depending on the apparatus and the operator, usually fall within certain limits (see Table 6).

Table 6: Typical Doses Received in Routine X-Ray Diagnosis

examination	dose per exposure
X-ray photograph	
Chest X-ray	0.04–1 rad
Gastronomical X-ray	1 rad
Extremities	0.25–1 rad
Fluoroscopy	10–20 rads per minute
X-ray movie	25 rads per examination

High-voltage power supplies for radar or television, X-ray machines in shoe stores, dental X-rays, luminous-dial watches, phonograph static eliminators, and television sets give significant doses of radiation, sometimes more than-is warranted. The yearly dose from diagnostic X-rays has become comparable to the dose received from cosmic rays. The United States leads in annual per capita dose as a result of diagnostic procedures, followed by Japan and the United Kingdom. Efforts are being made to reduce diagnostic exposures by design of more efficient X-ray equipment.

Further contributions of man that have become of concern are fallout resulting from nuclear-weapons testing (Table 7) and radioactivity produced in reactors.

Table 7: Worldwide Dose Commitment from Nuclear Tests Prior to 1970*

source	isotope	half-life	dose to bone surfaces (mrad)
External radiation	short-lived (*e.g.*, iodine-131)	8 days	36
	longer-lived (cesium-137)	30 years	36
Internal radiation	strontium-89 and -90	50 days	131
	cesium-137	28 years	21
	carbon-14t	5,730 years	16
Total			240

*North Temperate zones; doses calculated for bone surface.
†Calculated to year 2000 only.

Most of the radioactivity produced in power reactors is safely contained. **A** small percentage escapes as stack gas or liquid effluent and eventually may contaminate the atmosphere and water supplies. Similar problems exist in nuclear-fuel-reprocessing plants. Planned peaceful uses of atomic explosions in digging harbours or in mining minerals far underground would also produce radioisotopes. While atomic plants promise to be abundant and clean sources of energy, they may eventually contribute considerably to worldwide radiation background (Table 8).

The use of coal instead of atomic energy in future power plants will not solve the problem of radioactive contamination since many sources of coal contain natural radioactivity (*e.g.*, radium), which appears in stack gases, in addition to chemical pollutants.

Portable atomic batteries contain large amounts of dangerous radioisotopes (*e.g.*, polonium or plutonium), safely sealed. Atomic-power rockets will be available soon for interplanetary travel. These will produce some radioactive contamination (*e.g.*, hydrogen-3) in interplanetary space.

It appears from the tables that, if the current trends are maintained, by the end of the 20th century the entire human population may be exposed to twice the radiation level from natural sources. It is important to study and

Table 8: Estimate of Long-Lived Radioactive Gases Released in Air and Water by Power Reactors and Fuel-Reprocessing Plants, Global Scale*

type	1970	1980	2000
Curies, released per year krypton-85 equivalent†	17,000,000	200,000,000	1,900,000,000
Curies, released per year hydrogen-3'	340,000	4,000,000	37,000,000
Curies, hydrogen-3 ‡ accumulated in air and water	1,000,000	62,000,000	590,000,000
Curies, hydrogen-3 in air due to cosmic rays	40,000,000–80,000,000	40,000,000–80,000,000	
Curies, krypton-85 accumulated in air	50,000,000	3,000,000,000	25,000,000,000

*Courtesy of Thomas Pigford, University of California, Berkeley. †In units of krypton-85 equivalent, maximum permissible concentration 0.0003 microcurie per litre air. ‡Maximum permissible concentration 3 microcuries per litre air. Amounts shown are much smaller than the permissible concentration if one assumes uniform distribution in atmosphere. Care must be taken that local concentrations should not exceed allowable limits.

understand the possible consequences of these changes in radiation level. It is interesting to note that fusion processes, if they can be shown to be practical for power sources, will produce considerably less radioactivity than fission reactors at a similar power level.

INJURY FROM IONIZING RADIATIONS

Although all forms of radiation, if intense enough, may produce some adverse effects on man, the "hard," or penetrating, ionizing rays, including X-rays and moving atomic particles, present the greatest hazard. Paradoxically, these same ionizing rays, when judiciously used in medicine, constitute a formidable weapon against cancer and an invaluable aid in diagnosis (see RADIOLOGY).

Maximum permissible exposure limits

The concept of maximum permissible exposure to radiation is a much-debated point, subject to change as more information accumulates. Although the body may not be noticeably affected, genetic damage may occur and pass unnoticed for several generations, as was mentioned earlier. Maximum permissible exposures are proposed for persons who, by the nature of their work, are exposed to amounts of radiation beyond that to which the general population is exposed. (Recommendations vary somewhat in different countries; those given in Table 9 are for the U.S.)

Table 9: Maximum Permissible Exposures to Ionizing Radiation for Professional Radiation Workers

organ exposed	exposure in Rem*
Whole body	3 Rem in 13 weeks
Professional emergency dose (once in life)	25 Rem
Accumulated dose under 18 years	no professional exposure permitted
Accumulated dose over 18 years	average of 5 Rem per year, less than 15 Rern in any single year
Skin	8–10 Rem in 13 weeks or 30 Rem per year
Gonads	5 Rem per year
Bone	10 Rem in 13 weeks or 30 Rem per year

*One Rern denotes the dose of radiation of various kinds that cause effects equivalent to one rad of X-rays.

The limiting factor for persons in the general population is based on the genetic hazard. The greater worldwide concern for these heritable, nonthreshold types of radiation damage is expressed in the recommendation of the International Commission on Radiological Protection: for individuals not engaged in radiation–work, the exposure to ionizing radiation should not exceed five Rern in a lifetime, in addition to natural background radiation and to the lowest practical dosages from medical exposure. The permissible exposure levels are being constantly reviewed and adjusted as new evidence becomes available on biological effects.

Rontgec's discovery of X-rays

Historical background. In December 1895 the German physicist Wilhelm Conrad Rontgen demonstrated the first X-ray pictures, among them that of the left hand of Mrs. Rontgen. Within a few weeks the news of the discovery spread throughout the world, and the penetrating properties of the rays were soon exploited for medical diagnosis without immediate realization of possible deleterious effects. The first reports of X-ray injury to various human tissues and to vision came in 1896. In that same year Elihu Thomson, the physicist, deliberately exposed one of his fingers to X-rays and provided accurate scientific observations on the development of roentgen-ray burns.

Also in 1896, Thomas Alva Edison discovered fluoroscopy; he was engaged in developing a fluorescent roentgen-ray lamp when he noticed that his assistant, Clarence Dally, was so "poisonously affected" by the new rays that his hair fell out and his scalp became inflamed and ulcerated. By 1904 Dally developed severe ulcers on both hands and arms; these lesions, which soon afterward became cancerous, caused his early death. During the next few decades, many investigators and medical doctors developed radiation burns and cancer, and more than 100 persons died, presumably as a result of their exposure to X-rays. These sad early experiences eventually led to an awareness of radiation hazards for professional workers and stimulated development of a new branch of science —radiobiology.

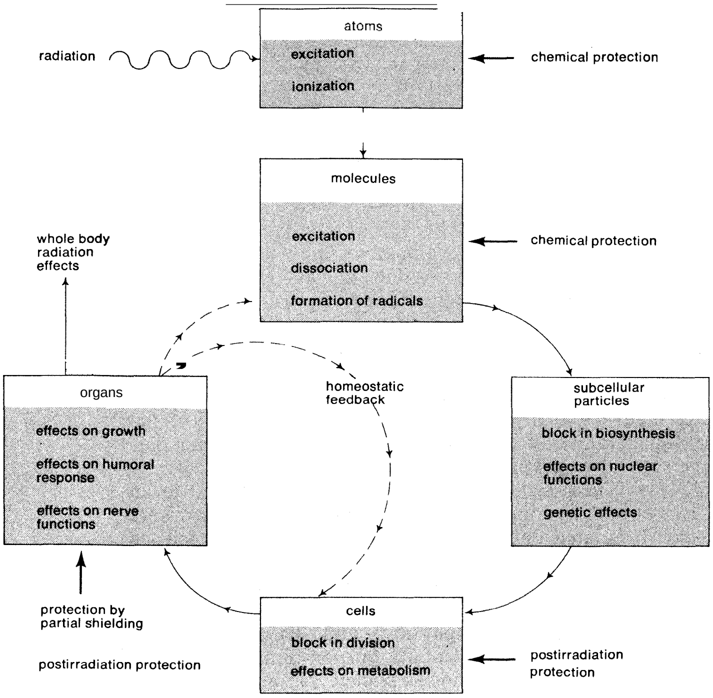
Radiations from radioactive materials were not immediately recognized as being related to X-rays. In 1906 A.-H. Becquerel, the French physicist and discoverer of radioactivity (1896), accidentally burned himself by carrying radioactive material in his pocket. Noting that, Pierre Curie, the codiscoverer of radium, deliberately produced a similar burn on himself. A few months later it was found that radium could be useful in medicine; this discovery led to the founding of the Radium Hospital in Paris in 1906. It was also soon recognized that radium could be very toxic. Beginning about 1925, a number of women in the paint industry, who were exposed to luminescent paint containing radium, became ill with anemia and lesions of the jawbones and mouth; some of these persons later developed bone cancer. The same symptoms appeared in some patients who had received radium internally to relieve arthritis and other diseases. In the 1930s attention was called to these radium hazards, and the practices were stopped.

Accelerator generation of radiation

In 1933 Ernest Lawrence and his collaborators completed the first full-scale cyclotron at the University of California at Berkeley. This machine was a copious source of the neutron rays that had at that time just been discovered by Sir James Chadwick in England. This time, however, human guinea pigs were spared. Ernest, his brother John, and Paul Aebersold exposed rats to the beam of the cyclotron; they found that fast-neutron radiation was about two and a half times more effective in killing power than were X-rays. These facts indicated the need for protection for investigators engaged in research with the cyclotron; thus, a form of shielding was devised for the machine. When the first atomic reactor, built in Chicago, proved Enrico Fermi's principle of self-sustaining fission chain reactions (1942), considerable knowledge of the biological effects of neutrons was already available. Atomic reactors are now rapidly becoming a prime source of power for the world.

Modern accelerators produce a wealth of different types of radiations, including leptons (electrons, positrons, muons, and neutrinos) and hadrons. There are more than 100 known baryons and about 60 known mesons, each with different interaction properties. In addition, since the advent of spaceflight, the effects of certain space radiations on man became of interest. Protons in the Van Allen radiation belts and protons and heavier ions in solar flares and near the top of the atmosphere are of particular importance. As a result, health physics (including radiation shielding and health protection) has become an elaborate discipline.

Radiation effects on cells. To understand radiation effects, it is necessary to study the cellular and subcellular events that succeed a radiation insult and that give rise to gross effects. The work done in this area has already contributed a great deal to the understanding of normal and diseased states. How radiation may inflict biological damage is shown in the accompanying diagram.



General scheme of events *in* radiation action.
Radiation acts on atoms, which in turn cause modification in molecules, cells, and organs as shown. The existence of homeostatic regulation in human body causes highly complex "feedback" interactions. Levels at which various protective agents act are also shown.

Determination of cellular radio-sensitivity

The radiosensitivity of individual human cells in laboratory cultures was first measured in the 1950s on the progeny of human cervical cancer cells (of the strain named HeLa) kept alive in tissue culture. An amazing aspect of radiation effects is that the actual energy needed for producing changes or mutations in genes is exceedingly small—often 1,000 times smaller than chemical or heat energy needed to produce similar changes. HeLa and other available human cells are so sensitive to radiation that half of them are killed by doses of 80 to 300 Rem. The killing effect manifests itself primarily when the cells attempt to divide but cannot do so since their chromosomes are broken and abnormally jointed as a result of radiation. Neuroblasts (nerve-cell progenitors) and cells of the intestinal epithelium, or mucous membrane, and of blood-forming organs exhibit the greatest sensitivity. Rapidly proliferating tissues of the reproductive system are particularly sensitive to radiation. It was demonstrated that spermatozoa in the mouse can be killed by a few roentgens. Given time, however, reproductive organs recover fully from sublethal doses of radiation. Most resistant are adult nerve cells, which owe their resistance to the fact that they do not normally divide; specialized nerve endings and synapses (*e.g.*, in the retina), however, may exhibit great radiosensitivity. At sublethal doses, human cells exhibit delay of cell division following radiation, perhaps as a consequence of a block in the synthesis of new genetic material (deoxyribonucleic acid, or DNA), while the synthesis of proteins goes on. Such cells frequently turn into "giants," having several hundred times normal volume. Late effects—*e.g.*, the initiation of cancerous growth—may have their origin in sublethally injured cells and are in part due to genetic and chromosomal alterations.

Cellular-radiation effects are the complex result of breakage and rejoining of chromosomes and of direct genetic damage of a more subtle nature. The synthesis of new DNA is also usually markedly affected. Effects on membranes and other cell constituents appear to be of secondary importance. In bacterial viruses it is known that radiation can cause breakage in both strands of the DNA molecule, with lethal effect. Many single-strand breaks also occur, but most of these are reparable by enzymes in the cell. Chromosome breakage may be followed by normal or abnormal rejoining of the broken ends. There are mechanisms available in mammalian cells that can repair sublethal injury. Thus, the final effect is dose-rate dependent and is less at low dose rates.

Radiation effects on tissues and organ systems. Since tissues are composed of cells, tissue effects reflect the lethal and sublethal changes in the cells. These initiate a complex defense pattern: cells injured by radiation may release toxic substances; *e.g.*, proteolytic enzymes and nucleases that are able to inflict further injury. The injury is often followed by a regeneration accomplished by increased rate of cell division: the sick cells are often dissolved and removed, and, if recovery is not complete, scar tissue may form.

Inhibition of the proliferation of epithelial tissues with cellular breakdown and increased permeability leads to invasion by pathogens, inflammation, ulceration, loss of fluids, nausea, and diarrhea.

Inhibition of blood formation leads, in the course of days or weeks, to leukopenia (reduction of the number of white cells), which lessens defense against infection; anemia (reduction in the number of red cells), which results in defective oxygen transport; lassitude; anoxia (reduction of oxygen in tissue); bleeding, due to a failure of blood clotting and of platelet synthesis; and some loss of immunity. None of these effects is as yet completely understood by medical science.

Varied effects of radiation

It is known that, because of varied cellular effects of radiation, the function of individual organs can become imbalanced. Since the body has widespread humoral and neuronal interconnections, irradiation of a single organ can modify the functions of the rest of the body. For example, radiation of the spleen and liver causes retardation of nucleic-acid synthesis elsewhere in the body. Another example of this effect is given by irradiation studies on the pituitary of animals and humans. Pituitary irradiation can arrest the growth of a young animal and can affect its sexual development and metabolism. In humans irradiation of the pituitary can arrest proliferation of certain hormone-dependent cancer cells, thereby causing regression of this type of cancer wherever it may be located in the body.

Radiation effects due to whole-body radiation can be lessened by the shielding of particular parts of the body. Shielding of part of the blood-forming system of mice, (*e.g.*, the spleen), increases the radiation tolerance by a factor close to 2. Shielding of other parts of the body also protects, but to a lesser extent. Least protection is afforded by head shielding.

Some investigators claim a chronic deterioration of the central nervous system from moderate dose levels (although others have shown in experiments with monkeys that radiation doses up to lethal levels do not impair learning ability).

Generally speaking, man does not sense a moderate radiation field, though very low doses of radiation, less than one rad, can produce phosphene, a light sensation on the dark-adapted retina. Astronauts on the first spaceship that landed on the moon (Apollo 11, July 20, 1969) observed irregular light flashes and streaks during their flight. These events are probably the result of single heavy cosmic ray particles striking the retina. In some food-preference tests, rats, when given a choice, will avoid radiation fields of even a few roentgens. Three roentgens, probably acting on the olfactory system, are sufficient to arouse a slumbering rat, and a few roentgens can accelerate seizures in mice genetically susceptible to them.

In man and other mammals, the retina and smooth muscle respond to intense pulses of radiation; the cornea and peripheral nerves require much higher dose rates to respond.

Acute lethal effects. Prior to August 6, 1945, there was no known case of widespread loss of human life due to penetrating radiation. On that day the first atomic bomb used in warfare was exploded over Hiroshima causing the death of about 75,000 persons, many of whom died of the effects of radiation. In persons exposed to very large doses of radiation, painful, ugly, and repulsive symptoms occur prior to a delayed death.

During the years since World War II, man has learned to make atomic weapons of more than a thousandfold the power of the Hiroshima bomb, and arsenals of hundreds of such bombs are available in various countries. A study by the United Nations has shown that, if atomic war were to be unleashed in full fury, the effects of the blast, heat, radiation, and fission-product isotopes could kill millions of persons and drastically alter plant and animal populations. This might mean the end of civilization as it is known.

Fatal levels of radiation

In the radiation syndrome leading to death, the greater the dose given, the sooner and more profound the radiation effects. Following a single dose of 400–800 Rem to the whole body, survival is improbable. Very high doses, 5,000 Rem or more, cause immediate and discernible effects on the central nervous system. States of intermittent stupor and incoherence vary with hyperexcitability resembling epileptic seizures. Death is certain within several days This syndrome has been very carefully studied in animals; a human case has been described.

When the dose is between 600 and 1,000 Kem, the earliest symptoms are loss of appetite, nausea, and vomiting, followed by prostration. watery and bloody diarrhea, abhorrence of food, and fever. The blood-forming tissues are profoundly affected, and, in 15 to 30 days, the white-blood-cell count may decrease from about 8,000 to as low as 200 per cubic millimetie. The body loses its defenses against microbial infection, and inflammation of mucous membranes and of intestinal lining may occur. As a result of the reduction in blood platelets, the blood loses its ability to clot, and spontaneous internal or external bleeding may result. Return of the early symptoms, often accompanied by delirium or coma, presages death. Theie can be great individual variation in the symptoms. At lower doses they may be delayed for a few days. Complete loss of hair within ten days has been taken as an indication of the lethal severity of the exposure.

In the dose range 150 to 600 Rem, survival is possible (though in the upper range improbable), and the symptoms appear as above, but in milder form and generally following some delay. Nausea, vomiting, and malaise may begin on the first day, then disappear, and a latent period of relative well-being follow. Anemia and leukopenia set in gradually. After three weeks, internal hemorrhages may occur in almost any part of the body, but particularly in mucous membranes. Susceptibility to infection remains a very great hazard, and some loss of hair occurs. Weight loss, lassitude, emaciation, and fever may last for many weeks before either recovery or death occurs.

Moderate radiation doses

Moderate doses of radiation can interfere seriously with the immunity mechanism. In animals this manifests itself in enhanced sensitivity to bacterial toxins, greatly decreased fixation of antigens, and reduced efficiency of formation of antibodies. Furthermore, antibiotics, unfortunately, appear to have limited effectiveness in combatting postirradiation infections. In the 1960s, plastic isolators became available that allow antiseptic isolation of a person from his environment and thus may protect him from infection from external sources during the period critical for his recovery.

Man is generally able to survive a single dose of less than 150 Rem. The symptoms are similar to those alieady described, but milder and delayed. In doses under 100 Rem, the discernible radiation effects may be so slight that the exposed person is able to continue his normal occupation, though there is measurable depression of his bone marrow, and some persons suffer subjective discomfort from doses as low as 30 Rem. Sublethal doses may have chronic effects many years later.

Prenatal exposure and exposure of children to radiation. During embryonic and fetal development, the human organism is extraordinarily sensitive to changes in the environment that may affect the offspring in spite of the protective mediation of the mother's body. Exposure of the mother to significant doses of radiation may result in chronically expressed injury to her offspring.

Most of the information on the effects of radiation during pregnancy is based on studies with animals. Increased sensitivity is apparent in the period following fertilization. Radiation effects strongly depend on the stage of development of embryo and fetus at the time of exposure. For example, congenital malformations are caused in rats and mice when exposure occurs during periods of initial formation of the organs. Exposure of the rat embryo to 200 rads during early embryonic life is more likely to kill the embryo than to cause congenital malformations, whereas irradiation in late pregnancy is more likely to produce functional abnormalities in the offspring than lethal effects or malformations.

Radiation-caused defects

A wide variety of radiation-induced malformations have been observed in small rodents. Many of these are malformations of the nervous system and include microcephaly (reduced size of brain), exencephaly (part of the brain outside the skull), hydrocephaly (enlargement of the head due to excessive fluid), and anophthalmia (failure of the eyes to develop). Such effects follow doses of 100–200 rads at the appropriate time period. At a dose of 25 rads, only eye anomalies are seen; 20–50 rads in late pregnancy cause disorganization in the microscopic structure of rat brain; ten rads cause some retardation and alteration in the rate of development of the cortex. Functional abnormalities seen in animals after prenatal radiation include abnormal reflexes, restlessness and hyperactivity, impaired learning ability, and susceptibility to externally induced seizures. The malformations induced by radiation are similar to those that can be caused by specific virus infection and certain types of neurotropic drugs, pesticides, and. mutagens.

In human populations, 1 to 2 percent are born with malfunctions of the nervous system. Several studies have established that children born to mothers who received radiation to the pelvic region during pregnancy have much higher incidence of anomalies, particularly of the microcephalic type. A study was conducted among several hundred pregnant women who survived the Hiroshima or Nagasaki atomic explosions. Their children were medically observed for more than 20 years. At the higher dose levels, many of these children developed smaller than normal head size, and among these there was a significant increase in the incidence of mental retardation (10 percent at 100-rad dose). The greatest sensitivity to such effects occurs between the second and sixth months of pregnancy. There is also an increased risk of leukemia and of malignancies for children whose mothers were exposed to radiation during pregnancy. This risk is greater when radiation exposure is prolonged over several weeks: for single doses, the risk of mental retardation is greater.

Radiation exposure of children during the first few years of life causes somewhat similar effects as irradiation of the fetus during late pregnancy. The major source of information is from studies of cases treated for tumours and conditions of the scalp. Therapeutic doses can lead to the induction of tumours, particularly those of the thyroid and of the nervous system, to leukemia, and to retarded mental development. Mongolism (Down's syndrome), a severe form of mental retardation, is associated with a known chromosome abnormality. It probably can be caused by radiation, viruses, and chemical agents.

It is clear from the above that the exposure of pregnant mothers to excess radiation entails risks for their offspring. It is prudent to avoid extensive X-ray diagnosis or therapy unless the mother's health warrants it. At the same time radiation should not be construed as a major current cause of developmental abnormalities or of mental retardation. The latter condition is clearly related to other causes: malnutrition of the mother and the newborn (often related to poverty) and genetic factors are responsible for the great majority of cases.

Protection against external radiation. Great efforts have been made by scientists of many countries to find agents that will increase radioresistance of the body.

Oxygen deprivation and radiation effects

Findings in 1933 that living systems deprived of oxygen are more radioresistant and, in 1942, that newborn rats kept in carbon dioxide atmospheres were twice as resistant as their litter mates kept in air went largely unnoticed until 1950. In that year it was demonstrated that large doses of the amino acid cysteine given prior to irradiation provided substantial protection against radiation effect. Following this discovery, many substances were found with some protective action (Table 10). Although the matter is still under debate, it appears that many of these substances act by producing anoxia or by competing for oxygen with normal cell constituents and radiation-produced radicals. Since anoxia is in itself a highly hazardous physiological state and all protective compounds tried thus far are toxic, it cannot be said that protection of humans by these drugs prior to radiation exposure is a practical matter.

Table 10: Some Chemical Protectors Effective in Mice or Rats

class	specific chemical	effective dose (in milligrams per kilogram of tissue)
Sulfur compounds	glutathione	1,000
	cysteine	1,000
	cysteamine	150
	AET*	350
Hormones	estradiolbenzoate	12
	ACTH	25 for 7 days
Enzyme inhibitors	sodium cyanide	5
	carbon monoxide	by inhalation
	mercaptoethylamine (MEA)	235
	para-aminopropiophenone (PAPP)	30
Metabolites	formic acid	90
Vasoconstrictors	serotonin	50
Nervous system drugs	amphetamine	1
	chlorpromazine	20

*Aminoethylisothiuronium bromide hydrobromide.

Beginning in 1961, some diurnal changes in the radiosensitivity of rodents were reported—an indication that the complex humoral factors responsible for daily rhythms also alter the response of tissues to radiation. On the other hand, a few classes of substances are known that render animals more sensitive to acute radiation effects. Several compounds sensitize mice or rats to radiation. Perhaps the most interesting of these is the hormone thyroxine, a normal secretion of the body. A large class of sensitizers at the cellular level include nucleic-acid analogues (*e.g.*, 5-fluoro-uracil).

It may appear from the foregoing that acute radiation syndrome in mammals is the result of complex interaction from many affected organ systems. Nevertheless, it was shown in the early 1960s that there is some genetic control of radiosensitivity, since susceptible and resistant mouse strains were developed. Inability of recovery of the hemopoietic (blood-forming) system appears to be linked with high radiation sensitivity.

The contribution of germ-free studies

Beginning with the 1950s, the Lobund laboratories of Notre Dame University, which pioneered in the study of germ-free animals, have shown that germ-free mice that spend their entire lives in a sterile environment exhibit greater resistance to radiation than do animals in a normal environment, no doubt due in part to elimination of the hazards of infection.

For many years it was held that cure of radiation disease was hopeless since events were thought to be irreversible once a person had received a lethal dose. It was historically important, therefore, when it was demonstrated in mice that ground substance of embryo, spleen, or bone marrow administered following irradiation allowed the animals to survive what would otherwise be a lethal dose of X-rays. It is hoped that detailed study of this will lead to understanding of various physiological processes.

It has been shown that most of the effect of transfused bone marrow depends on the provision of intact, living cells, which then migrate to the marrow of the irradiated host and proliferate there, repopulating the host's marrow with cells characteristic of the donor. Bone-marrow transfusion between animals of different strains is successful in these cases because the irradiated animal loses its ability to develop antibodies against the injected "foreign" tissue. If the injected bone marrow proliferates rapidly enough, it may save the host from acute lethal effects of radiation.

Mammals that carry successful grafts of bone marrow become chimeras; *i.e.*, organisms producing cells of more than one genotype. Several months later, the transplanted tissue may develop a disease sometimes called wasting disease (or homologous disease). This is an immune reaction between the proliferating donor cells and the irradiated host and is often rapidly fatal.

If bone marrow could be taken with ease (as blood is taken) from normal human volunteers and then cultured and preserved under refrigeration and injected only into immunologically compatible individuals, it might prove to be a powerful therapeutic agent against radiation disease. Transfusion of fresh, normal human bone marrow has already demonstrated some worth in alleviating radiation effects. In 1958 an unfortunate accident occurred at a nuclear reactor in Yugoslavia in which five persons were exposed to near-lethal doses of mixed neutron and gamma radiation. They were flown to Paris, where four of them received human bone-marrow transfusions. As judged by their recovery, several of these men were considerably benefitted. In time it will be feasible to culture cells in the laboratory for transplanting into a victim of radiation sickness. Cultured cells are not yet suitable for the therapy of radiation disease by transfusion.

Organ and tissue transplantation

Bone-marrow transfusion, as well as transplantation of other organs, may be of importance in future control of various seemingly incurable conditions; *e.g.*, leukemia and some types of anemia. The clue to successful bone-marrow transplantation is the complete understanding of the genetic and biochemical aspects of immunological compatibility. Only when this knowledge is attained will transplantation of bone marrow (and also of essential organs such as heart, pancreas, or kidney) become established as a safe and feasible medical practice.

Long-term effects, somatic and genetic. Much of man's radiation exposure is at very low levels. There are some components of natural and man-made radiation (see below) to which persons may be exposed continuously or intermittently through most of their lives. Usually the effects of such low-level radiations are so minute that often even the most refined methods fail to demonstrate immediate functional effects or late deleterious effects that can be unequivocally assigned as the effects of radiation on the individual. Instead, long-term effects must be studied on a statistical basis in a population of individuals, and they are usually described as altered statistical risks for the populations. These risks are: lower average life expectancy, reduction in fertility, acceleration of the rate of mutations, and increase in the occurrence of chronic diseases such as cancer or cataracts. While these effects may be brought about by radiation, they are also affected by a host of environmental and hereditary factors.

Significant doses of radiation delivered daily (*e.g.*, 0.5 rad per day or more) cause rats and mice to become physiologically older than their actual age would indicate, with subsequent shortening of life-span. Some experiments with mice, however, showed a life-span lengthening for 0.1 rad daily dose. It is exceedingly difficult to make valid deductions from animal data to human beings. Data for radiologists, whose radiation exposure is assumed to be greater than that of other segments of the population, provide some information on radiation effects in man. A study of British radiologists showed that their mortality generally was lower than that of other doctors, whereas the mortality of American radiologists was found higher than that of their colleagues in ophthalmology and otolaryngology. A sample of nearly 100,000 survivors of Hiroshima and Nagasaki yielded the anomalous result that groups exposed to doses between 11 and 120 rads actually had a lower death rate in the ensuing 15 years than those receiving a lesser dose. Results such as these point to the fact that there are great variations in life-span due to many factors. Despite conflicting data, for the sake of health-protection estimates, the life-shortening effect of one rad on man may be assumed to lie between two and 15 days.

There are several concepts of the nature of radiation-induced aging. One theory maintains that all disease ex-

Radiation-induced aging

periences from early childhood, as well as toxic factors in the environment, act together to determine the physiological age of an individual. According to this model, the fallout-radiation hazard currently is much less of an aging factor than cigarette smoking, lack of exercise, obesity, or childhood diseases. Another concept, one more specific to radiation injury, states: the aging effect produced by radiation is due to the number of irreversible chromosome aberrations produced in the somatic (tissue) cells of mammals. Many of the chromosome aberrations can be demonstrated in blood cells or in liver cells several months following exposure to radiation, and chromosome aberrations have generally been associated with impaired cellular function. Neither of these theories has thus far been able to account fully for the modifying effects of the hormonal system.

Radiation effects on fertility

A dose of a few hundred Rem is believed unlikely to impair the fertility of normally fertile women, and it is believed that up to 15 Rem yearly will not impair the fertility of normally fertile men. The doses that cause genetic effects to the population are delivered in the years prior to conception of children. Although there are many genetic data for mice, it is difficult to arrive at reliable genetic estimates for man, since genetic effects occur at the chromosome as well as the gene level. H.J. Muller, the discoverer of mutation induction by radiation, suggested that recessive lethal mutations constitute the greatest genetic hazard; they can penetrate the population in the course of many successive generations and exert their toll, mainly when they appear in individuals born to parents who both carry the recessive lethal gene. Furthermore, certain chromosome mutations (*e.g.*, when chromosomes are present in abnormal numbers) also represent important genetic risks. In any case, it appears that lethal mutations exert their greatest effect in the first generation born subsequent to radiation exposure but that in subsequent generations more lethals will appear due to the same recessive genes. In a population of 1,000,000 persons who have been exposed to an average dose of one Rem, about 2,000 fatalities will be observed in the course of ten successive generations. This is, however, less than 0.1 percent of the fatalities that will occur in the same population in the same time due to the normally present "genetic load of lethal genes in the population. The normal lethal genetic load will approximately double when a population is exposed to a dose of 60 Rem. Such estimates usually are made by assuming that all radiation doses are strictly additive and that the effect is proportional to the total dose. Evidence has been mounting, however, that genetic radiation damage is almost completely repaired in the female reproductive system at low dose rates of X- or gamma rays. The effect on the male reproductive system may be less than previously assumed, due to the fact that normal sperm has better viability and motility than sperm carrying radiation-induced damage. Genetic effects from neutrons have been reported to repair less than effects from gamma rays; neutrons also are responsible for many more chromosome aberrations.

"Good" and "bad" mutations

It is generally believed that most of the radiation-induced mutations are "bad" in the sense that they are lethal or usually impair some property. It is also known, however, that a fraction of induced mutations are "good" and that mutations are essential to maintain a steady rate of evolutionary change. Evolutionary changes occur at random and tend to better adapt the species to its environment. Some data on populations of fruit flies indicate that radiation can be useful in accelerating the rate of evolution when the population faces an environmental crisis.

Radioisotopes and fallout. Radioactive isotopes usually emit electrons or positrons, alpha particles or gamma rays, or even characteristic X-rays. The exposure may be external, in which case penetration is an important factor. Alpha particles do not penetrate deeply enough in the skin to cause damage. Beta particles or X-rays over 30 kilovolts can be hazardous to the skin, causing redness, loss of hair, or ulceration. Large doses may cause skin cancer.

Isotopes can also enter the body by ingestion, inhalation, or injection. Their radiation effects then depend on their internal distribution, duration of retention in the body, energies, and rate of radioactive decay (half-life or half-value period). The problem is enormously complicated since isotopes have different and sometimes elaborate distribution patterns.

ACCUMULATION OF ISOTOPES IN CRITICAL ORGANS

The term critical organ has been assigned to that part of the body that is most vulnerable to a given isotope. The critical organ for plutonium, radium, strontium, and many other fission products is bone and the adjacent bone marrow. For soluble fission products (and also some forms of uranium), which will distribute in the entire body, the critical organs are the gonads, or sometimes the kidneys. For iodine the critical organ is the thyroid gland. Insoluble airborne radioactive dust often deposits in the alveoli of lungs, while colloidal particles of very small size can reach the bone marrow, liver, or spleen. Table 11 is an abbreviated list of the maximum permissible concentrations (U.S. recommendations) for man of some radioisotopes. (Maximum permissible concentration is the greatest amount of a radioisotope that can be accumulated in the body without producing noticeable damaging effects; compare with maximum permissible dose, which is applicable to radiation received from external sources.)

Table 11: Maximum Permissible Concentration (MPC) of Some Radioisotopes

isotope	chemical form	critical organ	microcuries in body
Tritium (hydrogen-3)	water		2,000
Carbon-14	carbon dioxide		400
Strontium-90*	water soluble salt		40
		bone	4
Iodine-131	water soluble salt		50
		thyroid	0.7
Cesium-137	water soluble salt		30
Radon-2221	gas		
Radium-226‡	water soluble salt		0.2
		bone	0.1
Uranium	water soluble salt		0.2
		kidney	0.005
Plutonium-239	water soluble salt		0.4
		bone	0.04

*MPC in drinking water: 0.001 microcurie per litre. †MPC in air: 0.00001 microcurie per litre. ‡MPC in drinking water: 0.0001 microcurie per litre.

Since isotopes continuously deliver radiations to the surrounding tissue, one must distinguish between the effect of protracted continuous exposure and simple acute or periodically repeated exposures. For beta and gamma radiations and X-radiation, utilizing split-dose delivery, it has been found that up to about 60 percent of an acute radiation effect "disappears" within several hours; the body, therefore, is able to tolerate a greater total dose when the dose is protracted or when part of it is given at later times. For neutron and alpha radiation the recovery is less. (Neutrons are generally more effective agents of mutation than are X-rays: for acute effects, by a factor of 1 to 8; for long-term effects of chronic radiation, by a factor up to 35.) Studies of such repair processes are complicated by age effects of persons exposed.

Fallout is the deposition of airborne radioactive contamination on earth. Radioisotopes may be produced naturally in the air by cosmic radiation or may enter the air from stack gases of atomic reactors following industrial accidents or from bombs or bomb tests. After 1954 bomb tests carried out by several nations produced fallout measurable on the surface of the entire world, arousing great attention and controversy with respect to its health and genetic effects. It is within the realm of practical feasibility to cover a significant portion of the Earth's surface with fallout in time of war. While much of the bomb hazard is due to blast waves and heat, the radiation dose from fission products could be so intense that only persons remaining in underground shelters for some weeks could hope to survive such an attack. Usually the most prominent isotopes in fallout are fission products; however, all materials exposed to nuclear blasts may become radioactive. A list of exposures to fallout isotopes is given in Table 7.

The hazard of long-lived isotopes

Several of the radioisotopes contained in fallout are especially hazardous because they remain radioactive for relatively long periods (have long half-lives). Cesium-137, strontium-90, and plutonium may be the most important of these. On the ground, fallout material can cover external surfaces and foliage and later be washed into the soil, where plants incorporate strontium-90, along with the chemically similar calcium, and cesium-137 with potassium. Humans obtain these radioactive materials mostly from drinking water and from plant and animal foods, including milk. In the sea much of these materials can eventually lodge in the bodies or skeletons of fishes and in plants near the coast.

Strontium-90 becomes concentrated in bone and remains there in steadily decreasing amounts for almost 30 years, producing local irradiation. Its actual concentration within the body is difficult to measure because of the softness of its rays. There is generally lower concentration of strontium-90 in man than in other animals or in plants or soil. Newborn babies, who have actively growing bones, retain relatively greater amounts of strontium-90 than do older persons.

The most easily detectable fallout product in animals and in man is iodine-131; this isotope emits beta and gamma rays and is enriched about 100 times in the thyroid gland through selective accumulation. Because of its relatively short half-life (eight days), iodine-131 is probably not the most hazardous fallout isotope. If a large population, however, is affected by excessive amounts, radiations from this isotope could lead to metabolic disturbances and increased thyroid-cancer incidence, especially in children.

Discharges of atomic-power plants

Atomic-power reactors discharge a mixture of radioactive gases into the atmosphere. Reactors are placed at sites where the atmospheric mixing and transport are such that the short-lived gases decay and are diluted before they can be inhaled in appreciable amounts by the population. The long-lived gaseous products are usually quoted in units of krypton-equivalent. When inhaled, krypton-85, a radioisotope of ten years half-life (half the radioactivity will disappear in ten years), is preferentially deposited in body fat. Projected krypton-equivalent concentrations from power reactors for the end of this century appear high, suggesting the need for better containment of the gaseous products of reactors in the future (Table 8).

Many fallout isotopes that reach the sea and inland waterways eventually end up in concentrated form in the bodies of waterborne animals and plants and become a source for concern when they are part of the food chain for man. Radioactive iodine, for example, has shown up in many fish and shellfish.

Methods that have been developed for biological protection from fallout extend from efforts for keeping isotopes out of the body to biochemical means for rapid elimination of isotopes from tissues. At times of atomic emergencies, airborne radioactive particles may be kept from the lungs by masks having suitable filters. Ingested isotopes may be prevented from being absorbed in the intestinal tract by certain mucoprotein substances that show great surface affinity for adsorption of strontium and other substances. In 1964 Canadian investigators demonstrated that sodium alginate prepared from the brown seaweed kelp is particularly useful in this regard. It is possible to remove virtually all radioactive strontium from cow's milk without affecting the essential nutritive components. Certain chelates—for example, EDTA (ethylenediaminetetraacetic acid)—will react with strontium and "cover" this atom. As a result, when EDTA is present in the blood, deposition of strontium in bones is reduced (elimination of already deposited isotopes also is somewhat accelerated). Most chelating agents are not specific for strontium; unfortunately, they also chelate the closely related and important element calcium.

Radiation and cancer. The role of large doses of radiation in increasing the incidence of cancer and leukemia cannot be questioned—it has been demonstrated in all animal species tested thus far. Studies of animal populations have indicated that the onset of cancer can be delayed by many years following radiation exposure and that protracted radiation exposure is more carcinogenetic than single exposures. Initially, radiation may produce precancerous lesions in the cells of the body, but it may take some additional factors to turn those lesions into malignant tumours.

It is interesting to note that radiation-induced leukemia in certain strains of mice can later be propagated to other members of the mouse population by a virus factor. It appears that the role of radiation in this instance might be to activate a leukemia virus already present in mice.

Radiation as one of many factors in cancer induction

Generally speaking, the increased cancer incidence in irradiated populations parallels radiation-induced aging and may be a manifestation of the older physiological age of the irradiated individuals. Studies of the origin of cancer point to a number of factors, including genetic, hormonal, and injurious factors (chemical, radiation, or virus), that usually act together to produce cancer. The great majority of cancer and leukemia cases in the Earth's population is probably not caused by radiation but by other agents. Nevertheless, the following generalizations are true in most instances:

1. There is often a long delay between radiation exposure and the appearance of tumours (in human beings the delay may be many years).
2. Higher radiation doses usually lead to more tumours, and sooner, than do lower doses.
3. Tumours often arise from precancerous groups of cells, not necessarily from a single cell.
4. Densely ionizing radiation (*e.g.*, neutrons, alpha particles) are more effective in inducing cancer than lightly ionizing X- or gamma rays.

The relationship of radiation to human cancer has been studied almost exclusively in persons who, by their occupations, are exposed to high radiation doses. Cancer induction from radiation below the permissible exposure levels has not been identified. It is the task of health-protection experts and special government agencies (*e.g.*, the U.S. Federal Radiation Council) to set the permissible radiation levels so that the risk for cancer induction is either negligible compared to other risks or at least tolerable in view of the beneficial effects of radiation to humanity.

Radiation induction of leukemia

The role of induction of leukemia in man at doses higher than 50 rads appears proportional to dose. Only acute leukemia and chronic myeloid leukemia seem to be induced. A number of such leukemias appeared among the survivors of Hiroshima and Nagasaki and in a special group of British patients who received large doses of radiation for the rheumatic condition of ankylosing spondylitis. From these data it appears that one rad induces 1.5 cases of leukemia per 1,000,000 individuals per year, or a total of 20 cases per rad per 1,000,000. The incidence in children, exposed under ten years of age, is about twice as high; the induction of acute lymphatic leukemia is particularly age dependent.

In addition to leukemia, other malignancies also are induced. After large doses of radiation, the number of all induced cases combined appears to be about twice the number of leukemias. There are many questions to be settled in connection with these figures; among these are the problem of accumulation of low daily doses and whether or not cancers that have a high spontaneous incidence are also more prone to be induced by radiation.

Incidence of bone and lung tumours

The human population carries a body burden of about 0.0001 microcurie of radium and has a natural incidence of four to seven bone sarcomas per 1,000,000 individuals per year. Radium, an alpha-particle emitter, has somewhat similar chemical properties as calcium, and it remains deposited in bone for many years. Some hundreds of individuals exist who have several microcuries in their bodies; these were radium-dial painters or individuals who may have received radioactivity as a therapeutic measure (in the 1930s) for arthritis and other conditions.

The incidence of bone tumours in this population appears to be higher when the body burden is more than 0.5 microcurie.

It has been known for some time that radium and uranium miners have a high incidence of lung tumours. Radioactivity can find its way to the lungs by inhalation of gas (radon) or of radioactive-dust particles that may become lodged in the lungs. Since a single dust particle can give off significant local radiation, the radioactive-dust content of the mines must be kept at a safe level. In the initiation of lung tumours, several factors interact; nonradioactive dust can cause silicosis and, perhaps, some cancer. Smoking and radioactive-mine exposure caused more lung cancer than either of these factors singly.

Chromosome aberrations and radiation. It has long been known that a major expression of radiation effect occurs in the form of chromosome aberrations. These have been studied in human-cell cultures, prepared from a variety of organs, and in leucocytes, or white blood cells. Chromosome aberrations, at the rate of a few cases per 1,000 cells, can be found in human blood from very low doses of radiation of only a few rads, and some of these aberrations persist in the body for several years. With the establishment of accurate dose-effect relationships, the frequency of certain types of aberrations can give information on the degree of radiation exposure of individuals.

In addition to causing chromosome breaks and rejoinings, radiation can also produce aneuploidy (abnormal number of chromosomes in cells) and polyploidy (multiple chromosome numbers).

Tumour cells in many forms of cancer also carry chromosome aberrations, and, in advanced forms of cancer, there is often extensive fragmentation and regrouping of chromosomes. Whether radiation-induced chromosome breaks cause cancer or are a consequence of cancer is not known with certainty.

There is only one type of human neoplasm, chronic granulocytic leukemia, where a specific chromosome abnormality is seen in over half of the cases. Chromosome 21, the "Philadelphia" chromosome, named after the city where it was discovered, lacks about one-half of its longer arm. Other conditions where specific chromosome abnormalities are seen include Down's syndrome (see above), Waldenstrom's macroglobulinemia, Burkitt's lymphoma, and multiple myeloma. The study of chromosome aberrations in human cells and tissues is extremely tedious because of the smallness of chromosomes and the need to carefully observe several thousand in each individual case. The use of computers may help to differentiate normal from abnormal patterns.

EFFECTS OF HERTZIAN WAVES, INFRARED RAYS

Effect of long radio waves

Hertzian waves. The effects of Hertzian waves (electromagnetic waves, radio waves, and microwaves or radar waves) and of infrared rays usually are regarded as equivalent to the effect produced by heating. The longer radio waves induce chiefly thermal agitation of molecules and excitation of molecular rotations, while infrared rays excite vibrational modes of large molecules and release fluorescent emission as well as heat. Both types of radiation are preferentially absorbed by fats containing unsaturated carbon chains.

The fact that heat production resulted from bombardment of tissue with high-frequency alternating current (wavelengths somewhat longer than the longest radio waves) was discovered in 1891, and the possibility of its utilization for medical purposes was realized in 1909, under the term diathermy. This method of internal heating is beneficial for relieving muscle soreness and sprain. Diathermy can be harmful, however, if so much internal heat is given that the normal cells of the body suffer irreversible damage. Since man has heat receptors primarily in his skin, he cannot be forewarned by pain when he gets a deep burn from diathermy. Sensitive regions easily damaged by diathermy are those having reduced blood circulation; *e.g.,* bone or the lens of the eye. Cataracts of the eye lens have been produced in animals by microwave radiation applied in sufficient intensity to cause thermal denaturation of the lensprotein.

Microwaves and their application

Recently, microwave ovens have found widespread use in commercial kitchens and some homes. These can heat and cook very rapidly and, if used properly, constitute no hazard to operators. In the radio–television industry and in the radar division of the military, persons are sometimes exposed to high densities of microwave radiation. The hazard is particularly pronounced with exposure to masers, capable of generating very high intensities of microwaves (*e.g.,* carbon dioxide masers). The biological effects depend on the absorbency of tissues. At frequencies higher than 150 megahertz, significant absorption takes place. The lens of the human eye is most susceptible to frequencies around 3,000 megahertz, which can produce cataracts. At still higher frequencies, microwaves interact with superficial tissues and skin, in much the same manner as infrared rays.

Acute effects of microwaves become significant if a considerable temperature rise occurs. Cells and tissues eventually die at temperatures of about 43° C (109" F). Microwave heating is minimized if the heat that results from energy absorption is dissipated by radiation, evaporation, and heat conduction. Normally one-hundredth of a watt (ten milliwatts) can be so dissipated, and this power limit generally has been set as the permissible dose. Studies with animals have indicated that, below the permissible levels, there are negligible effects to various organ systems. Microwaves or heat applied to testes decrease the viability of sperm. This effect, however, is not significant at the "safe" levels.

Some investigators in the Soviet Union have documented a variety of nonthermal effects of microwaves and recommend about 1,000 times lower safe occupational dose levels than are in force in the United States. Most prominent among the nonthermal effects appear to be those on the nervous system. These have resulted in untimely tiring, excitability, and insomnia registered by persons handling high-frequency radio equipment. Nonthermal effects have been observed on the electroencephalogram of rabbits. These may be due to changes in the properties of neural membranes or to denaturation of macromolecules.

The heating effect of high-frequency waves and currents has had some use in surgery. The first surgical use of shortwaves (eight megacycles) was made in destroying a patient's tonsils without the necessity of giving an anesthetic. This procedure produces cutting by heat coagulation of tissue without producing bleeding. It has been used to cut certain structures (*e.g.,* the dura membranes) in brain surgery and in minor skin operations. Higher frequency electromagnetic waves (about ten megahertz) are also used in brain surgery for producing "heat lesions" in the treatment of Parkinson's disease and cerebral palsy. Such heat lesions produced in certain parts of the brain may temporarily relieve the tremors accompanying these diseases. (For laser therapy, see below Effects of lasers.)

Hazards of infrared radiation

Infrared rays. An important part of solar energy reaches the earth in the form of infrared rays. Absorption and emission by the human body of these rays plays an important part in temperature exchange and regulation of the human body. Principles of infrared emission and absorption must be considered in the design of air conditioning and of clothing.

Overdosage of infrared radiation, usually resulting from direct exposure to a hot object (including heating lamps) or flame, can cause severe burns (flame burns). While infrared exposure is a hazard near any fire, it is particularly dangerous in the course of atomic chain reactions. In the course of an atomic detonation, a brief but very intense emission of infrared occurs, together with visible and ultraviolet light emitted from the fireball (flash bums). Of the total energy of nuclear explosion, as much as one-third may be in the form of thermal radiation, moving with the velocity of light (186,000 miles per second). The rays will arrive almost instantaneously at regions removed from the source by only a few miles. Smoke or fog can effectively scatter or absorb the infrared components, and even thin clothing can greatly reduce the severity of bum effects. See also BURNS.

EFFECTS OF VISIBLE AND ULTRAVIOLET LIGHT

Life could not exist on Earth without light from the sun. Plants utilize the energy of the sun's rays in the process of photosynthesis to produce carbohydrates and proteins, which serve as basic organic sources of food and energy for animals. Light is a powerful regulating influence on many biological systems. Most of the strong ultraviolet rays of the sun, which are hazardous, are effectively absorbed by the upper atmosphere. At high altitudes and near the Equator, the ultraviolet intensity is greater than at sea level or at northern latitudes.

Very shortwave ultraviolet light, below 2200 angstroms (A), is highly toxic for cells; in the intermediate range, the greatest killing effectiveness on cells is at about 2600 angstroms (one angstrom equals $\frac{1}{10,000,000}$ millimetre). The nucleic acids of the cell, of which the genetic material is composed, strongly absorb rays in this region. This wavelength, easily available in mercury vapour, xenon, or hydrogen arc lamps, has great effectiveness for germicidal purification of the air.

Since penetration of visible and ultraviolet light in body tissues is small, only the effects of light on skin and on the visual apparatus are of consequence. When incident light exerts its action on the skin without additional external predisposing factors, scientists speak of intrinsic action. In contrast, a number of chemical or biological agents may condition the skin for action of light; these latter phenomena are grouped under photodynamic action. Visible light, when administered following lethal doses of ultraviolet, is capable of causing recovery of the cells exposed. This phenomenon, termed photorecovery, has led to the discovery of enzyme systems capable of restoring damaged nucleic acids in genes to their normal form. It is probable that photorecovery mechanisms are continually operative in some plants exposed to direct action of sunlight.

The surface of the Earth is protected from the lethal ultraviolet rays of the sun by the top layers of the atmosphere, which absorb far ultraviolet, and by ozone molecules at the stratosphere, which absorb most of the near ultraviolet. Even so, it is believed that an enzymatic mechanism operating in the skin cells of individuals continually repairs the damage caused by ultraviolet to the nucleic acids of the genes. Some scientists believe that exhaust gases from supersonic planes flying above the stratosphere could diminish the thickness of the ozone layer, thus exposing persons to more intense ultraviolet radiation at ground level. Others fear that pollution produced at ground level may increase the ozone layer and eventually diminish the overall intensity of light radiation.

Quality of light and its effects

There is some evidence to indicate that not only overall light intensity but also special compositions have differential effects on organisms. For example, in pumpkins, red light favours the production of pistillate flowers, and blue light leads to development of staminate flowers. The ratio of females to males in guppies is increased by red light. Red light also appears to accelerate the rate of proliferation of some tumours in special strains of mice. The intensity of incident light has an influence on the development of light-sensing organs; the eyes of primates reared in complete darkness are much retarded in development.

Intrinsic action. Light is essential to man because of its biosynthetic action. Ultraviolet light induces the conversion of ergosterol and other vitamin precursors present in normal skin to vitamin D, an essential factor for normal calcium deposition in growing bones. While some ultraviolet light appears desirable for formation of vitamin D, an excess amount is deleterious. Man has a delicate adaptive mechanism that regulates light exposure of the more sensitive deeper layers of his skin. The transmission of light depends on the thickness of the upper layers of the skin and on the degree of skin pigmentation. All persons except albinos are born with varying amounts of melanin pigment in their skins. Exposure to light further enhances the pigmentation already present and can induce production of new pigment granules. The therapeutic possibilities of sunlight and ultraviolet light became apparent around 1900, with popularization of the idea that exposure of the whole body to sunlight promotes health.

By that time, it was already known that large doses of ultraviolet radiation caused sunburn, the wavelength of about 2800 angstroms being most effective. It induces reddening and swelling of the skin (due to dilation of the blood vessels), usually accompanied by pain. In the course of recovery, epidermal cells are proliferated, melanin is secreted, and the outer corneal layer of dead cells is thickened. In 1928 it was first shown clearly that prolonged or repeated exposure to ultraviolet light leads to the delayed development of skin cancer. The fact that ultraviolet light, like X-radiation, is mutagenic (induces mutations) may explain its ability to cause skin cancer, but the detailed mechanism of cancer induction is not yet completely understood. There seems very little doubt, however, that skin cancer in man is in some cases correlated with prolonged exposure to large doses of sunlight. Among Negroes who are protected by rich melanin formation and thickened corneal structure of the skin, incidence of cancer of the skin is several times less frequent than it is among white people living at the same latitude.

Light sensitization

Photodynamic action. There are a number of diseases in man and animals in which light sensitivity is involved; for example, hydroa, which manifests itself in blisters on parts of the body exposed to sunlight. It has been suggested that this disease is due to a light-sensitive porphyrin compound found in the blood.

Actually there are many organic substances and various materials of biological origin that make cells sensitive to light. When eosin is added to a suspension of human red blood corpuscles exposed to light, the red corpuscles will break up in a process named hemolysis. Other typical photodynamic substances are rose bengal, hematoporphyrin, and phylloerythrin—all are dyes capable of fluorescence. Their toxicity manifests itself only in the presence of light and oxygen.

Some diseases in domestic animals result from ingestion of plants having photodynamic pigments. For example, St. Johnswort's disease is caused by the plant *Hypericum.* Fagopyrism results from eating buckwheat. In geeldikopp ("yellow thick head"), the photodynamic agent is produced in the animal's own intestinal tract from chlorophyll derived from plants. In humans the heritable condition of porphyria frequently is associated with light sensitivity, as are a number of somewhat ill-defined dermatological conditions that result from exposure to sunlight. The recessively inherited rare disease xeroderma pigmentosum also is associated with light exposure; it usually results in death at an early age from tumours of the skin that develop on exposed areas. The cells of such individuals possess a serious genetic defect: they lack the ability to repair nucleic-acid lesions caused by ultraviolet light.

Certain drugs (*e.g.*, sulfanilamide) sensitize some persons to sunlight. Many cases are known in which ingestion of or skin contact with a photodynamic substance was followed by increased light sensitivity.

Effects on development and biological rhythms. In addition to its photosynthetic effect, light exerts an influence on growth and spatial orientation of plants. This phototropism is associated with yellow pigments and is particularly marked in blue light. The presence of illumination is a profound modifier of the cellular activities in plants as well. For example, while some species of blue-green algae carry out photosynthesis in the presence of light, they do not undergo cell division.

Diffuse sensitivity to light exists also in several phyla of animals. Many protozoans react to light. Chameleons, frogs, and octopuses change colour under the influence of light. Such changes are ascribed to special organs known as chromatophores, which are under the influence of the nervous system or endocrine system. The breeding habits and migration of some birds are set in motion by small consecutive changes in the daily cycle of light.

Light is an important controlling agent of recurrent daily physiological alterations (circadian rhythms) in many animals and probably in humans (see PERIODICITY, BIOLOGICAL). Lighting cycles have been shown to be important in regulating several types of endocrine function: the daily variation in light intensity keeps the secretion of adrenal steroids in synchrony; the annual breeding cycles of many mammals and birds appear to be regulated by light. Ambient light somehow influences the secretions of a tiny gland, the pineal body, located near the cerebellum. The pineal body, under the action of enzymes, produces melanotonin, which in higher concentrations slows down the estrous cycle; low levels of melanotonin, caused by exposure of animals to light, accelerates estrus. It is believed that light stimulates the retina, and information is then transmitted by sympathetic nerves to the pineal body.

Effects of light on biological rhythms

Effects on the *eyes*. The wavelength of light that produces sunburn can also cause inflammation of the cornea of the eye. This is what occurs in snow blindness or after exposure to strong ultraviolet-light sources. Unusual sensitivities have been reported. Ultraviolet light, like infrared or penetrating radiations, can also cause cataract of the eye lens, a condition that is characterized by denatured protein in the fibrous cells forming the lens. The retina usually is not reached by ultraviolet light, but large doses of visible and infrared light can irreversibly bleach the visual pigments, as in sun blindness. Numerous pathological conditions of the eye are accompanied by abnormal light sensitivity and pain, a condition that is known as photophobia. The pain appears to be associated with reflex movements of the iris and reflex dilation of the blood vessels of the conjunctiva. Workers in situations where they can expect to be exposed to ultraviolet-light sources or to atomic flashes should wear protective glasses.

EFFECTS OF LASERS

Masers and lasers are devices capable of generating exceedingly intense microwave and light beams, respectively, that travel in narrow beams. Laser beams can carry radio broadcasts and have made possible refined studies of the surface of the Moon. They also have potentially useful applications in biology and medicine.

The destructive action of lasers

The highly intense laser beam can instantly vaporize the surface of a target. This has led to its application in research as well as in surgery. The laser microprobe is used for microanalysis of surface composition. Laser beams have been found to have a selective effect on cellular components, or organelles: those components that absorb light of the wavelength of the beam are destroyed, whereas transparent parts of the cells remain unaffected. Organelles such as mitochondria, which are responsible for cell respiration, or chloroplasts, which are involved in plant-cell photosynthesis, can be separately studied in this manner.

The use of special dyes can alter laser action. The availability of high-pulse-intensity laser beams is also revolutionizing medical photography and microscopy. It is becoming possible to photograph microaction in a very small fraction of a second and to use the relatively new technique of holography for image synthesis.

It was found that, at the appropriate intensity, initial local thermal effects of a laser beam are followed by formation of scar tissue. This property is used in microsurgery of the eye, particularly for treating retinal detachment. Since light can be focussed to a small spot on the retina, damage to lens and cornea can often be avoided; vascular tumours and retinoblastomas also have been treated. In performing such procedures, care must be taken to avoid hemorrhage; the nature and extent of scars and the hazard of developing cataract must be also considered.

Lasers have been used in investigative therapy for a number of conditions of the skin, including hemangiomas and various types of skin cancer. The cells of malignant melanoma tumours carry a dark pigment, which, it is believed, might selectively absorb laser wavelength and thereby destroy the cell. Nerve cells are more sensitive to laser beams than are a variety of the supporting cells, suggesting neurosurgical applications. Laser treatment of the surface of teeth may have an eventual role in the control of tooth decay.

BIBLIOGRAPHY. General information is given in C.W. SHILLING (ed.), ***Atomic Energy Encyclopedia in the Life Sciences*** (1964); and, on the effects of atomic weapons, by S. GLASSTONE and P.J. DOLAN (eds.), ***The Effectsof Nuclear Weapons,*** 3rd ed. (1977). The national and international nuclear science field is monitored by the INTERNATIONAL ATOMIC ENERGY AGENCY (IAEA), ***INIS Atomindex*** (semimonthly), which includes abstracts and indexes of U.S. government reports and other documentation. Early history of the development of the field is found in O. GLASSER, ***Wilhelm Conrad Röntgen and the Early History of the Roentgen Rays*** (1933); and P. BROWN, ***American Martyrs to Science Through the Roentgen Rays*** (1936). Historically important technical books on radiobiology are D.E. LEA, ***Actions of Radiations on Living Cells*** (1946); and A. HOLLAENDER (ed.), ***Radiation Biology,*** 3 vol. (1954–56). Radiological units and measurements are discussed in R.E. LAPP and H.L. ANDREW$ ***Nuclear Radiation Physics,*** 4th ed. (1972), and in publications of the International Commission on Radiation Units and Measurements (ICRU). Introductory information on radiation biology is given in J.E. COGGLE, ***Biological Effectsof Radiation*** (1971); and E.J. HALL, ***Radiarion and Life*** (1976). More advanced texts include: M. ERRERA and A.G. FORSSBERG (eds.), ***Mechanisms in Radiobiology,*** 2 vol. (1960–61); INTERNATIONAL ATOMIC ENERGY AGENCY, ***Biological and Environmental Effects of Low-Level Radiation,*** 2 vol. (1976); J.M. YUHAS, R.W. TENNANT and J.D. REGAN (eds.), ***Biology of Radiation Carcinogenesis*** (1976); W.H. BLAHD (ed.), ***Nuclear Medicine,*** 2nd ed. (1971); L.G. TALIAFERRO and B.N. JAROSLOW, ***Radiation and Immune Mechanisms*** (1964); D.S. GROSCH, ***Biological Effects of Radiations,*** 2nd ed. (1979); D.J. KIMELDORF and E.L. HUNT, ***Ionizing Radiation: Neural Function and Behavior*** (1965); A.C. UPTON, ***Radiation Injury*** (1969); and M.M. ELKIND and G.F. WHITMORE, ***The Radiobiology of Cultured Mammalian Cells*** (1967). Radiation genetics is discussed in C. STERN, ***Principles of Human Genetics,*** 3rd ed. (1973); B. WALLACE and T.G. DOBZHANSKY, ***Radiation, Genes, and Man*** (1959), as well as in WORLD HEALTH ORGANIZATION, ***Effect of Radiation on Human Heredity*** (1957).

Radiation effects on human beings are discussed in J.W. GOFMAN, ***Radiation and Human Health*** (1981); H. BLATZ, ***Introduction to Radiological Health*** (1964); and E.P. CRONKITE and V.P. BOND, ***Radiation Injury in Man*** (1960). Examples of works calling attention to the hazards of radiation include the following: A.R. TAMPLIN and J.W. GOFMAN, ***Population Control Through Nuclear Pollution*** (1970); R. CURTIS and E. HOGAN, ***Perils of the Peaceful Atom: The Myth of Safe Nuclear Power Plants*** (1969). The problems of atomic power for society are discussed in HARRY FOREMAN (ed.), ***Nuclear Power and the Public*** (1970); and in G.T. SEABORG and W.R. CORLISS, ***Man and Atom; Building a New World Through Nuclear Technology*** (1971).

Radiation accidents are discussed in L.H. LANZL, J.H. PINGEL, and R.H. RUST (eds.), ***Radiation Accidents and Emergencies in Medicine, Research, and Industry*** (1965); and in U.S. HOUSE OF REPRESENTATIVE$ COMMITTEE ON SCIENCE AND TECHNOLOGY, ***Three Mile Island Nuclear Plant Accident*** (1979). Fallout and its biological consequences are covered by F.C. MCLEAN and A.M. BUDY in ***Radiarion. Isotopes, and Bone*** (1964); E.B. FOWLER (ed.), ***Radioactive Fallout, Soils, Plants, Foods, Man*** (1965); and INTERNATIONAL ATOMIC ENERGY AGENCY, ***Transuranium Nuclides in the Environment*** (1976). The effects of the Hiroshima and Nagasaki explosions on humans are discussed in a series of reports of the Atomic Bomb Casualty Commission. The United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) issues occasional reports on the biological effects of radiation and on fallout. The Office of Radiation Programs, a branch of the U.S. Environmental Protection Agency (EPA), issues studies of the effects of radiation on the environment. Standards and permissible levels of radiation exposure are available through reports of the National Council on Radiation Protection and Measurements (NCRP), and the International Commission on Radiological Protection (ICRP). The U.S. Nuclear Regulatory Commission issues many technical reports and educational materials, including films. IAEA also has a series of publications. Radiation problems in space flight are discussed in NATIONAL RESEARCH COUNCIL, SPACE RADIATION STUDY PANEL, ***Radiobiological Factors in Manned Space Flight*** (1967).

Ultraviolet and associated radiations are covered in A. HOLLAENDER (ed.), ***Radiation Biology,*** vol. 2, ***Ultraviolet and Related Radiations,*** vol. 3, ***Visible and Near-Visible Light*** (1956); and W. HARM, ***Biological Effects of Ultraviolet Radiation*** (1980). An international journal, ***Photochemistry and Photobiology*** (monthly), is devoted to this field.

Laser research is discussed in the following publications: S. FINE and E. KLEIN, "Biological Effects of Laser Radiations,"

Advances Biol. Med. Phys., 10:149–226 (1965); and S.J. WYARD (ed.), ***Solid State Biophysics: Applications of Electron Spin Resonance, Dielectric Measurements, the Mössbauer Effect, and Lasers to Biology and Medicine*** (1969).

There are additional specialized journals available in the field, including ***Radiation Research*** (monthly); ***International Journal of Radiation Biology*** (monthly); ***Radiobiologiya*** (bimonthly); ***Journal of Nuclear Medicine*** (monthly); ***Environmental and Experimental Botany*** (quarterly); and ***Nuclear News*** (monthly).

(C.A.T.)

Radiation Detection and Characterization

Radiation detection is concerned with a great variety of atomic and subatomic particles moving at different speeds and with electromagnetic waves of different wavelengths moving always at the speed of light. Both waves and particles are constantly present on Earth and in space. They emanate from the Sun, other stars, naturally occurring radioactive elements, synthetically produced radioactive isotopes, nuclear reactors, nuclear fission, and in general from interactions between colliding particles, between waves and particles, and between bulk matter and particles or waves.

The particles considered in this context are long-lived electrons, protons, alpha particles (*i.e.*, helium nuclei), neutrons, several hundred short-lived subatomic particles, and all antiparticles. The whole spectrum of electromagnetic waves is also considered in this context, from long radio waves through light waves and X-rays, to the shortest of gamma rays.

For details of the phenomena treated here in general terms relevant to radiation detection, see RADIOACTIVITY; X-RAYS; METALS, THEORY OF; SEMICONDUCTORS AND INSULATORS, THEORY OF; SOLID STATE OF MATTER; RADIATION EFFECTS ON MATTER.

The fundamental mechanism underlying the operation of all radiation detectors is the dissipation of energy by a charged particle in a suitable medium and the distribution of this energy among atoms and molecules of the detecting material. Radiations such as electromagnetic waves and neutral particles that react only weakly with the electronic structure of many detection media must first interact with some suitable material in a way that will produce a charged particle, which can then dissipate its energy in the detection media. The essential concept of detection lies in the fact that energy is provided by the radiation, and in this article it is therefore appropriate first of all to consider general aspects of matter and energy in relation to the charged particles that have already been mentioned and to electromagnetic waves and neutrons; following this, the characteristics of the various types of radiation are described and, finally, the detectors themselves are briefly reviewed.

THE STRUCTURE OF MATTER AND INTERACTIONS WITH RADIATION

A normal atom consists of a central nucleus, composed of electrically neutral neutrons and positively charged protons, around which orbit several negatively charged electrons. The number of electrons equals the number of protons in the nucleus, the negative charge on each electron just balancing the positive charge on each proton, so that the atom is electrically neutral; this number is called the atomic number of the atom, Z, and determines its chemical identity. Electrons around the nucleus exist in stable groups of orbits, the innermost, called the K shell, containing two electrons, the next, the L shell, containing eight electrons when it is complete, and each larger shell, M, N, O, P, etc., being complete with larger numbers of electrons. Each orbit corresponds to a definite state of energy, and movement of an electron from one orbit to another can take place only by the emission or absorption of a fixed amount of energy, an amount that depends on the initial and final orbits. It takes energy to move an electron away from the nucleus. When an electron is removed entirely from the atom, excess energy can be absorbed by it in such a way that this excess is manifested as kinetic energy (motion) of the free electron. A convenient unit of energy for describing atomic phenomena is the electron volt (eV), which is equal to 1.6×10^{-12} erg. The minimum energy that is needed to remove an outer electron from its orbit is called the ionization potential (E_i) and is about 15 electron volts for atoms in the gaseous state. To eject one of the innermost electrons from the innermost or K shell, a minimum energy, E_K, of much greater value is needed, and E_K is approximately equal to 10 times the square of the quantity, atomic number minus one ($E_K = 10\,[Z - 1]^2$eV). An atom from which an electron has been removed is left with a positive charge and is called a positive ion.

Energy of ionization

The energy required to ionize an atom can be provided either by collision with a charged particle, such as an electron or a proton, or by absorption of electromagnetic radiation in the form of light, X-rays, or gamma rays. In the latter type of interaction, electromagnetic energy behaves as though it existed in packets, or quanta, of energy proportional to the frequency of the radiation. Such quanta are called photons. Blue light, for instance, has a quantum energy of about three electron volts, whereas photons of deep-red light of twice the wavelength and half the frequency interact as quanta of 1.5 electron volts. An X-ray photon capable of providing just enough energy to remove one of the K electrons from an atom of argon, ($Z = 18$), has an energy of 3,000 electron volts. A photon of higher energy than E_K removes one of the K electrons and provides an excess kinetic energy to the electron, but it is rather less likely to cause ionization by photoelectric absorption than radiation having a quantum energy just equalling E_K; the probability of interaction by this process decreases as the photon energy increases. Usually, radiations of the highest frequencies and photon energies are the gamma (γ) rays. The probability of photoionization as a function of photon energy is expressed as the effective cross-section area presented to the radiation by each atom. The unit of cross section is the barn, equal to 10^{-24} square centimetre.

When a photon of high energy strikes an atom, it can remove an electron by a scattering process called Compton scattering, the energy of the primary photon being shared between the released electron and a scattered photon. The cross section for Compton scattering is proportional to the number of electrons in the atom and inversely proportional to the square root of the photon energy.

A fundamental mode of interaction for high-energy photons is by pair production, in which a photon of energy slightly greater than 1,000,000 electron volts (MeV) materializes into an electron and its antiparticle, a positive electron, or positron, in the proximity of an atomic nucleus. This interaction is possible because of the equivalence of matter and energy first enunciated by Albert Einstein in his theory of relativity (*i.e.*, energy equals mass times speed of light squared: $E = mc^2$). An electron (or a positron) of mass m_0 has a self-energy m_0c^2 equal to just over 500,000 electron volts (in fact, 0.51 MeV), and a photon of energy 1.02 MeV can thus provide the energy for two electron masses, one negative and one positive, because the net charge must remain zero. The cross section—*i.e.*, the probability—for pair production increases as the square of the atomic number of the target atom and roughly as the square of the excess energy (photon energy less 1.02 MeV).

Ionization caused by a charged particle colliding with an atom is a much less well-defined process than those described above, and electrons may be ejected with a range of energies depending on the mass and energy of the incident-charged particle. The average energy abstracted from the particle is about twice the ionization potential of the atom ionized, and the symbol w is used for this parameter (a constant that is characteristic for different conditions); w, the average energy per ionizing event, depends to a slight extent on the mass and energy of the incident particle. The probability of ionization at collision is much larger for slow-moving particles; therefore, as an energetic charged particle traverses a given thickness of material, it is slowed down more and more rapidly as energy is dissipated in the ionizing process, until all the energy is lost and the particle is absorbed. Massive particles, such as protons, which have a mass about 1,800 times that of an electron, are not greatly deflected by the electron structure of the atoms through which they pass and a fairly definite range for them can be assigned, depending on their initial energy. Electrons, on the other hand, are easily deflected; a beam of electrons initially having the same energy (*i.e.*, being monoenergetic) undergoes exponential absorption as it passes through matter, owing to electrons being scattered out of the beam. A maximum range, however, can still be measured.

Neutron reactions

Neutrons, having no electric charge, cannot interact with the electron structure of atoms but can penetrate the nucleus and cause a reaction. In some cases, the disturbed nucleus splits apart, as in the case of the uranium isotope of mass 235 (^{235}U), which breaks up into two fission fragments carrying kinetic energy of 150 MeV. In other cases, a charged particle, either a proton or an alpha particle (a; a helium nucleus) is ejected; for example, a boron-10 (^{10}B) nucleus ejects an alpha particle, and a lithium-7 nucleus is left, after reaction with a neutron. The reaction is written: $^{10}\text{B} + n \rightarrow \text{Li} + \text{a}$ (2.5 MeV); or $^{10}\text{B}(n, \text{a}) + {}^{7}\text{Li}$. Sometimes a neutron is absorbed by a nucleus. When a stable nucleus is formed by absorption of a neutron, the excess energy is released as a gamma ray. Some nuclei that result, however, are unstable and decay by what is called radioactive emission of electrons (usually referred to as beta [β] particles) and gamma rays. The radiation resulting from such neutron capture and reaction appears many seconds, days, or years later. In the other examples mentioned, the radiation is prompt; *i.e.*, it is emitted in a time less than a nanosecond (10^{-9} second) after the reaction.

In one class of interaction, an elastic collision takes place, and the energy of the neutron is shared between a scattered neutron and a projected nucleus. Scattering of fast neutrons on hydrogen nuclei gives rise to protons.

DETECTION PROCESSES AND DETECTION MEDIUMS

The distribution of energy among the atoms and molecules of a suitable detection medium is the fundamental process in a radiation detector.

A charged particle (electron, proton, alpha particle, or fission fragment) of energy E that dissipates all its energy in a gas, by knocking electrons out of atoms, leaves a column of positive ions and free electrons along its track. The number of such electron–ion pairs produced is equal to the energy E, divided by the value of w, the energy per ion pair appropriate to the gas and the particle. If electrodes are set up on either side of the track, to produce an electric field, free electrons are attracted to the positive electrode, or anode, and positive ions are attracted to the negative electrode, or cathode. If all of the separated ions and electrons are collected, the electric charge available for measurement is equal to E/w multiplied by e, the electron charge, of value 1.6 X 10^{-19} coulomb (unit of charge). For an energy of 1,000,000 electron volts and energy per ion pair of 30 electron volts, the charge collected would be 5×10^{-15} coulomb. Because the minimum value of a pulse of electric charge that can be measured by available electronic apparatus is about 1,000 electron charges (1.6×10^{-16} coulomb), an ionization chamber can be used for detecting the passage of a charged particle that dissipates a minimum energy of about 30,000 electron volts in the detection medium; *i.e.*, the gas in the chamber.

One source of charged particles

The charged particle providing energy to ionize the gas may be produced by decay of a radioactive material emitting an alpha or beta particle; and the radioactive source may be placed outside the detector, so that the radiation must pass through a suitably thin portion of the chamber wall (a window) to reach the gas. If the absorption of particle energy in a window would be inconveniently large, the source is mounted inside the detector, directly in contact with the gaseous detection medium. In some circumstances, the radioactive material is available in the form of a gas—*e.g.*, carbon dioxide with carbon-14 (^{14}CO$_2$)—which can be mixed with the ionization-chamber gas or can itself be used as the detection medium.

When the charged particles are produced as a result of gamma ray or neutron interactions, similar considerations apply; the interacting material may be in the form of a radiator outside the detector, inside the detector but separate from the detection medium, or distributed throughout the detection medium. For example, a neutron detector might consist of an ionization chamber containing a layer of boron-10, or it might be made by filling an ion chamber with boron trifluoride containing boron-10 (^{10}BF$_3$), a gas at room temperature. In either case, neutron interactions would be detected by ionization produced in the gas by the $^{10}\text{B}(n, \text{a})$ reaction.

Ionization detection. A detection system meant to produce electrical signals when a particular type of radiation is present thus consists of a source of radiation, an interacting medium (for neutrons and gamma rays), and a detecting medium together with some means of converting the energy dissipated in this material into the required signal. Such an active detector may make use of ionization and charge collection in the detection process; the ionization medium may be a gas or may be a suitable semiconducting solid. The class of detectors comprises ionization chambers and conduction counters.

Geiger-Muller counter tubes. In some gas-filled devices, a sufficiently high electric field is applied to cause the free electrons produced by the primary particle to move sufficiently fast to ionize more atoms, thereby increasing the original charge by the process of gas multiplication. Proportional counters use this process to give internal amplification factors of up to 1,000,000 times while still giving an output proportional to the initial signal. At higher values of gas multiplication, complex processes come into play that give a large output signal from the smallest levels of initial ionization, and the output is then independent of the primary signal. Such tubes are called Geiger-Miiller counter tubes.

Spark chamber. A gas detector using high values of gas multiplication is the spark counter, in which a spark jumps between two electrodes along the track of an ionizing particle. Because the spark indicates the position of the particle, assemblies of spark counters are arranged inside spark chambers to indicate the paths of penetrating particles by observation of the sparks produced.

Cloud *and* bubble chambers. Rather less active track delineating devices depending directly on the production of ions along the path of the particle are cloud chambers and bubble chambers. In the former, vapour from a supersaturated atmosphere of water or alcohol condenses on ions to form visible droplets, which can be photographed. In the bubble chamber, a superheated liquid is used, and ions form nuclei for local boiling, again giving a visible track.

The detailed physics of the various types of ionization detector summarized above depends to a large degree on the behaviour of electrons and ions under the applied electric field.

Fluorescent detectors. A second main class of radiation detectors depends on the conversion of the distributed energy of the primary ionizing particle into light, which is then used to generate photoelectrons (electrons emitted from a surface as a result of light falling on that surface) from the cathode of a photomultiplier tube. The photoelectron current is multiplied in the tube to give an adequately large output signal. Detection mediums that convert the energy of moving electrons into light are known as phosphors. The light of a television picture tube is produced by the impact of energetic electrons on a phosphor deposited on the inner surface of the faceplate of the picture tube. The process of light production is known as fluorescence. Early examples of its use in radiation detection, albeit with visual rather than instrumental observation of the light, are the detection of X-rays by the German physicist Wilhelm Conrad Rontgen in 1895 using barium-platinocyanide, and the detection of alpha particles by observation of the scintillations produced in zinc sulfide by the British physicist Ernest Rutherford during the early 1900s.

Fluorescence

As well as the rather transient effects so far considered, the passage of an ionizing particle through certain types of material can give rise to more or less permanent changes along the track. A familiar example is the photographic film. Perhaps less well known is the blackening of glass under intense irradiation from beta particles or gamma rays. The degree of darkening either of film or glass can be used to determine the total exposure of the sample to radiation, and badges containing sensitive substances are sometimes worn to indicate the amount of invisible radiation to which their wearers are exposed.

The properties of ionization **mediums.** Gases. Once a gas has been ionized. the free electrons must be separated from the positive ions by an electric field if a useful charge signal is to be produced. In the absence of a field, the attraction between positive and negative charges, combined with the random motion of the gas molecules, causes the oppositely charged particles to recombine. At room temperature, molecules of a gas are in constant random motion with an average kinetic energy of 0.04 electron volt. This energy is proportional to the absolute temperature (°K) of the gas, which is equal to the temperature in °C plus 273. For a given mass of gas in a fixed volume (*i.e.*, a given gas density), the pressure exerted by the gas on the walls of the vessel is proportional to the absolute temperature and to the density. The average distance travelled by a gas molecule before it collides with another molecule is called the mean free path and is inversely proportional to the number of molecules per unit volume.

In a gas at the standard temperature of 0° *C* and the standard pressure of one atmosphere, one cubic centimetre contains 2.7×10^{19} molecules, and the mean free path (symbolized by the Greek letter lambda, A) is about 2×10^{-5} centimetre.

Drift velocity

Between collisions with gas molecules, a positive ion obtains energy from an electric field produced between electrodes in an ion chamber and thus is accelerated toward the negative electrode, or cathode. This energy is small compared with the average kinetic energy that the ion possesses in common with the surrounding gas molecules, and because the ion and the molecules are similar in mass, velocity (the downfield velocity component) is lost at each collision. The positive ion thus drifts toward the cathode in a series of steps that have an average value proportional to the field strength multiplied by the square of the free time between collisions. The drift velocity is proportional to this average step distance divided by the free time. Since the free time is inversely proportional to the mean free path, the drift velocity, $W+$,is proportional to field strength, E, multiplied by the mean free path, λ. At a constant temperature, mean free path is inversely proportional to the gas pressure, *p,* and the ion drift velocity can be written $W^+ = \mu E/p$, in which the constant symbolized by the Greek letter mu, μ, is called the mobility of the ion for the particular gas. With the electric field, E, expressed in volts per centimetre and the velocity, $W+$, in centimetres per second, the mobility, μ, is given in cm/sec per V/cm, otherwise written as centimetres squared per volt-second.

The electric field between plane parallel electrodes is equal to their potential difference (P.D.) divided by their spacing. With a spacing of x cm and a P.D. of v volt, E $= v/x$ volts per centimetre, and for a gas pressure of p atmospheres, the ion drift velocity (see above) is $W^+ = \mu(E/p)$, which, substituting for E, gives $W^+ = \mu v/px$ centimetres per second; the time *t* for the ion to travel a distance *s* centimetres is t $= s/W^+ = spx/\mu v$ second (*i.e.*, the distance multiplied by spacing and pressure, divided by field, multiplied by mobility). For an ion formed close to the anode, so that $s = x$, the transit time is proportional to the square of the electrode spacing. With argon at one atmosphere, an applied voltage of 500 volts and a spacing of five centimetres, E is 100 V/cm; W^+, for the argon ion Ar_2^+, is 190 cm/sec; and the transit time is 26 milliseconds. With a gas filling pressure of one-tenth atmosphere, the velocity would be 10 times higher and the transit time 10 times smaller.

A free electron, having a mass only 1/70,000 of that of an argon molecule, cannot transfer much kinetic energy to a gas molecule in an elastic collision, being somewhat analogous to a table-tennis ball bouncing off a cannon ball. The energy acquired by the electron from the electric field thus accumulates, and the agitation energy becomes much larger than that of the thermal energy (heat) of the gas. The electron changes direction at each collision with a molecule, so that the motion remains random; with a superimposed drift velocity, as with the ion, and, by a similar argument, the drift velocity, W–, is found to be proportional to field multiplied by mean free path. The mean free path of an electron in a gas, λ, is affected by the value of E, because this determines the agitation energy, and a simple value of mobility is not applicable. In fact, it can be shown that the drift velocity is proportional to the field E multiplied by electron mean free path λ divided by the square root of the parameter symbolized by the Greet letter eta, η, which is the ratio of average agitation energy of electron and gas molecule: $E\lambda/\eta^{1/2}$. A free electron thus has an agitation energy of 0.04η electron volt in a gas at room temperature, and, under sufficiently high values of electric field, this energy can increase to a value at which inelastic collisions with molecules occur, when the electron transfers energy to the structure of the molecule. In a monatomic gas such as argon, the lowest level at which energy can be transferred from an electron is 11.57 electron volts, so η can rise to about 300. In a more complex gas, such as carbon dioxide, energy can be exchanged at low values with the vibrational and rotational energy systems of the polyatomic molecule, keeping the electron agitation energy down to a low value for fields below 2,000 volts per centimetre.

Mean free path in argon

The mean free path of free electrons in argon decreases markedly as the agitation energy increases. As a result, the drift velocity of electrons in pure argon at a pressure of one atmosphere does not exceed 10^6 centimetres per second. The admixture of a small proportion of a complex gas to argon reduces the agitation energy of electrons, however, and allows much higher values of drift velocity to be obtained. In an ion chamber containing argon plus 1 percent nitrogen, at one atmosphere, an electron has a drift velocity of 0.5×10^6 centimetres per second in a field of 100 volts per second and crosses a distance of five centimetres in 10 microseconds (10^{-5} second). A knowledge of the value of drift velocity and the way it depends on electric field strength is necessary to understand the mechanism of ionization detectors.

The effectof small admixtures of other gases to a noble gas such as argon has been described in relation to electron drift velocity. A similar profound effect is produced on the ionizing efficiency of fast charged particles.

The value of w, the energy per ion pair abstracted from, say, an alpha particle in pure neon, is 36.3 electron volts. (The ionization potential of neon is 21.5 electron volts.) If, however, about 0.1 percent of argon is added to the neon, w is reduced to 26.3 electron volts, which is the energy per ion pair for pure argon. The improvement in energy transfer is caused by the fact that much of the energy abstracted from the primary particle in pure neon is used in exciting a level in the neon electron system that requires 16.6 electron volts but that does not result in ejection of an electron. This excited state in neon lasts for a long time (on the time scale of atomic phenomena) and has an average lifetime of about a millisecond, after which it decays by emission of an ultraviolet quantum. Such long-lived excited conditions are called metastable states. The mean free time between collisions is about 10^{-7} second, so that a neon atom excited into its metastable state will undergo many collisions in its life, and in the gas mixture containing argon it will certainly collide with an argon atom, having an ionization potential of 15.7 electron volts. Because the excitation energy, E_r, of the neon atom is 16.6 electron volts, there is a high probability that the energy can be transferred to the argon atom, causing ionization. Such collisions of the second kind, as this type of energy transfer is called, greatly increase the efficiency of production of ionization in a mixture of a noble gas with another gas having an ionization potential less than E, the energy of excitation of the metastable state of the noble gas.

The same principle that allows energy transfer from a metastable state of one molecule to ionize a second molecule allows a charge transfer from a positive ion to a molecule of lower ionization potential. Thus, in an argon–alcohol mixture, collision between an argon ion and a molecule of ethyl alcohol results in neutralization of the argon ion by extraction of an electron from the alcohol molecule. This process is of particular importance in understanding the operation of a Geiger counter tube.

At high values of electric field, electrons obtain high average energy, the instantaneous value being distributed statistically about the average. Occasionally an electron attains an energy larger than the ionization potential of the gas molecule with which it collides, and ionization occurs. The actual number of ions produced by an electron in moving one centimetre through a gas under the influence of a field E (in units of volts per centimetre per atmosphere [V/cm/atm] is given by a coefficient, called the first Townsend coefficient and symbolized by the Greek letter alpha, ***a.*** Values are given in the Table, as is the effect of adding argon to neon.

Townsend Coefficient for Various Gases

E	*a,* ion pairs per centimetre			
(V/cm/atm)	argon	neon	neon + 0.1% argon	air
3,800	0.2	6	120	—
7,600	3.8	38	230	—
19,000	40	200	380	—
38,000	530	530	830	38

Electron avalanche

In such high electric fields, electrons produced by an ionizing collision are themselves accelerated and cause further ionization, and the tertiary electrons thus produced go on to produce further ionizing collisions, an electron avalanche rapidly building up. The process can be visualized by considering two parallel electrodes in a chamber filled with argon at a pressure of one atmosphere and supporting a voltage difference of just over 10,000 volts. At this field, the first Townsend coefficient in argon produces about 10 electron–ion pairs per centimetre. An electron produced at the cathode—for instance, by an incident ultraviolet photon—is accelerated toward the anode; after passing through one millimetre of gas, it is probable that one ionizing collision has occurred, and so there are now two electrons; after two millimetres each electron has ionized again, and there are four electrons. This process continues, and, at the nth millimetre, there are 2^n electrons; when the avalanche reaches the anode it contains 2^{10} electrons ($2^{10} =$ 1,024). This analysis is only illustrative. An accurate analysis shows that an electron formed at the cathode under the given conditions produces 150 times more ion pairs on its way to the anode than does an electron formed halfway between the plates. Some possible arrangements, therefore, are not all suitable for use in a gas multiplication detector, because the amount of multiplication varies markedly with the position of the initial ionization in the chamber. A cylindrical geometry. consisting of an anode wire stretched along the axis of a cylindrical cathode, provides an electric field that has a high value close to the anode surface and falls away rapidly as distance from the anode increases; this arrangement allows of controllable gas multiplication in the proportional counter.

At higher values of electric field than the 10,000 volts per centimetre per atmosphere applied across a parallel plate system in an argon filled chamber that gives a coefficient α equal to 10, the initial part of the electron avalanche is so intense that ultraviolet photons of high enough energy to ionize the gas are produced, generating electrons at a little distance from the primary region. A second bead of intense ionization is produced, generating photons in its turn, and the discharge is propagated across the gap between the electrodes to form a spark, in a time of about 10^{-8} second, short compared with the transit time of an electron, which is about one microsecond (10^{-6} second). Providing the high voltage supply to the electrodes is capable of supplying the required current without reduction in voltage, the spark becomes a sustained discharge, fed by electrons generated from the cathode by photoemission (from the ultraviolet photons produced in the spark) and by electrons produced at the cathode by the arrival of positive ions. An ion of argon having a potential energy of 15.7 electron volts, extracts an electron from the metal of the cathode against the surface force (symbolized by the Greek letter phi, ϕ) of phi electron volts and becomes a neutral atom with an excitation energy of (15.7 $-$ ϕ) electron volts. All this happens at a distance of about 5 $\times$ 10^{-8} centimetre from the surface, and in a time of 10^{-12} second the excited atom strikes the metal. If the excitation energy is greater than the surface force, ϕ, which is called the work function, then there is some probability that a further electron will be extracted and move into the discharge. Typical values of the work function lie around four electron volts, and, because $\frac{1}{2}E_i$ (E_i is the ionization potential) for many gases is greater than five electron volts, this process is quite effective in generating electrons.

Certain complex gas molecules, such as ethyl alcohol, dissipate the excitation energy left after neutralization among the vibrational states; in this case, no electron is extracted.

If the high voltage supply to the electrodes is not capable of supplying a continuous current to the discharge, the spark persists until the charge drawn from the capacitance here, the capacitance is formed by the electrodes and the connections to the system) causes the field to drop below the level at which ionization by collision can occur: the spark is then extinguished.

The processes described above are of importance in Geiger counters and spark chambers.

Solids. In an ionization chamber using gas as the detection medium, no current passes between the electrodes in the absence of ionization in the gas; the neutral gas is an insulator. It is equally difficult, however, to make a gas detector with adequate stopping power to absorb the energy from a penetrating particle, such as a high-energy electron, or with a high enough atomic number and atomic density to provide a high detection efficiency for gamma rays, and it would be desirable to use a solid detection medium. In such solid material, atoms are relatively immobile, and their electronic structures interact with each other to such an extent that it is not possible to assign specific outer orbital electrons to a particular atom. The concept of ionization is thus not as clear-cut as it is with a gas, in which the atoms or molecules are essentially independent; furthermore, in metals and semiconductors at room temperature, some electrons are always free to move under the influence of an electric field, and so these materials show electrical conductivity. An insulating solid has the outer electrons sufficiently firmly bound to the array of atoms so that they are not free to move. Absorption of energy from a photon or a charged particle may break the bonds holding an electron in place, and it is then free to move in a field. The vacancy thus created in the electron structure of the atomic array leaves a net positive charge in the locality, which charge is referred to as a positive hole. The imbalance in forces produced by an electric field causes a neighbouring valence electron to move toward the anode, filling up the original vacancy; the hole has now migrated one step toward the cathode. Conduction in an insulator can thus take place by migration of holes and electrons. In some materials, however, the holes have low or zero mobility.

Ionization process in solids

The minimum energy needed to free an electron from the valence band of energy to the conduction band of energy, so that it can move in an electric field, is called the band gap, which is about five electron volts in diamond and in silver chloride. In semiconductors, the band gap is lower; in silicon, it is 1.1 electron volts, and in germanium it is 0.72 electron volt. In these materials, which are used for making transistors, the conductivity is lower as the material is purer (pure silicon has a resistivity in excess of 100 megohms at room temperature, between

The band gap

opposite faces of a cube with sides of one centimetre). This conductivity results from the presence of electrons raised into the conduction band by the statistical fluctuation of thermal energy and corresponds to an electron density of 10^{10} per cubic centimetre and a lifetime of 100 microseconds before the electron is trapped. Each electron produced under these circumstances generates a hole, so that the hole density in silicon of a high purity is also 10^{10} per cubic centimetre at room temperature. High-purity material in which all charge carriers (both electrons and holes) are thermally generated is said to have intrinsic conductivity. The rate of generation is 10^{14} per second per cubic centimetre at room temperature.

The movement of a free electron in a solid under the influence of an electric field is similar to the description given for the movement of an ion in a gas; energy is obtained from the field and is lost when the electron is scattered by the pattern of forces formed by the interlocking electron bonds that hold the solid together. The drift velocity thus turns out to be proportional to the product of field and mean free path, as in the case of the gas ion, and a mobility, μ, can be assigned. For silicon at room temperature μ_- is 1,350 centimetre squared per volt-second, whereas μ_+, hole mobility, is 500 centimetres squared per volt-second.

The range of an electron of initial energy of two million electron volts in solid silicon is 0.5 centimetre, compared with 500 centimetres in argon gas at atmospheric pressure, and the energy per electron-hole pair, w, is 3.5 electron volts. A conduction counter made from a one-centimetre cube of intrinsic conductivity silicon could thus be used to absorb the energy from high-energy particles. An energy of 250,000 electron volts absorbed generates about 75,000 electron-hole pairs, giving a collected charge of 1.2×10^{-14} coulomb, which is well above the limit of detection set by amplifier noise. With a potential difference of 100 volts between opposite faces of the cube, this charge is collected in a time of about 10 microseconds.

During this time, however, 10^9 charge carriers generated by thermal energy have been collected, and this number is not constant for each 10-microsecond period but is subject to statistical variation about this average value. For 5 percent of the time the value will deviate from the mean by more than twice the square root of the average number; in this case by more than 6 × 104. If the conduction detector is imagined as a device sampling the charge collected in successive periods of time, each of which is equal to the time taken for charge carriers to cross the block from anode to cathode (10 microseconds in the example considered), then the statistical fluctuation described would make it difficult to identify with any certainty the extra signal caused by the absorption of 0.25 million electron volts from a charged particle in the sensitive volume of the silicon block. In order to reduce this conductivity noise, it is necessary to cool the material because the density of thermally generated carriers is proportional to the thermal agitation energy of the electron in the solid. At the temperature of liquid nitrogen, the density of free carriers is reduced from 10^{10} per cubic centimetre of silicon to effectively zero, and the cooled silicon should then be suitable for detecting electrons.

Need for cooling a medium

Unfortunately, it is difficult to make silicon of intrinsic purity, and the impurities that are commonly present decrease the resistivity. Silicon is tetravalent, which means that the outer shell of the atom contains four electrons. In the lattice of the crystalline form in which silicon is normally prepared, each atom shares its valence electrons with four neighbouring atoms, so that each electron pairs with an electron from the adjacent atom. An atom of phosphorus is pentavalent (*i.e.*, has five outer shell electrons); when it is incorporated in the silicon lattice as an impurity, it forms electron pair (or covalent) bonds with the four adjacent silicon atoms. This bonding leaves a spare electron that has a binding energy of only 0.054 electron volt and so is readily freed from its parent atom by thermal energy at room temperature and is free to take part in conduction. Trivalent impurities, such as boron, bind three adjacent atoms with covalent bonds, leaving a positive hole that can drift away under thermal excitation and having an effective binding energy of 0.08 electron volt.

Silicon of the highest purity contains about 10^{13} atoms per cubic centimetre of excess impurity of either type, giving rise to resistivity of about 10^4 ohms per centimetre cube at room temperature, which increases to about 10^6 ohms per centimetre cube if the temperature is lowered to 77° K. Material with excess of pentavalent (donor) impurity conducts by electrons and is called an n-type semiconductor. With trivalent (acceptor) impurity and conduction by hole migration, the material is called a p-type semiconductor.

Various methods have been developed to manufacture materials with low electronic conductivity, with p- and n-type layers, and with p–n junctions. Although silicon and germanium are by far the most commonly used materials for solid-state ion chambers, the classical work in this area was carried out on single crystals of silver chloride cooled to 77° K.

Scintillation counter. The distributed energy of a charged particle may be used to cause the emission of light from a suitable medium; the absorption of this light by a cathode, called a photocathode, produces electrons (photoelectrons) that escape from the surface of the cathode into the evacuated interior of an electron tube called a photomultiplier. The photoelectrons are accelerated by an electric field and strike an electrode coated with secondary emitting material. Each electron striking this surface with an energy of 1–500 electron volts knocks out secondary electrons that can be accelerated in turn onto a second electrode, or dynode, as the coated electrodes are called, and so on through a series of dynodes that are at successively more positive potential.

Photo-multiplier

With a secondary emission coefficient of four electrons per primary electron of energy 100 electron volts, a 10-dynode multiplier gives a gain of four raised to the 10th power, (4^{10}), which is slightly more than 1,000,000. Each primary photoelectron produced by the absorption of a light photon then produces at the anode of the tube a pulse of charge containing 1,000,000 electrons and of magnitude 1.6×10^{-13} coulomb. Two primary electrons striking the first dynode simultaneously produce a pulse of 2,000,000 electron charges at the output, and so on. The photomultiplier is a linear amplifier—*i.e.*, giving an output proportional to the primary photoelectron signal. A light-emitting detection medium (a phosphor), coupled to the photocathode of a photomultiplier operated to give a suitable value of gain, forms a scintillation counter; and in this type of detector the useful primary signal, comparable to the electrons and ions formed along the track of a particle in a gas, is the emission of photoelectrons. A high yield of photoelectrons for a given energy dissipated in the phosphor requires a photocathode of high quantum efficiency (efficiency measured by photoelectrons per incident photon), coupled with minimum light loss to a transparent phosphor that converts ion excitation energy into photons, with high efficiency. The wavelength of the fluorescent light must match the photocathode efficiency curve. In practice it is found that the most useful photocathodes give high quantum efficiency in the blue region of the spectrum and that the most useful phosphors emit blue fluorescence.

Photocathodes. Cathodes useful for scintillation counters are formed by semiconducting compounds of antimony and the alkali metals cesium, sodium, and potassium. Most commonly used are cesium–antimony (Cs_3Sb) and cesium–potassium antimony (K_2CsSb).

Secondury *emitters.* The two cathodes described above are good secondary emitters and are conveniently formed in photomultiplier tubes by reacting the alkali metals (cesium and potassium) with antimony-coated dynodes at the same time as the cathode is formed. Another useful material is beryllium oxide.

In a photomultiplier tube, dynodes are arranged so that an electric field accelerates secondary electrons from the surface and directs them to the next dynode in line. One method of doing this is to make a dynode in the form of parallel inclined slats with the upper plane covered by a

Types of dynodes

fine, transparent mesh. Electrons pass through the mesh of the first dynode to strike the inclined slats. The field from the mesh of the second dynode, spaced two millimetres below, extracts the secondaries from the space between the slats of the dynode of origin and accelerates them through the mesh to strike the second dynode slats, and so on. The efficiency of extraction of secondary electrons in this "venetian-blind" geometry lies between 90 and 95 percent.

Other types of dynode in common use are the box-and-grid and the linear-focussed structure. The latter consists of carefully shaped electrodes arranged in facing but offset pairs and designed to direct the secondaries from their point of origin to the next dynode without producing low-field regions, which are inherent in the venetian-blind and box-and-grid systems.

Photonzultiplier tubes. The great majority of scintillation counters use end-window photomultiplier tubes having semitransparent photocathodes of diameter ranging from two to 30 centimetres. All of these employ a dynode system fitting into a neck of five centimetres diameter, the cathode being spaced away from the first dynode by a distance of about half the end-window diameter. An electron optical system provides shaped electric fields to direct photoelectrons into the first dynode, and larger tubes in general require higher voltages between cathode and first dynode. With the gain of an individual dynode varying as the seven-tenths power of the inter-dynode voltage, a 10-dynode tube has a gain varying as the seventh power of the overall tube voltage, so that it is necessary to provide stable high voltage supplies to operate photomultiplier tubes at constant gain.

The secondary emission coefficient of dynode surfaces is slightly dependent on temperature, and most photomultiplier tubes decrease in gain by about 0.5 percent as the temperature (around 20° C, or 68° F) increases by one degree Celsius (1.8 degrees Fahrenheit). For accurate measurement, it is therefore necessary to hold the system at constant temperature or to provide a circuit that makes slight adjustments to the tube operating voltage to keep the gain constant. The transit time of an electron from the cathode to its arrival at the anode as a pulse of charge depends on the tube design and the operating voltage.

Because scintillation counters are often used to measure the strength of small sources of radioactive isotopes, it is necessary to construct the photomultiplier tube with the minimum amount of radioactive contamination. In practice this condition requires the use of glass containing the minimum amount of potassium-40, which is a beta- and gamma-emitting isotope, and sometimes calls for the tube envelope to be made, as far as possible, of fused silica.

Phosphors. In a scintillation counter, a large, transparent block of phosphor is coupled to the end window of a photomultiplier tube and surrounded by an efficient diffuser to ensure efficient transfer to the photocathode of the fluorescent light provided by particle excitation. With such an arrangement the energy, absorbed from a charged particle that results in the emission of a single photoelectron from the cathode of 24 percent quantum efficiency for three-electron–volts photons, can be measured. This figure is analogous to ***w***, the energy per ion pair, for a conduction medium, and the most efficient phosphors in use require about 200 electron volts to give a single photoelectron.

One useful class of phosphors is provided by fluorescent organic materials derived from benzene, typified by anthracene and naphthalene, that can be made as transparent crystals of useful size. As molecules excited by absorption of energy from a primary particle return to their ground state, light is emitted, and the decay time of this fluorescence is a few nanoseconds (10-9 second). If a small amount of anthracene is dissolved in naphthalene, however, the composite crystal emits light characteristic of anthracene fluorescence—but with the longer decay time of naphthalene. It appears that energy is transferred from molecule to molecule without radiation until radiation occurs, and this will take place from molecules having the highest emission probability.

This effect is exploited in liquid and plastic phosphors, which consist of efficient organic fluors, such as p-terphenyl dissolved in toluene or in polymerized polyvinyl–toluene.

Quenching

All of these materials show a marked loss of efficiency as the rate of energy loss from the primary particle increases, a loss that has been explained in terms of nonradiative dissipation of energy by damaged molecules. The reduction of fluorescent efficiency by competing routes for energy dissipation is called quenching; oxygen dissolved in the solvent of a liquid phosphor must be carefully removed to prevent quenching.

Inorganic phosphors are transparent insulating crystals containing a small proportion of a suitable impurity. Production of an electron-hole pair by an energetic ionizing particle occurs as described above. The hole is neutralized by extracting an electron from a nearby impurity centre, and this ionized centre in turn attracts the free electron from the conduction band to form a neutralized but excited impurity centre. The excess energy is emitted as a photon of energy which is less than the band-gap of the crystal and dependent on the properties of the impurity.

The phosphor in greatest use in scintillation counters is thallium activated sodium iodide; this can be grown into single crystals up to 75 centimetres in diameter and 25 centimetres thick. The material is extremely hygroscopic (moisture absorbing) and must be protected from the water vapour in the atmosphere by mounting it in a hermetically sealed can, with a window for coupling to a photomultiplier tube. In order to maximize the collection of fluorescent light, all the surface of the crystal except that facing the window is covered with a tightly packed thin layer of white diffusing titanium dioxide. This method of mounting is applied to many types of phosphor, whether hygroscopic or not.

Non-scintillating phosphors. Some phosphors are used for radiation detection because of their ability to store up energy in the crystal lattice over a period of exposure to nuclear radiation. Such materials are inefficient scintillators because the metastable state of the excited impurity centre is long-lived at room temperature and requires thermal excitation to temperatures perhaps 200"—300" *C* above room temperature to cause de-excitation. Calcium fluoride activated with manganese is one such phosphor. Energy stored in the lattice is estimated at some later time by rapidly heating the material, which starts to emit light at a temperature of about 200° *C* (392" F). The light rapidly rises in intensity to a peak at about 260" (500" F), then falls off as the temperature rises further and all the excited states are returned to ground level. The light output from the thermoluminescent phosphor is observed by a photomultiplier tube and plotted against the temperature of the sample. The total amount of light emitted is proportional to the energy absorbed.

The energy dissipated in a solid can be stored, after a fashion, by disrupting the bonds holding atoms to their neighbours, thus changing their potential energy. A special glass made from equal proportions of aluminum and lithium phosphate incorporating a few percent silver phosphate is nonfluorescent under ultraviolet light. Exposure to ionizing radiation breaks the silver atom bonds and produces silver impurity centres that turn the glass into a phosphor that fluoresces under ultraviolet light with an intensity depending on the energy dose absorbed. The weak fluorescence is orange in colour, and the intensity is measured in a photoelectric instrument.

Photographic emulsion

A mechanism similar to the above operates in photographic emulsion, in which the bonds of silver atoms to bromine atoms are disrupted by energy absorption. In this material, however, the evaluation technique involves development of the emulsion that causes each grain or small crystal of silver bromine (AgBr) to contain two or three free silver atoms reduced to atomic silver, thus becoming visible. Photographic films are widely used for measuring exposure to radiation, as are thermoluminescent and radiation-activated phosphors.

Ionization counters. The basic detection process, the collection of separated charge carriers produced by energy from a fast ionizing particle, and the properties of conduction mediums, have been described. In the case of a gas, it has been shown that the value of w, the energy per ion pair, is about 30 electron volts but depends on the gas and is modified by small admixtures of other gases. The movement of electrons in the collecting field also depends on the gas composition and pressure.

Detectors using variations of the same principles are pulse ion chambers, silicon and germanium detectors, Geiger counters, spark chambers, and cloud chambers.

Spark chambers, cloud chambers, and bubble chambers. The production of a spark, initiated by the presence of one or more primary electrons, has been described. In a spark chamber, an assembly of accurately parallel plates spaced perhaps one centimetre apart has plates 1, 3, 5, etc. connected together and to electrical ground. Plates 2, 4, 6, etc. are connected via a resistor to a relatively low voltage supply of 100 volts or so and also to a pulsed high-voltage unit, which applies a field of about 20,000 volts per centimetre across the plates in the opposite sense to the low-voltage field when triggered from a coincidence detector. Such chambers are made with dimensions up to three metres on a side and contain up to 80 plates. They are mounted between the poles of large magnets to measure particle momentum by forcing a particle to move in a circular path the radius of which depends on the magnetic field and the particle's mass, velocity, and charge. The gas used in a spark chamber is commonly air.

Spark location is usually obtained by photographing the chamber, but electronic methods of locating the spark within the chamber have also been developed.

The Wilson cloud chamber

A venerable precursor to the spark chamber for track delineation was the cloud chamber, developed by the Scottish physicist C.T.R. Wilson between 1896 and 1912. A chamber with a glass window opposite a movable floor is filled with a gas saturated with ethyl alcohol vapour. Expansion of the chamber volume by rapidly lowering the floor of the chamber, thus increasing its volume and thereby reducing the pressure suddenly, causes supersaturation of the alcohol vapour, which then condenses as droplets on any ions (or dust particles) present. With a dust-free filling and a clearing field between electrodes in the chamber. which is switched off when an observation is to be made; droplet-delineated particle tracks show up clearly against the black floor of the chamber and can be photographed. The reset time after expansion is five to 10 seconds, a time that compares unfavourably with the few milliseconds of the spark chamber.

The bubble chamber, developed by a U.S. nuclear physicist, Donald A. Glaser, in 1952, uses a superheated liquid in place of a vapour to obtain better stopping power in the track medium. Diethyl ether, compressed to a pressure of several atmospheres and heated well above the boiling point, remains superheated when the pressure is released, long enough for tiny bubbles to form around the ions in a particle track. Liquid hydrogen is used in neutron detection chambers. Being built up to two metres in diameter, bubble chambers, with their large magnet systems, are impressive engineering achievements. Their recycle time is about one second.

APPLICATIONS IN SCIENCE, INDUSTRY, AND TECHNOLOGY

Dosage measurement

The dissipation of energy by ionizing radiation in matter causes damage, either temporary, as in the ionization of a gas, or more permanent, as when chemical bonds are broken or atoms are displaced from their position in the lattice structure (see RADIATION EFFECTS ON MATTER). The unit used for the measurement of radiation dosage is called the rad. This unit corresponds to the dissipation of an energy of 100 ergs per gram of material at the point of interest. Devices for measuring dosage are ion chamber dosimeters; photographic dosimeters; thermoluminescent phosphor dosimeters; and radio-photoluminescent glass dosimeters. The use of a Geiger counter for uranium prospecting is part of general public knowledge. It is one example of an instrument used to detect gamma rays emitted by a radioactive material; it gives an electrical pulse each time a gamma ray interacts in the sensitive volume of the counter tube. Radioactive elements may also emit beta particles, which are high-velocity electrons, and alpha particles, which are positively charged nuclei of helium atoms, and specialized detectors are necessary for these radiations. The distinction between alpha and beta particles, which are components of ordinary matter, and gamma rays is that the latter consist of electromagnetic waves, essentially similar to X-rays and light waves but extending to higher energies and shorter wavelengths.

In nuclear reactors, uranium undergoes fission when bombarded with neutrons, and the two (or sometimes three) fragments of the uranium nucleus fly apart with great velocity, generating heat as they lose energy to the surrounding material. Neutrons are uncharged particles, about 1,840 times more massive than electrons, and are produced in fission and in a number of nuclear reactions. Detectors are needed for the detection and measurement of neutrons and sometimes for fission reactions.

Extraterrestrial radiation

The surface of the Earth is bombarded day and night with cosmic rays, possessing extremely high energy, even after passing through the full depth of the atmosphere. Outside this protective mantle there is a continuous bombardment of ultraviolet rays and X-rays, protons from the Sun, and, passing through space and through the Earth as though it were transparent, a cosmic flux of neutrinos, which are almost indetectable uncharged particles produced during radioactive decay.

Scientists seek to detect these different radiations, to measure their energies and to determine the instant at which they interact with instruments, and to measure the total amount of particles or radiation that has passed through a particular position over a period of time. The absorption of radiation may also be used to provide information on the thickness of the absorbing material.

The widespread use of X-rays and radioisotopes in medicine and industry, the lethal concentrations of radioactive material in nuclear reactors, the possibilities of widespread contamination following nuclear warfare, all combine to generate a need for detectors and instruments that measure the degree of hazard existing in any given environment—the so-called health-physics instruments. The possibility of tagging individual molecules with radioactive atoms, and so extending the sensitivity of analytical methods, has produced a huge demand for detectors used in analytical instruments that find application in medicine, biology, and chemical analysis. Improvements in spectrographic techniques both on the surface of the Earth and in satellite environments give rise to a need for detectors of hard (high-energy) ultraviolet radiation, which merges into the soft (low-energy) X-ray region; these in general have to be specially developed for the experiments in hand.

Nuclear physics experiments demand increasingly complex arrays of detectors that are connected together in elaborate computer-controlled systems and that may involve the precise measurement of position of interaction of a particle of known energy and its time of interaction to an accuracy of a fraction of a nanosecond. In the industrial field, the control of nuclear reactors demands the use of specialized detectors, and the application of radioisotopes (radioactive atoms) in thickness gauges and in analytical instruments generates a need for a wide range of equipment. In the search for raw materials for nuclear power, a range of portable instruments involving rugged detectors has become necessary, and, of course, similar instruments merge into the health-physics aspects of civil-defense instrumentation. More complex instruments and detectors are used in geological evaluation of rock strata for the petroleum industry and in the search for other raw materials.

BIBLIOGRAPHY. JACK SHARPE, *Nuclear Radiation Detectors*, 2nd ed. rev. (1964), is a compact survey of the basic physics of detectors; W.H. TAIT, *Radiation Detection* (1980), is an introductory text; GLENN F. KNOLL, *Radiation Detection and Measurement* (1979), is more extensive, covering techniques. Specific detectors are dealt with in the following: GEOFFREY DEARNALEY and D.C. NORTHROP, *Semiconductor Counters for Nuclear Radiations*, 2nd ed. rev. (1966); C.H. WANG, DAVID L. WILLIS, and WALTER D. LOVELAND, *Radiotracer Methodology in the Biological, Environmental and Physical Sciences* (1975),

which includes a detailed explanation of liquid scintillation counting; D.H. WILKINSON, ***Ionization Chambers and Counters*** (1950, reissued 1970); BRUNO B. ROSSI and HANS H. STAUB, ***Ionization Chambers and Counters*** (1949, reprinted 1977). Of the last two books, Wilkinson's is predominantly theoretical in nature, whereas Rossi and Staub's deals with experimental techniques developed in the Manhattan Project. The International Atomic Energy Agency (IAEA), Vienna, publishes a number of books and reports that are readily available. In the field of health physics, reports no. 109 and 120 in the IAEA "Technical Reports Series" deal with personnel dosimetry systems and the monitoring of radioactive contamination on surfaces, respectively, whereas ***Nuclear Accident Dosimetry Systems*** is one of the IAEA "Proceedings Series" (all 1970). Papers on detectors are regularly published in many journals, including the ***Review of Scientific Instruments*** (monthly), of which the May 1961 issue was devoted entirely to spark chambers; ***Nuclear Instruments and Methods*** (bimonthly); and ***IEEE Transactions: Nuclear Science*** (bimonthly), which include the proceedings of the Scintillation and Semiconductor Counter Symposium held every other year since 1948.

Radiation Effects on Matter

The term radiation implies a flow or streaming of atomic and subatomic particles and of waves, such as those that characterize heat rays, light rays, and X-rays. A simple way of classifying radiation is by the speed of particles or waves in *vacuo:* all waves travel with the speed of light, whereas particulate matter moves with somewhat slower speeds. A comprehensive study of radiation includes the emission, propagation, and absorption of the particles or waves. This article is concerned only with the absorption of radiation as it encounters matter; that is, the behaviour of the radiation and the matter with which it interacts, and how energy is transferred from the radiation to its surroundings.

I. General background

RADIATION PHYSICS

Types of radiation. The first type of radiation to be considered consists of waves and is called electromagnetic radiation, because the concept of wave motion is an outgrowth of the theory of electromagnetic fields. Electromagnetic radiation travels with the velocity of light, approximately 3 × 10^{10} centimetres (186,000 miles) per second in free space. The second general type, particulate radiation, includes all radiations travelling at less than the speed of light in free space.

Electromagnetic radiation. According to the theory of relativity (*q.v.*), the velocity of light is a fixed quantity independent of the velocity of the emitter, the absorber, or a presumably independent observer, all three of which do affect the velocities of other kinds of waves, such as sound. In an extended definition, the term light embraces all waves that have the velocity of visible light, whether they be longer or shorter. It includes the following: the long electromagnetic waves predicted by the British physicist James Clerk Maxwell in 1864 and discovered by the German physicist Heinrich Hertz in 1887 (now called radio waves); infrared and ultraviolet light waves; the X-rays discovered in 1895 by Wilhelm Rontgen, a German physicist; the gamma rays that accompany many radioactive-decay processes; and some even more energetic (with higher energy) X-rays and gamma rays produced as the normal accompaniment of the operations of ultrahigh-energy machines (*i.e.*, particle accelerators, once erroneously called atom smashers, such as the Van de Graaff generator, the cyclotron and its variants, and the linear accelerator).

X-rays and gamma rays

Particulate radiation. Unlike X-rays and gamma rays, some high-energy radiations travel at less than the speed of light. Some of these were identified initially by their particulate nature and only later shown to travel with wavelike character. One example of this class is the electron, first established as a negatively charged particle in 1897 by the English physicist Joseph John Thomson and later as the component of beta rays emitted by radioactive elements. The electron was shown by the U.S. physicist Robert Millikan in 1910 to have a fixed charge and by George Paget Thornson, an English physicist, and the U.S. physicists Clinton Joseph Davisson and Lester Halbert Germer (1927) to have wavelike as well as particulate character. Electrons classified as radiation have velocities that range from as low as 10^8 centimetres per second to almost the speed of light. The negative electron, still commonly called an electron, is identified more precisely as a negatron. In 1932 the U.S. physicist Carl Anderson demonstrated the existence of a positive electron, generally called a positron and identified as one of the antiparticles of matter (every subatomic particle has an antiparticle, and a collision between them annihilates both, ultimately producing energy). The collision of a positron and an electron results in the intermediate production of a short-lived atom-like system called positronium, which decays in about 10^{-7} second into two gamma rays. Other particles commonly classified as radiation when travelling with high velocity include the positively charged nucleus of the hydrogen atom, or proton; the nucleus of deuterium (*i.e.*, heavy hydrogen, the nucleus of which has double the mass of normal hydrogen's nucleus), or deuteron, also positively charged; and the nucleus of the helium atom, or alpha particle, which has a double positive charge. The more massive positive nuclei of other atoms show similar wavelike properties when sufficiently accelerated in an electric field. All charged-particle radiations have a charge exactly equal to that of the negative or positive electron or to some integral multiple of that charge.

Annihilation radiation

The neutron is also a radiation; it is a particle emitted in certain radioactive-decay processes and in fission, the process in which a nucleus splits into two smaller nuclei; it decays in free space with a 12- to 13-minute half-life; *i.e.*, one-half of any given number of neutrons decay within 12–13 minutes, each into a proton and a negatron plus an antineutrino. The antineutrino is an uncharged particle that has zero mass when at rest and a velocity equal to that of light. The mass of the neutron approximates that of the hydrogen atom, about 1,850 times the mass of the electron.

Another class of elementary particles, called mesons, includes both positively charged particles and negatively charged particles (*i.e.*, with the same charge as that of an electron), as well as uncharged particles. The masses of mesons are always greater than those of electrons, and most have a mass less than that of the proton; a few have slightly greater mass. Although all these particles are classified as radiations, they are so few that their chemical effects are not presently studied. Because they are part of the constant bombardment from free space to which all matter is constantly exposed, however, they may have considerable effects, such as contributing to the processes of aging and evolution.

Radiation effects of mesons

The structure and properties of matter. Matter in bulk is composed of particles that, compared to radiation, may be said to be at rest, but the motion of the molecules that compose matter, which is attributable to its temperature, is equivalent to travel at the rate of hundreds of metres per second. Although matter is commonly considered to exist in three forms, solid, liquid, and gas, a review of the effects of radiation on matter must also include mention of the interactions of radiation with glasses, attenuated (low-pressure) gases, plasmas, and matter in states of extraordinarily high density. A glass appears to be solid but is actually a liquid of extraordinarily high viscosity, or a mixture of such a liquid and embedded microcrystalline material, which unlike a true solid remains essentially disorganized at temperatures much below its normal freezing point. Low-pressure gases are represented by the situation that exists in free space, in which nearest neighbour molecules, atoms, or ions may be literally centimetres apart. Plasmas, by contrast, are regions of high density and temperature in which all atoms are dissociated into their positive nuclei and electrons.

The capability of analyzing and understanding matter depends upon the details that can be observed and to an important extent upon the instruments that are used. Bulk, or macroscopic, matter is detectable directly by the senses supplemented by the more common scientific instruments, such as microscopes, telescopes, and balances.

It can be characterized by measurement of its mass and, more commonly, its weight, by magnetic effects, and by a variety of more sophisticated techniques, but most commonly by optical phenomena—by the visible or invisible light (*i.e.*, photons) that it absorbs, reflects, or emits or by which its observable character is modified (cf. PHOTOCHEMICAL REACTIONS). Energy absorption, which always involves some kind of excitation, and the opposed process of energy emission depend on the existence of ground-state and higher energy levels of molecules and atoms. A simplified system of energy states, or levels, is shown schematically in Figure 1. Such a system is exactly fixed for each atomic and molecular system by the laws of quantum mechanics; the "allowed," or "permitted," transitions between levels, which may involve energy gain or loss, are also established by those same laws of nature. Excitation to energy levels above those of the energetically stable molecules or atoms may result in dissociation or ionization: molecules can dissociate into product molecules and free radicals, and, if the energy absorption is great enough, atoms as well as molecules can yield ions and electrons (*i.e.*, ionization occurs). Atomic nuclei themselves may exist in various states in which they absorb and emit gamma rays under certain conditions, and, if the nuclei are raised to, or by some process left in, energy states that are sufficiently high, they may themselves emit positrons, negatrons, alpha particles, or neutrons (and neutrinos), or dissociate into the nuclei of two or more lighter atoms. The resulting atoms may be similarly short-lived and unstable or they may be extremely long-lived and quite stable.

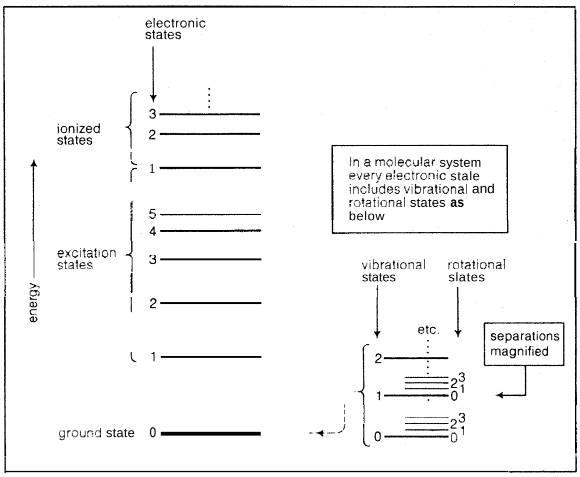


Figure 1: Energy states in molecular systems (see text).

Absorption and quantum laws

CHEMICAL AND BIOLOGICAL EFFECTS OF RADIATION

All matter is subject to the effects of radiation. Effects of radiation on chemical structure and physical properties of matter are, respectively, the concerns of radiation chemistry and of radiation physics. The two subjects are not sharply delineated, however, and, apart from such phenomena as induced radioactivity, are included in the general province of the radiation chemist. Photochemistry is the study of the chemical effects of light. Although radiation chemistry can be considered to include photochemistry, the latter has sufficient history and scientific content to be studied and reported separately. Common photochemical changes are exhibited in the photographic processes: photosynthesis of sugars, amino acids, and related compounds in living plants; and the tanning of living (light-coloured) skin exposed to the sun's rays. In certain cases the excited states that result from light absorption lose their energy in luminescence (the spontaneous emission of light, as in the processes called fluorescence and phosphorescence). Many of the photochemical changes induced by visible and ultraviolet light can also be produced to some degree by higher energy radiations. The rays and fast charged particles of high-energy radiation can be more penetrating than light. They may fog apparently unexposed photographic plates (Becquerel discovered radioactivity in this way); they may cause cancers in inadequately protected flesh or, on the other hand, be used to cure cancers under controlled conditions; and they can cause both destructive and (very infrequently) viable mutations in living organisms at all stages of complexity, from single cells to fruit flies, plants, and highly developed mammals.

Radiation sickness

Radiation sickness is a disagreeable and potentially dangerous consequence of exposure to large doses of high-energy radiation. The more obvious form of such sickness is accompanied by indigestion and nausea of varying intensity, lasting from hours to a month or more. General overexposure (as in accidents) also results in more subtle damage to the bone marrow, which in extreme cases can require total replacement to effect a cure—if a cure is possible at all.

Evolution of life forms. From the human point of view, the interaction of radiation with matter can be considered the most important process in the universe. When the universe began to cool down, at an early stage in its evolution, stars, such as the Sun, and planets appeared, and elements such as hydrogen (H), oxygen (0), nitrogen (N), and carbon (*C*) combined into simple molecules such as water (H_2O), ammonia (NH,), and methane (CH,). As a result of the action of far-ultraviolet light (wavelength less than 185 nanometres) before oxygen appeared in the atmosphere, and of penetrating alpha, beta, and gamma radiations, as well as of electric discharges from lightning storms when the temperature dropped and water began to condense, compounds such as the larger hydrocarbons, alcohols, aldehydes, acids, and amino acids were ultimately built. These simple compounds interacted and eventually developed into living matter. To what degree—if at all—the radiations from radioactive decay contributed to the synthesis of living matter is not known, but the existence of high-energy-irradiation effects on matter at very early times in the history of this world is recorded in certain micas as microscopic, concentric rings, called pleochroic haloes, produced as the result of the decay of tiny specks of radioactive material that emitted penetrating products, such as alpha particles. At the termini of their paths, such particles produced chemical changes, which can be seen microscopically as dark rings. From the diameters of the rings and the known penetrating powers of alpha particles from various radioactive elements, the nature of the specks of radioactive matter can be established. In some cases, alpha particles could not have been responsible for the effects observed; in other cases, the elementary specks that occupied the centres of the haloes were not those of any presently known elements.

It can be readily surmised that some of the elements that participated in the evolution of the world were not originally present but were produced as the result of external high-energy bombardment; that some disappeared as the result of such processes; and that many compounds required for the living processes of organisms evolved as a consequence of the high-energy irradiation to which all matter is subjected.

Role of radiation in life systems. Most of the effects of radiation on living matter are destructive. Mutations produced by radiation are rarely viable and persistent, but those that are can be useful, as in the production of new strains of seeds in botany and agriculture. X-rays, gamma rays, and beta and alpha particles are used in cancer treatment because the mutants produced by their action on a cancer may be nonviable; *i.e.*, the cancer can be stimulated to its self-destruction. On the other hand, healthy tissue may be so affected by penetrating radiation that a viable cancer, ultimately destructive to the host, is produced. This last effect is the danger against which those who work with radioactive materials or penetrating radiation must protect themselves.

Necessity of radiation to life

Radiation effects are essential to the existence of life on Earth. Radiation in its various forms is ultimately responsible not only for the existence of life but for the variety of its forms. Light from the Sun, through photosynthesis, is the primary source of foodstuffs and of the fossil fuel deposits represented in coal, petroleum, and natural gas. Higher energy radiations, both from the Sun and from outer space, affect weather and radio communications and, supplemented by natural and artificially produced radioactivity, are sources of both adventitious and deliberately induced mutations. The probability of ultimately lethal effects has stimulated such work as research on tsetse flies and malarial mosquitoes, in which it has been found an effective control procedure to expose males to gamma radiation and thus to make them sterile. Released into a general population of females that mate but once, the irradiated males mate frequently and thus reduce the number of offspring.

Radiation in photographic process. The effect of radiation on silver halides is the basis of the photographic process. It was originally confined to recording merely what can be seen with the eye, but the same process was also soon found to report effects produced by the more energetic (and invisible) ultraviolet light. Improvements involving the use of photosensitizers made it possible to take photographs in the dark with infrared light and with suitable filters to photograph the limited areas of a scene that reflect infrared to some special degree. From the military point of view, infrared photographic surveillance greatly diminishes the effectiveness of protective camouflage. Weather satellites are another example of an effective use of infrared surveillance.

The discovery of X-rays in 1895 was soon followed by employment of X-ray photography in medicine, and later by the use of both X-rays and gamma rays in routine examination of metallurgical products for voids and defects.

STUDIES OF RADIATION CHARACTERISTICS

The behaviour of light seems to have interested ancient philosophers but without stimulating them to experiment, though all of them were impressed by vision. The Greek philosopher Aristotle considered vision to involve a light source, the passage of a beam from the eye of the observer to the thing observed, and passage of a beam from the thing observed to the eye of the observer. The first meaningful optical experiments on light were performed by the English mathematician Isaac Newton (beginning in 1666), who showed (1) that white light diffracted by a prism into its various colours can be reconstituted into white light by a prism oppositely arranged and (2) that light of a particular colour selected from the diffracted spectrum of a prism cannot be further diffracted into beams of other colour by an additional prism. Newton hypothesized that light is corpuscular in its nature, each colour represented by a different particle speed, an erroneous assumption. Furthermore, in order to account for the refraction of light, the corpuscular theory required, contrary to the wave theory of the Dutch scientist Christiaan Huygens (developed at about the same time), that light corpuscles travel with greater velocity in the denser medium. Support for the wave theory came in the electromagnetic theory of Maxwell (1864) and the subsequent discoveries of H.R. Hertz and of Rontgen of both the very long and the very short waves Maxwell had included in his theory. Max Planck, a German physicist, proposed a quantum theory of radiation to counter some of the difficulties associated with the wave theory of light, and in 1905 Einstein proposed that light is composed of quanta (later called photons). Thus, experiment and theory had led around from particles (of Newton) that behave like waves (Huygens) to waves (Maxwell) that behave like particles (Einstein), the apparent velocity of which is unaffected by the velocity of the source or the velocity of the receiver. Furthermore it was found, in 1922, that the shorter wavelength electromagnetic radiations, such as X-rays, have momentum such as may be expected of particles, part of which can be transferred to electrons with which they collide (*i.e.*, the Compton effect).

Quantum theory of light

Electrons themselves had already been shown to be particles of fixed charge. In 1927 experiments were reported that demonstrated the wave nature of the electron: diffracted from crystals, electrons behave like waves of measurable wavelength; in passage through thin metallic foils, electrons are scatteied like waves also of measurable wavelength. As mentioned earlier, the wavelike character of all fast particles has been established.

II. Fundamental processes involved in radiation interaction with matter

THE PASSAGE OF ELECTROMAGNETIC WAVES

The field concept. A discussion of this subject requires preliminary definition of a few of the more common terms. Around every particle, whether it be at rest or in motion, whether it be charged or uncharged, there are potential fields of various kinds. As one example, a gravitational field exists around the Earth and indeed around every particle of mass that moves with it. At every point in space, the field has direction in respect to the particle. The strength of the gravitational field around a specific particle of mass, m, at any distance, r, is given by the product of g, the universal gravitational constant, and m divided by the square of r, or gm/r^2. The field extends indefinitely in space, moves with the particle when it moves, and is propagated to any observer with the velocity of light. Newton showed that the mass of a homogeneous spherical object can be assumed to be concentrated at its centre and that all distances can be measured from it. Similarly, electric fields exist around electric charges and move with them. Magnetic fields exist around electric charges in motion and change in intensity with all changes in the accompanying electric field, with the magnetic field at any point being perpendicular to the electric field in free space. Any regular oscillation is time-dependent, as is any change in field strength with time.

Time-dependent electric and magnetic fields occur jointly; together they propagate as what are called electromagnetic waves. In an assumed ideal free space (without intrusion from other fields or forces of any kind, devoid of matter, and, thus, in effect without any intrusions, demarcations, or boundaries), such waves propagate with the speed of light in the so-called transverse electromagnetic mode; *i.e.*, the directions of the electric field, the magnetic field, and the propagation of the wave are mutually perpendicular. They constitute a right-handed coordinate system; *i.e.*, with the thumb and first two fingers of the right hand perpendicular to each other, the thumb points in the direction of the electric field, the forefinger in that of the magnetic field, and the middle finger in that of propagation. A boundary may be put on the space by appropriate physical means (bound space), or the medium may be something other than a vacuum (material medium). In either case, other forces and other fields come into the picture, and propagation of the wave is no longer exclusively in the transverse electromagnetic mode. Either the electric field or the magnetic field (a matter of arbitrary choice) may be considered to have a component parallel to the direction of propagation itself. It is this parallel component that is responsible for attenuation of energy of the waves as they propagate.

Light: transverse electromagnetic waves

Frequency range. Electromagnetic waves span an enormous range of frequencies (*i.e.*, number of oscillations per second), only a small part of which fall in the visible region. Indeed, it is doubtful that lower or upper limits of frequency exist, except in regard to the applicability of present-day instrumentation. Figure 2 indicates the usual terminology employed for electromagnetic waves of different frequency or wavelength. Customarily, scientists designate electromagnetic waves by fields, waves, and particles in increasing order of the frequency ranges to which they belong. Traditional demarcations into fields, waves, and particles (*e.g.*, gamma-ray photons) are shown in the figure. The distinctions are largely of classical (*i.e.*, nonquantum) origin; in quantum theory there is no need for such distinctions. They are preserved, however, for common usage. The term "field" is used in a situation in which the wavelength of the electromagnetic waves is larger than the physical size of the experimental setup. For wave designation, the wavelength is comparable to or smaller than the physical extent of the setup, and at the same time the energy of the photon is low. The particle description is useful when wavelength is small and photon energy is high.

Classification of electromagnetic regions

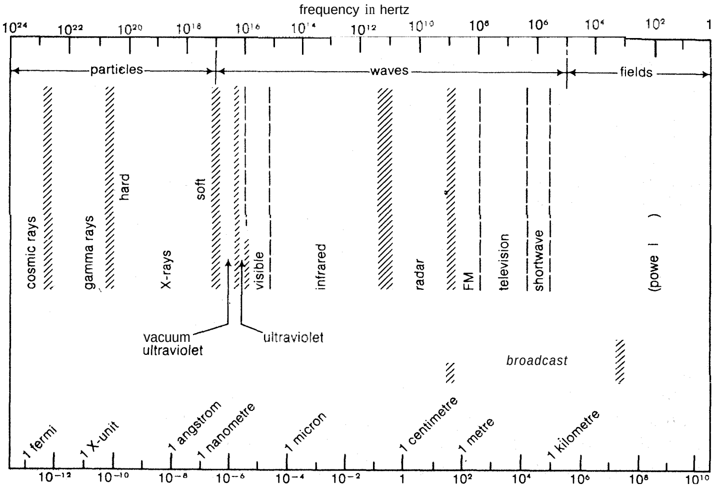


Figure 2: Electromagnetic spectrum.
Two scales are shown: the frequency scale, expressed in hertz (Hz), or cycles per second, and the wavelength scale, expressed in centimetres. Other units of wavelength customarily used to define certain regions of the spectrum are also given; *e.g.,* the nanometre (nm, 1/1,000,000,000 metre) the unit useful in the visible region, with an approximate range from 760 nm (red) to 380 nm (violet). Demarcations between the various classifications of radiations (*e.g.,* gamma rays and X-rays) are not sharp.

Properties of **light.** The ordinary properties of light, such as straight-line propagation, reflection and refraction (bending) at a boundary or interface between two mediums, and image formation by mirrors or lenses, can be understood by simply knowing how light propagates, without inquiring into its nature. This area of study, essentially, is geometrical optics. On the other hand, the extraordinary properties of light do require answers to questions regarding its nature (physical optics). Thus, interference, diffraction, and polarization relate to the wave aspect, while photoelectric effect relates to the particle aspect of light. Apparently light has dual character. It is the duality in the nature of light, as well as that of matter, that led to quantum theory.

In general, radiation interacts with matter; it does not simply act on nor is it merely acted upon. Understanding of what radiation does to matter requires also an appreciation of what matter does to radiation.

When a ray of light is incident upon (falls on) a plane surface separating two mediums (for example, air and glass) it is partly reflected (*i.e.*, thrown back into the original medium) and partly refracted (*i.e.*, transmitted into the other medium). Laws of reflection and refraction state that all the rays (incident, reflected, and refracted) and the normal (a perpendicular line) to the surface lie in the same plane, called the plane of incidence. Angles of incidence and reflection are equal; for any two mediums the sines of the angles of incidence and refraction have a constant ratio, called the mutual refractive index. All these relations can be derived from the electromagnetic theory of Maxwell, which constitutes the most important wave theory of light; but electromagnetic theory is not necessary to demonstrate these laws.

Maxwell's theory of radiation

Double refraction. In double refraction, light enters a crystal the optical properties of which differ along two or more of the crystal axes. What is observed depends on the angle of the beam with respect to the entrant face. Double refraction was first observed in 1669 by Erasmus Bartholin in experiments with Iceland spar crystal and elucidated in 1690 by Huygens.

If a beam of light is made to enter an Iceland spar crystal at right angles to a face, it persists in the crystal as a single beam perpendicular to the face and emerges as a single beam through an opposite parallel face. If the exit face is at an angle not perpendicular to the beam, however, the emergent beam is split into two beams at different angles, called the ordinary and extraordinary rays, and they are usually of different intensities. Clearly, any beam that enters an Iceland spar crystal perpendicular to its face and emerges perpendicular to another face is of changed character—although superficially it may not appear to be changed. Dependent on the relative intensities and the phase relationship of its electric components (*i.e.*, their phase shift), the beam is described as either elliptically or circularly polarized. There are other ways of producing partially polarized, plane-polarized, and elliptically (as well as circularly) polarized light; but these examples illustrate the phenomena adequately.

Polarization of an electromagnetic wave can be shown mathematically (see ELECTROMAGNETIC RADIATION) to relate to the space–time relationship of the electromagnetic-field vector (conventionally taken as the electric vector, a quantity representing the magnitude and direction of the electric field) as the wave travels. If the field vector maintains a fixed direction, the wave is said to be plane-polarized, the plane of polarization being the one that contains the propagation direction and the electric vector. In the case of elliptic polarization the field vector generates an ellipse in a plane perpendicular to the propagation direction as the wave proceeds. Circular polarization is a special case of elliptic polarization in which the so-described ellipse degenerates into a circle.

Kinds of polarization

An easy way to produce circularly polarized light is by passage of the light perpendicularly through a thin crystal (*e.g.*, mica). The mica sample is so selected that the path difference for the ordinary and the extraordinary rays is one-quarter the wavelength of the single-wavelength, or monochromatic, light used. Such a crystal is called a quarter-wave plate, and the reality of the circular polarization is shown by the fact that, when the quarter-wave plate is suitably suspended and irradiated, a small torque —that is, twisting force--can be shown to be exerted on it. Thus, the action of the crystal on the light wave is to polarize it; the related action of the light on the crystal is to produce a torque about its axis.

The ratio of the intensity of the reflected light to that of the incident light is called the reflection coefficient. This quantitative measure of reflection depends on the angles of incidence and refraction. or the refractive index, and also on the nature of polarization.

It can be shown that the reflection coefficient at any angle of incidence is greater for polarization perpendicular to the plane of incidence than for polarization in the plane of incidence. As a result, if unpolarized light is incident at a plane surface separating two media, reflected light will be partially polarized perpendicular to the plane of incidence, and refracted light will be partially polarized in the plane of incidence. An exceptional case is the Brewster angle, which is such that the sum of the angles of incidence and refraction is 90". When that happens, the reflection coefficient for polarization in the plane of incidence equals zero. Thus, at the Brewster angle, the reflected light is wholly polarized perpendicular to the plane of incidence. At an air–glass interface, Brewster's angle is approximately 56°, for which the reflection coefficient for perpendicular polarization is 14 percent. Another important angle for refraction is the critical angle of incidence when light passes from a denser to a rarer medium. It is that angle for which the angle of refraction is 90° (in this case the angle of refraction is greater than the angle of incidence). For angles of incidence greater than the critical angle there is no refracted ray; the light is totally reflected internally. For a glass-to-air interface the critical angle has a value 41"48'.

Brewster's angle

Dispersion. The variation of the refractive index with frequency is called dispersion; it is this property of a prism that effects the colour separation, or dispersion, of white light. An equation that connects the refractive index with frequency is called a dispersion relation. **A** simple classical dispersion relation can be obtained by consideration of the joint effect of (1) the oscillation of the electrons bound to the atoms in the absence of an external field and (2) the frequency of the external electromagnetic field imposed on the system, with appropriate consideration (3) of the effect of the number of electrons per cubic centimetre of the material and (4) of the charge and mass of the electron. For visible light the index of refraction increases slightly with frequency, a

Propagation velocity of electromagnetic energy

phenomenon called normal dispersion. The degree of refraction depends on the refractive index. The increased bending of violet light over red by a glass prism is, therefore, the result of normal dispersion. If experiments are done, however, with light having a frequency close to the natural electron frequency, some strange effects appear. When the radiation frequency is slightly greater, for example, the index of refraction becomes less than unity (<1) and decreases with increasing frequency; the latter phenomenon is called anomalous dispersion. A refractive index less than unity refers correctly to the fact that the speed of light in the medium at that frequency is greater than the speed of light in vacuum. The velocity referred to, however, is the phase velocity or the velocity with which the sine-wave peaks are propagated. The propagation velocity of an actual signal or the group velocity is always less than the speed of light in vacuum. Therefore, relativity theory is not violated. An example is shown in Figure 3. in which a light source is initially pointed in the direction *A*. The source rotates in such a way that the velocity of the light image moves from D to E with a velocity v approximating c. Thus, the phase velocity with which the image moves from *A* to B is greater than c, but the relativity principle is not violated because the velocity of transmission of matter or energy does not exceed the velocity of light.

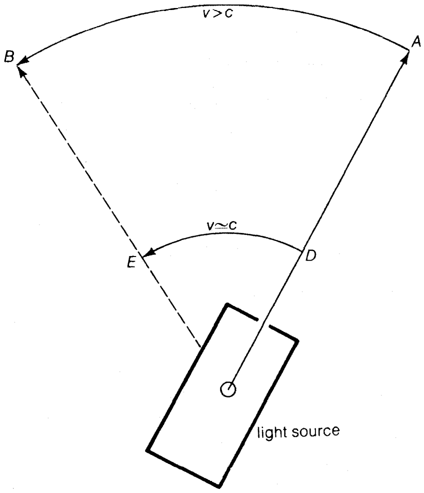


Figure 3: Contrast of phase velocity, v, and wave velocity, *c*. As the light source turns counterclockwise on its axis, the velocity with which the signal moves (*e.g.*, from *A* to *B*) can exceed the velocity of light without violation of the relativity principle (see text).

The quantum-mechanical form of the dispersion relation is similar to the form of that relation in classical theory but is different in concept and details.

Electromagnetic waves and atomic structure. Quantum concepts. Quantum mechanics includes such concepts as "allowed states"; *i.e.*, stationary states of energy content exactly stipulated by its laws. The energy states shown in Figure 1 are of that kind. A transition between such states depends not only on the availability (*e.g.*, as radiation) of the precise amount of energy required but also on the quantum-mechanical probability of such a transition. That probability, the oscillator strength, involves so-called selection rules that, in general terms, state the degree to which a transition between two states (which are described in quantum-mechanical terms) is allowed. As an illustration of allowed transition in Figure 1, the only electronic transitions permitted are those in which the change in vibrational quantum number accompanying a change in electronic excitation is plus or minus one or zero,-except that a 0 ↔ 0 (zero to zero) change is not permitted. All electronic states include vibrational and rotational levels, so that the probability of a specific electronic transition includes the probabilities of transition between all the vibrational and rotational states that can conceivably be involved. Figure 1 is, of course, a simplified picture of a compendium of energy states available to a molecule (polyatomic structure)—and the selection rules are accordingly more involved in such a case. The selection rules are worked out by scientists in a process of discovery; the attempt is to state them systematically so that the applicable rules in an experimentally unstudied case may be stated on the basis of general principle.

Oscillator strengths

Resonance. Calculations based on quantum-mechanical principles yield resonance frequencies or oscillator strengths between two states that, as an example, describe dispersion.

It is notable that the refractive index in the classical and quantum-mechanical relations is complex; that is, it includes a factor of the square root of minus one, which is imaginary. Its imaginary part is associated with the damping factor symbolized by the Greek letter gamma, Γ, a measure of the rate at which the intensity of radiation diminishes along its path. The real and imaginary parts of the refractive index can be separated into a general form so that the index of refraction (n) for a given frequency symbolized by the Greek letter omega, ω, expressed as $n(\omega)$, is equal to the difference between a real index $n_1(\omega)$ and an imaginary index, written as $in_2(\omega)$, in which i is the imaginary $\sqrt{-1}$; *i.e.*, as $n(\omega) = n_1(\omega) - in_2(\omega)$. The real part of the refractive index, $n_1(\omega)$, represents no attenuation but only a time delay; the imaginary part, $n_2(\omega)$, represents a true attenuation. The intensity of light decreases exponentially with distance; that is, it decreases in the same ratio of final to initial intensity in travel through the same distance. The energy loss from the light wave appears as energy added to the medium, or what is known as absorption. The coefficient of absorption, *a*, is equal to the reciprocal of the distance over which the light intensity is reduced by a factor of *e* (the logarithmic base e $=$ 2.718). The index of absorption is simply the imaginary index n_2. They are both dependent on frequency and are related so that the coefficient of absorption is equal to the frequency times the index of absorption divided by the velocity of light (c), or a $= \omega n_2/c$.

Reflection and transmission coefficients

Reflection and transmission coefficients (ratio of intensity of the reflected and transmitted beams to the incident) at a boundary of an absorbing medium are valid as long as the refractive index is considered to be complex. In addition to attenuation it implies a phase shift between the two polarization components, parallel and perpendicular to the plane of incidence, for both the reflected and the transmitted light. Thus, even if the incident light is unpolarized, both the reflected and the transmitted light are elliptically polarized. Another effect of absorption on reflection is that, at Brewster's angle (at which the sum of the angles of incidence and refraction is 90°), the reflection coefficient for the parallel polarization component is not zero but finite.

Absoiption *and* emission. Absorbing mediums are classified as weskly or strongly absorbing depending upon whether the real index is much greater or much less than (symbolized $<<$) the imaginary index ($n_2 << n_1$ or $n_1 << n_2$). The intermediate situation is characterized by the two indices being approximately (symbolized $\simeq$) the same ($n_1 \simeq n_2$). A medium can, of course, be weakly absorbing at one region of the electromagnetic spectrum and strongly absorbing at another. Most substances that are transparent in the visible range also absorb weakly in the visible range. Frequently, however, they absorb strongly in the ultraviolet at or close to their resonance frequencies. If a medium is weakly absorbing, its dispersion and absorption can be measured directly from the intensity of refracted or transmitted light. If it is strongly absorbing, on the other hand, the light does not survive even a few wavelengths of penetration. The refracted or transmitted light is then so weak that measurements are at best difficult. The absorption and dispersion in such cases, nevertheless, may still be determined by studying the reflected light only. This procedure is possible because the intensity of the reflected light contains the refractive index of which the real and imaginary parts separate neatly into dispersion and absorption, respectively. In the far ultraviolet, it is the only practical means of studying absorption, a study that has revealed valuable information about electronic energy levels and collective energy losses (see below *Molecular* activation) in condensed material.

Laser source of radiation

Experimental studies of the chemical effects of radiation on matter can be greatly forwarded by the use of beams of high intensity and very short duration. Such studies are made possible by employment of the laser, a light source developed by United States physicists A.L. Schawlow and C.H. Townes (1958) from the application of one of the Einstein equations. Einstein suggested (on the basis of a principle of detailed balancing, or microscopic reversibility) that, just as the amount of light absorbed by a molecular system in a light field must depend on the intensity of the light, the amount of light emitted from excited states of the same system must also exhibit such dependency. In this fundamentally important idea of microscopic reversibility can be seen one of the most dramatic illustrations of the physical effects of radiation.

Under any circumstance, the absorption probability in the ground state is given by the number of molecules (or atoms), N_i, in that state multiplied both by the probability, B,$_j$, for transition from state i to state *j* and by the light intensity, $I(\nu)$, at frequency symbolized by the Greek letter nu, ν; *i.e.*, $N_iB_{ij}I(\nu)$. Light emission from an excited state to the ground state depends on the number of molecules (or atoms) in the upper state, N,, multiplied by the probability of spontaneous emission, A_{ji}, to the ground state plus the additional induced emission term, $N_jB_{ji}I(\nu)$, in which B_{ji} is a term that Einstein showed to be equal to B,, and that relates the probability of such induced emission, so that in the general case in any steady-state situation (in which light absorption and emission are occurring at equal rates):

$$N_iB_{ij}I(\nu) = N_j[A_{ji} + B_{ji}I(\nu)].$$

There is a well-developed theoretical relationship (not here presented) of a quantum-mechanical nature between **A,,** and B_{ij}. Ordinarily, the light intensity, $I(\nu)$, is so low that the second term on the right can be neglected. At sufficiently high light intensities, however, that term can become important. In fact, if the light intensity is high, as in a laser, the induced-emission probability can easily exceed that of spontaneous emission.

Spontaneous emission of light is random in direction and phase. Induced emission has the same direction of polarization and propagation as that of the incident light. If by some means a greater population is created in the upper level than in the lower one, then, under the stimulus of an incident light of appropriate frequency, the light intensity actually increases with path length provided that there is enough stimulated emission to compensate for absorption and scattering. Such stimulated emission is the basis of laser light. Practical lasers such as the ruby or the helium–neon lasers work, however, on a three-level principle.

Mechanism of pair production

The Compton effect. The energy required to remove an orbital electron from an atom (or molecule) is called its binding energy in a given state. When light of photon energy greater than the minimum binding energy is incident upon an atom or solid, part or all of its energy may be transformed through the photoelectric effect, the Compton effect, or pair production—in increasing order of importance with increase of photon energy. In the Compton effect, the photon is scattered from an electron, resulting in a longer wavelength, thus imparting the residual energy to the electron. In the other two cases the photon is completely absorbed or destroyed. In the pair-production phenomenon, an electron–positron (positively charged electron) pair is created from the photon as it passes close to an atomic nucleus. **A** minimum energy (1,020,000 electron volts) is required for this process because the energy of the electron–positron pair at rest—the total mass, $2m$, times the velocity of light squared ($2mc^2$)—must be provided. **If** the photon energy (hv) is greater than the rest mass, the difference ($h\nu - 2mc^2$), called the residual energy, is distributed between the kinetic energies of the pair with only a small fraction going to the nuclear recoil.

The photoelectric effect. The photoelectric effect is caused by the absorption of electromagnetic radiation and consists of electron ejection from a solid (or liquid) surface, usually of a metal, though nonmetals have also been studied. In the case of a gas, the term photo-ionization is more common, though there is basically little difference between these processes. Despite experimental difficulties connected with surface-adsorbed gas and energy loss of ejected electrons in penetrating a layer of the solid into vacuum, early experimenters established two important features about the photoelectric effect. These are: (1) although the photoelectric current (*i.e.*, the number of photoelectrons) is proportional to the incident-light intensity, the energy of the individual photoelectrons is independent of light intensity; and (2) the maximum energy of the ejected electron is roughly proportional to the frequency of light. On the basis of these observations, Einstein argued that the light is absorbed in quanta of energy equal to Planck's constant (h) times light frequency, $h\nu$, by electrons, one at a time. **A** minimum energy symbolized by the Greek letter psi, ψ, called the photoelectric work function of the surface, must be supplied before the electron can be ejected. When a quantum of energy is greater than the work function, photoelectric emission is possible with the maximum energy symbolized by the Greek letter epsilon, ε, of the photoelectron (ε_{max}) being stated by Einstein's photoelectric equation as equalling the difference between the photon energy and the work function; *i.e.*,,,& $= h\nu -$ $. Einstein's interpretation gave strong support for the quantum theory of radiation. Early experiments determined Planck's constant, h, independently through the above equation and also established the fact that an immeasurably small time delay is involved between absorption of a quantum of light and the ejection of an electron.

Photoelectric threshold frequency

Accurate and reliable values of the work function and ejection energy are now available for most solids; the chief obstacles to the development of such data were the difficulty of preparation of clean surfaces and the energy loss of electrons in penetration into vacuum. The photoelectric threshold frequency, symbolized by the Greek letter nu with subscript zero, ν_0, is one at which the effect is barely possible; it is given by the ratio of the work function symbolized by the Greek letter psi, ψ, to Planck's constant ($\nu_0 = \psi/h$). The photoelectric yield, defined as the ratio of the number of photoelectrons to that of incident photons, serves as a measure of the efficiency of the process. Photoelectric yield starts from a zero value at threshold, reaches a maximum (about 1/1,000) at about twice the threshold frequency, and falls again when frequency is further increased. Some unusual alloys exhibit yields a hundred times greater than normal (*i.e.*, about 0.1). Normally the yield depends also on polarization and angle of incidence of the radiation. Parallel polarization (polarization in the plane of incidence) gives higher yield than does perpendicular polarization, sometimes by almost ten times.

Light interacts with matter in a variety of ways (*e.g.*, reflection, refraction, interference, scattering); it can induce chemical change in a molecule almost exclusively by absorption only.

Cross section. **A** useful concept in describing the absorption of radiation in matter is called cross section; it is a measure of the probability that photons interact with matter by a particular process. When the energy of each individual photon ($h\nu$) is much smaller than the rest energy of the electron (its mass times the velocity of light squared [mc^2]), the scattering of photons is described by a cross section derived by the English physicist J.J. Thomson. This cross section is called the Thomson cross section, symbolized by the Greek letter sigma with subscript zero, ***a,,*** and is equal to a numerical factor times the square of the term, electric charge squared divided by electron rest energy, or $\sigma_0 = (8\pi/3)(e^2/mc^2)^2$. When the photon energy is equal to or greater than the electron's rest energy of ($h\nu \geq mc^2$), inelastic (*i.e.*, energy loss) scatterings begin to appear. One such is Compton scattering, in which an X-ray or gamma ray (electromagnetic radiation from an atomic nucleus) experiences an increase in wavelength (reduction in energy) after being scattered through an angle. Arthur Holly Compton, a physicist in the United States, correctly interpreted the effect by using laws of classical relativistic mechanics. He showed that the increase in wavelength symbolized by the Greek letters delta and lambda, *Ah,* is independent of the energy of the photon and is given by an expression in which the product of two terms appears. The first is a universal constant symbolized by the Greek letter lambda with subscript zero, λ_0, generally called the Compton wavelength, and itself equal to Planck's constant, *h*, divided by the mass of the electron at rest and the velocity of light; *i.e.*, X, $= h/mc = 2.4 \times 10^{-10}$ centimetre. The second is a term dependent on the angle symbolized by the Greek letter theta, θ, through which the photon is scattered; it is one minus the cosine of that angle, or $1 - \cos \theta$. The increase in wavelength observed at that angle is simply AX = X, $(1 - \cos \theta)$. In discussion of the Compton effect the electron is treated as free—that is, not bound to a nucleus—because, in the study of that effect for most materials of low atomic number, the incident photon has energy much greater than the binding energy. For bound electrons, the corrections to the Compton relation are small but complicated. When photons are scattered, the concept of differential cross sections may be used; differential cross section is a measure of the probability that a photon will be scattered within a given small angle.

Compton wavelength

The differential cross section for the Compton process was derived by the Swedish physicist Oskar Klein and the Japanese physicist Yoshio Nishina. The Klein–Nishina formula shows almost symmetrical scattering for low-energy photons about 90° to the beam directions. As the photon energy increases, the scattering becomes predominantly peaked in the forward direction, and, for photons with energies that are greater than five times the rest energy of the electron, almost the entire scattering is confined within an angle of 30". When averaged over the angle, the Klein–Nishina cross section shows variation with the incident photon energy. At low energy this cross section increases uniformly and approaches the classical Thomson value as energy is decreased; at high energy the cross section is inversely proportional to the energy. The energy distribution of Compton electrons (recoil or scattered electrons) and outgoing photons may also be derived from the Klein–Nishina theory. The result shows a wide distribution; for atoms of low atomic number and incident photon energies in the region of importance (*i.e.*, 1,000,000 to 100,000,000 electron volts), the probability of scattering per unit energy interval is fairly constant—except that, for the case of nearly total conversion of the photon energy into electron kinetic energy, a plot of energy versus angle shows a sharp, narrow peak. Thus, as a crude approximation, the average energy of a Compton electron is about half the incident photon energy.

Compton scattering plays a key role in the interaction of matter with intermediate-energy gamma rays and high-energy X-rays. For these radiations, it is almost the exclusive mechanism by which energy is transferred from the radiation and added to the matter. An example may be cited of the penetration of gamma rays from the radioactive substance cobalt-60 into a sample of water or aqueous solution. The electron density is approximately 3×10^{23} per millilitre. Taking the Compton cross section as approximately 3×10^{-25} square centimetre per electron, calculation yields a mean free path for Compton scattering of about ten centimetres (*i.e.*, a photon will move about ten centimetres between successive encounters with electrons). The dominant radiation effect produced by a gamma ray, therefore, is attributable to the recoil electron and the vast number of progeny (such as secondary and tertiary electrons) that are produced. These higher generation electrons are produced through electron-impact ionization (*i.e.*, an electron is removed from an atom by the collision of another electron), a process that continues until barred either by energetic limitation or by low cross section. For cobalt-60 gamma rays the average Compton energy in a material of low atomic number, such as water, is approximately 600,000 electron volts. The study of chemical transformation resultant from interaction of radiation with matter is called radiation chemistry. Normally the incident radiation is an X-ray or a gamma ray or a charged particle. There are areas of overlap between photochemistry and radiation chemistry, particularly in the range of wavelengths from about 150 to about 85 nanometres.

Energy transfer from gamma rays

Puir production. Pair production is a process in which a gamma ray of sufficient energy is converted into an electron and a positron. A fundamental law of mechanics, given by Newton, is that in any process total linear (as well as angular) momentum remains unchanged. In the pair-production process a third body is required for momentum conservation. When that body is a heavy nucleus, it takes very little recoil energy, and therefore the threshold is just twice the rest energy of the electron; *i.e.*, twice its mass, m, times the square of the velocity of light, c^2, or $2mc^2$. Pair production can also occur in the field of an atomic electron, to which considerable recoil energy is thereby imparted. The threshold for such a process is four times the rest energy of an electron, or $4mc^2$. The total pair-production cross section is the sum of the two components, nuclear and electronic. These cross sections depend on the energy of the gamma ray and are usually calculated in an electron theory proposed by British physicist P.A.M. Dirac through a method of approximation that is a simplification of a method (a "first approximation") given by the German physicist Max Born (*i.e.*, a "first Born approximation"). The process is envisaged by Dirac as the transition of an electron from a negative to a positive energy state. Corrections are required for these cross sections at high energy, at high atomic number, and for atomic screening (*i.e.*, the intrusion of the field of the electrons in an atom); these are normally made via numerical procedures. The fraction of residual energy, symbolized by the Greek letter alpha, unexpended in conversion of energy to mass, that appears in any one particle (*e.g.*, the electron) is thus given by the kinetic energy of that electron E_e minus its rest energy mc^2 divided by the energy of the gamma ray hv (*i.e.*, the product of Planck's constant and the frequency of the gamma ray) minus twice the rest energy of the electron $2mc^2$, or $\alpha =$ (E, $- mc^2)/(h\nu - 2mc^2)$. Because the same equation applies to each of the two electrons that are formed, it must be symmetric about the condition that each of the particles has half the residual energy, symbolized by the Greek letter alpha, α (in excess of that conveyed to the "third body"); *i.e.*, that $\alpha = 0.5$. Below an energy of about 10,000,000 electron volts for the gamma ray, the probability for pair production (*i.e.*, the pair-production cross section) is almost independent of the atomic number of the material, and, up to about 100,000,000 electron volts of energy, it is also almost independent of the quantity α. Even at extremely high energies the probability that a certain fraction of the total available energy will appear in one particle is almost independent of the fraction as long as energy is comparably distributed between the two particles (*i.e.*, excepting in cases in which almost all energy is dumped into one particle alone). Typical pair-production cross sections at 100 MeV (million electron volts) are approximately 10^{-24} to 10^{-22} square centimetre, increasing with atomic number. At high energies, approximately equal to or greater than 100 MeV, pair production is the dominant mechanism of radiation interaction with matter.

Born's "first approximation"

Clearly, as the photon energy increases, the dominant interaction mechanism shifts from photoelectric effect to Compton scattering to pair production. Rarely do photoelectric effect and pair production compete at a given energy. Compton scattering, however, at relatively low energy competes with the photoelectric effect and at high energy competes with pair production. Thus, in lead, interaction below 0.1 MeV is almost exclusively photoelectric; between 0.1 MeV and 2.5 MeV both photoelectric and Compton processes occur; between 2.5 MeV and 100 MeV Compton scattering and pair production share the interaction. In the pair process the photon is annihilated, and an electron-positron pair is created. On the other hand, an electron or positron with energy approximately

Brems-strahlung production

equal to or greater than 100 MeV loses its energy almost exclusively by production of high-energy bremsstrahlung (X-rays produced by decelerating electric charges) as the result of interaction with the field of a nucleus. The cross section for bremsstrahlung production is nearly independent of energy at high energies, whereas at low energies the dominant energy-loss mechanism is by the creation of ionizations and excitations. A succession of bremsstrahlung and pair-production processes generates a cascade or shower in the absorber substance. This phenomenon can be triggered by either an electron, a positron, or a photon, the triggering mechanism being unimportant as long as the starting energy is high. A photon generates a pair through pair production, and the charged particles generate photons through bremsstrahlung, and so on repeatedly so long as the energy is kept sufficiently high. With penetration into the substance, the shower increases in size at first, reaches a maximum, and then gradually reduces. Loss of particles by degradation to lower energies (in which the yield of bremsstrahlung is low), ionization loss, and production and absorption of low-energy photons eventually reduce the size of the cascade. The mathematical theory of cascades has been developed in great detail.

X-rays and gamma rays. When light of sufficiently high frequency (or energy equal to $h\nu$), independent of its source, is absorbed in a molecular system, the excited molecular state so produced, or some excited state resultant from it, may either interact with other molecules or decompose to produce intermediate or ultimate products; *i.e.*, chemical reactions ensue. Study of such processes is encompassed in the subject of photochemistry (see also below Molecular activation).

Electromagnetic waves of energy greater than those usually described as ultraviolet light (see Figure 2) are included in the classes of X-rays or gamma rays. X-ray and gamma-ray photons may be distinguished by definition on the basis of source. They are indistinguishable on the basis of effects when their energy is absorbed in matter.

The total effect of X-ray or gamma-ray irradiation of matter, in the almost immediate time interval, is the production of high-energy electrons of energy related to that of the incident ray. Such electrons behave like beta rays (electrons emitted from atomic nuclei) or electrons from a machine source of the same energy. They lose energy by excitation and ionization of atoms and molecules of the systems they traverse. The amount of energy such an electron gives to an atom or molecule tends to exceed that deposited in photochemical processes, and the variety of initial physical (and consequent chemical) effects is more numerous and diverse. The situation is further complicated by the fact that the secondary electrons produced in ionization processes in which the input of energy is high may themselves initiate other ionization and excitation processes that can yield further chemistry, the totality of which is embraced in the title of radiation chemistry (see below Molecular activation and Ionization and chemical change).

THE PASSAGE OF NUCLEAR PARTICLES AND RADIATION

Ionization tracks

Heavy charged particles. Charged particles, such as atomic or molecular ions or molecular fragments, that travel in a material medium deposit energy along their paths, or tracks. If the medium is sufficiently thick, the velocity of the charged particle is reduced to near zero so that its energy is practically totally absorbed and is totally utilized in producing physical, chemical, and, in viable (living) matter, biological changes. If the sample is sufficiently thin, the particle may ultimately emerge, but with reduced energy.

Linear energy transfer.. The stopping power of a medium toward a charged particle refers to the energy loss of the particle per unit path length in the medium. It is specified by the differential $—dE/dx$, in which $—dE$ represents the energy loss and dx represents the increment of path length. For extremely high-speed particles (velocities 99 percent the speed of light or greater), a part of the deposited energy is carried away from the vicinity of the track in the time scale for electronic transition (about 10^{-15} second) in the form of energetic secondary electrons (delta rays) and radiation (bremsstrahlung). The radiation chemist is interested in the distance in the track in which a specific amount of energy is deposited. In approximate terms, it is customary to refer to the LET (or linear energy transfer), the energy actually deposited per unit distance along the track (*i.e.*, $—dE/dx$). For not-so-fast particles, stopping power and LET are numerically equal; this situation covers all heavy particles studied so far in chemistry and biology, but not electrons. In a refined study and redefinition of linear energy transfer or restricted linear collision stopping power, a quantity symbolized by the letter L with subscript Greek letter delta, L_Δ, is defined as equal to the fractional energy lost ($—dE$) per unit distance traversed along the track (dl), or $L_\Delta = -(dE/dl)_\Delta$, in which the subscript delta (A) indicates that only collisions with energy transfer less than an amount Δ are included (the quantity L_Δ is expressed in any convenient units of energy per unit length). The symbol L_{100} would indicate—as nearly as theory is useful or helpful—that only the energy actually deposited in the track (*e.g.*, in the spurs or regions of high ionic and excitation density) is of interest. Energy deposited in "blobs" or "short tracks," to the side of the main track, as described in the Mozumder–Magee theory of track effects (after A. Mozumder, an Indian physicist, and J.L. Magee, a United States chemist), is purposefully not included in the definition of L,,,,.

Stopping *power*. By use of classical mechanics, Bohr developed an equation of stopping power, $—dE/dx$, given as the product of a kinematic factor and a stopping number.

The kinematic factor includes such terms as the electronic charge and mass, the number of atoms per cubic centimetre of the medium, and the velocity of the incident charged particle. The stopping number includes the atomic number and the natural logarithm of a term that includes the velocity of the incident particle as well as its charge, a typical transition energy in the system (cf. Figure 1; a crude estimate is adequate because the quantity appears within the logarithm), and Planck's constant, h. Bohr's stopping-power formula does not require knowledge of the details of atomic binding.

Bethe's stopping number

For a heavy incident charged particle in the nonrelativistic range (*e.g.*, an alpha particle, a helium nucleus with two positive charges), the stopping number B, according to Hans Bethe, a physicist in the United States, is given by quantum mechanics as equal to the atomic number (Z) of the absorbing medium times the natural logarithm (ln) of two times the electronic mass times the velocity squared of the particle, divided by a mean excitation potential (I) of the atom; *i.e.*, $B = Z \ln (2mv^2/I)$.

Bethe's stopping number for a heavy particle may be modified by including corrections for particle speed in the relativistic range ($\beta^2 + \ln [1-\beta^2]$), in which the Greek letter beta, β, represents the velocity of the particle divided by the velocity of light, and polarization screening (*i.e.*, reduction of interaction force by intervening charges, represented by the symbol $\delta/2$), as well as an atomic-shell correction (represented by the ratio of a constant C to the atomic number of the medium); that is, $B = Z (\text{In } 2mv^2/I - \beta^2 - \ln [1 - \beta^2] - C/Z - \delta/2)$.

The most important nontrivial quantity in the equation for stopping number is the mean excitation potential, I. Experimental values of this parameter, or quantity, are known for most atoms, but no single theory gives it over the whole range of atomic numbers because the calculation would require knowledge of the ground states and all excited 'states. Statistical models of the atom, however, come close to providing a theory. Calculations by the United States physicist Felix Bloch in **1933** showed that the mean excitation potential in electron volts is about 14 times the atomic number of the element through which the charged particle is passing ($I = 14Z$). A later calculation gives the ratio of the potential to atomic number as equal to a constant (a) plus another constant (b) times the atomic number raised to the $-\frac{2}{3}$ power in which a = 9.2 and b = 4.5; *i.e.*, $I/Z = a + bZ^{-2/3}$. This formula is widely applicable. Other exact quantum-mechanical calculations for hydrogen give its mean excitation potential as equal to 15 electron volts.

Bragg's rule

Even though the basic stopping-power theory has been developed for atoms, it is readily applied to molecules by virtue of Bragg's rule (after a British physicist, W.H. Bragg), which states that the stopping number of a molecule is the sum of the stopping numbers of all the atoms composing the molecule. For most molecules Rragg's rule applies impressively within a few percent, though hydrogen (H_2) and nitrous oxide (NO) are notable exceptions. The rule implies: (1) similarity of atomic binding in different molecules having one common atom or more and (2) that the vacuum ultraviolet transitions, in which most electronic transitions are concentrated under such irradiation, involve energy losses much higher than the strengths of most chemical bonds.

The charge on a heavy positive ion fluctuates during penetration of a medium. In the beginning, it captures an electron, which it quickly loses. As it slows down, however, the cross section of electron loss decreases relative to that for capture. Basically, the impinging ion undergoes charge-exchange cycles involving a single capture followed by a single loss. Ultimately, an electron is permanently bound when it becomes energetically impossible for the ion to lose it. A second charge-exchange cycle then occurs. This phenomenon continues repeatedly until the velocity of the heavy ion approximates the orbital velocity of the electron in Bohr's theory of the atom, when the ion spends part of its time as singly charged and another part as a neutral atom. On further slowing down, the electronic-energy-loss mechanism becomes ineffective, and energy loss by nuclear collision dominates.

Range. The total path length traversed by a charged particle before it is stopped is called its range. Range is considered to be taken as the sum of the distance traversed over the crooked path (track), whereas the net projection measured along the initial direction of motion is known as the penetration. The difference between range and penetration distances results from scattering encountered by the particle along its path.

Computation of particle ranges

Particle ranges may be obtained by (numerical) integration of a suitable stopping-power formula. Experimentally, range is more easily measured than is stopping power. For heavy particles a critical incident energy in low-atomic-number mediums is 1,000,000 electron volts divided by the mass of the particle in atomic mass units (numerically the same as its atomic weight). For incident energies higher than this critical value, range is usually well-known, and computation agrees with experiment within about 5 percent. In the case of aluminum, which is the best studied material, the accuracy is within about 0.5 percent. For incident energies less than the critical value, however, range calculations are usually uncertain, and agreement with experiment is poor. The range–energy relation is often given adequately as a power law, that range (R) is proportional to energy *(E)* raised to some power (n); that is, $R \propto E^n$. Protons in the energy interval of a few hundred MeV conform to this kind of relation quite well with the exponent *n* equal to 1.75. Similar situations exist for other heavy particles. Measurements of range and stopping power are of great importance in particle identification and measurement of their energies. Many experimental data and computations are available for ranges of heavy particles as well as of electrons. The theory by which Bethe derived a stopping number is generally accepted as providing the framework for understanding the variation of range with energy, though in practice the mean excitation potential, I, must be obtained in many cases by experimental curve fitting.

Both stopping power and range should be understood as mean (or average) values over an ensemble of atoms or molecules, because energy loss is a statistical phenomenon. Fluctuations are to be expected. In general, these fluctuations are called straggling, and there are several kinds. Most important among them is the range straggling, which suggests that, for statistical reasons, particles in the same medium have varying path lengths between the same initial and final energies. Bohr showed that for long path lengths the range distribution is approximately Gaussian (a type of relationship between number of occurrences and some other variable). For short path lengths, such as those encountered in penetration of thin films, the emergent particles show a kind of energy straggling called Landau type (after Lev Landau, a Soviet physicist). This energy straggling means that the distribution of energy losses is asymmetric when a plot is drawn, with a long tail on the high-energy-loss side. The intermediate case is given by a distribution according to Sergey Ivanovich Vavilov, a physicist in the Soviet Union, that must be evaluated numerically. There are evidences in support of all three distributions in their respective regions of validity.

Braggpeak

The ionization density (number of ions per unit of path length) produced by a fast charged particle along its track increases as the particle slows down. It eventually reaches a maximum called the Bragg peak close to the end of its trajectory. After that, the ionization density dwindles quickly to insignificance. In fact, the ionization density follows closely the linear energy transfer (LET). With slowing, the LET at first continues to increase because of the strong velocity denominator in the kinematic factor of the stopping-power formula. At low speeds, however, LET goes through a maximum because of: (1) progressive lowering of charge by electron capture and (2) the effect of the logarithmic term in the stopping-power formula. In general, the maximum occurs at a few times the Bohr orbital velocity. A curve of ionization density (also called specific ionization or number of ion pairs—negative electron and associated positive ion—formed per unit path length) versus distance in a given medium is called a Bragg curve. The Bragg curve includes straggling within a beam of particles; thus, it differs somewhat from the specific ionization curve for an individual particle in that it has a long tail of low ionization density beyond the mean range. The mean range of radium-C′ alpha particles in air at NTP (normal temperature and pressure), for example, is 7.1 centimetres; the Bragg peak occurs at about **6.3** centimetres from the source with a specific ionization of about 60,000 ion pairs per centimetre.

Electron stopping number

Electrons. In the first Born approximation, inelastic cross section depends only on velocity and the magnitude of the charge on the incident particle. Hence, an electron and a positron at the same velocity should have identical stopping powers, which should be the same as that of a proton (a nuclear particle having about 1,850 times the mass of a positron, but the same charge) at that velocity. In practice there is some difference in the case of an electron because of the indistinguishability of the incident and atomic electrons. In describing an ionization caused by an incident electron, the more energetic of the two emergent electrons is called, by convention, the primary. Thus, maximum energy loss (ignoring atomic binding) is half the incident energy. Incorporating this effect, the stopping number of an electron is given by a complicated expression that involves a different arrangement of the parameters found in the stopping number of heavy charged particles; *i.e.*,

$$B_e = \frac{Z}{2}\Big[\ln mc^2\beta^2E/2I^2(1-\beta^2) - (2\sqrt{1-\beta^2} - 1 + \beta^2)$$
$$\ln 2 + (1-\beta^2) + \frac{1}{8}(1-\sqrt{1-\beta^2}) - 2\frac{C}{Z} - \delta\Big].$$

This stopping-power formula has a wide range of validity, from approximately a few hundred electron volts to a few million electron volts in materials of low atomic number. For low velocities, the Born approximation gradually breaks down, and highly excited states begin to be inaccessible to transitions by virtue of small maximum energy transfer. Yet, with some corrections the electron-stopping-power formula may be extended down to about 50 electron volts. Below that value any stopping-power formula is of doubtful validity, even though it is certain that most of the energy is still being lost to electronic states down to a few electron volts of energy.

On the high-velocity side, relativistic effects increase electron-stopping power from about 1,000,000 electron volts upward. Except for the term δ (Greek letter delta), attributable to polarization screening, the relativistic stopping power tends to infinity as the electron velocity approaches the speed of light $(v/c = \beta \rightarrow 1)$. One-half of the stopping power, called the restricted stopping power, is numerically equal to the linear energy transfer and changes smoothly to a constant value, called the Fermi plateau, as the ratio β approaches unity. The other half, called the unrestricted stopping power, increases without limit, but its effect at extreme relativistic velocities becomes small compared with energy loss by nuclear encounters.

At extremely high velocities an electron loses a substantial part of its energy by radiative nuclear encounter. Lost energy is carried by energetic X-rays (*i.e.*, bremsstrahlung). The ratio of energy loss by nuclear radiative encounter to collisional energy loss (excitation and ionization) is given approximately by the incident electron energy (E) in units of *1,000,000* electron volts times atomic number (Z) divided by *800; i.e.*, $EZ/800$. For a large class of mediums (atomic number equal to or greater than 8; *i.e.*, that for oxygen) the electron stopping is dominated by bremsstrahlung radiation for energies greater than *100* MeV.

Cherenkov radiation

Cherenkov radiation. When the speed of a charged particle in a transparent medium (air, water, plastics) is so high that it is actually greater than the group velocity of light in that medium, then a part of the energy is emitted as Cherenkov (Cerenkov) radiation, first observed in 1934 by Pavel Alekseyevich Cherenkov, a physicist in the Soviet Union. Such radiation rarely accounts for more than a few percent of the total energy loss. Even so, it is invaluable for purposes of monitoring and spectroscopy. Cherenkov radiation is spread over the entire visible region and into the near ultraviolet and the near infrared. The direction of its propagation is confined within a cone, the axis of which is the direction of electron motion.

Energy-transfer mechanism. At the low-velocity end of its path, an electron continues to excite electronic levels of atoms or molecules until its kinetic energy falls below the lowest (electronically) excited state (see Figure 1). After that it loses energy mainly by exciting vibrations in a molecule. Such a mechanism proceeds through the intermediary of temporary negative ion states, for direct momentum-transfer collisions are very inefficient. In a condensed medium (liquid, solid, or glass) very low-energy (less than one electron volt) electrons continue to lose energy by a process called phonon emission and by interaction with other low-frequency intermolecular motions of the medium.

An electron and a singly charged heavy particle with the same velocity have about equal stopping powers. Because of the smali mass of the electron, however, the relative retardation (*i.e.*, decrease in velocity per unit path length) is much more for it. This larger retardation for an electron means that, if an electron and a heavy particle start with the same velocity, the electron will have a much smaller range. Electron tracks show much more straggling and scattering compared with that of a heavy particle. The first effect results from the fact that the electron can lose a large fraction of its energy in a single encounter; the second is the result of small mass. A power law may be used to connect range and energy of electrons in a given medium—*i.e.*, the range is proportional to energy raised to a power n; as in the case of a heavy particle, the index n is slightly less than two at high energies. At low energies the relationship is such that the exponent is one or less. Many formulas and tables are available for stopping powers and for ranges of electrons as well as of heavy particles over a wide range of energies.

Electron and heavy-particle straggling

Neutrons. A neutron is an uncharged particle with the same spin as an electron and with mass slightly greater than a proton mass. In free space it decays into a proton, an electron, and an antineutrino (a fundamental particle of radioactivity), and has a half-life (the time required for half of a large number of neutrons to decay) of about *12–13* minutes, which is so large compared with lifetimes of interactions with nuclei that the particle disappears predominantly by such interactions.

Neutron beams may be produced in a variety of ways. A modern method is to extract a high-intensity beam from a nuclear reactor. A simpler, but expensive, device is one that employs a mixture of radium and beryllium. The reaction of the alpha (a) particles emitted by the radium with beryllium nuclei produces a copious output of neutrons. The neutron is a major nuclear constituent and is responsible for nuclear binding. A free neutron interacts with nuclei in every conceivable way, more or less depending on its velocity and the nature of the target. Ordinary interactions include scattering (elastic and inelastic), absorption, and capture by nuclei to produce new elements. Pure absorption does not result in a new element, even though it is sometimes accompanied by emission of gamma rays. In certain cases of capture, radioactivity follows, often with production of beta (β) particles. In another class of interaction, a heavy charged particle is ejected (such as an α-particle or proton); the resultant nucleus is often but not always radioactive. As an example, the reaction of neutrons on boron to produce alpha particles provides the basis for alpha-particle welding. The principle of such welding, invented by the Soviet chemist V.I. Goldansky, is to deposit a thin layer of a boron (or lithium) compound in the interface between diverse materials, which is thereafter irradiated with neutrons. The high-energy a-particles produced from the nuclear reaction weld the materials together.

Neutron diffraction

Extraordinary interactions of the neutron are represented by diffraction, nuclear fission, and nuclear fusion. Diffraction, exhibited by low-energy neutrons (approximately equal to or less than *0.05* electron volt), demonstrates their wave nature and is consistent with de Broglie's hypothesis. Neutron diffraction complements X-ray technique in locating the positions of atoms in molecules and crystals, especially atoms of low atomic number, such as hydrogen. Fission is the breakup of a heavy nucleus (either spontaneously or under the impact, for example, of a neutron) into two smaller ones with liberation of energy and neutrons. Spontaneous-fission rates and cross sections of fission induced by agencies other than the neutron are so small that in most applications only neutron-induced fission is important. Also, the neutron-induced-fission cross section depends on the particular isotope (species of an element with the same atomic number and similar chemical behaviour but different atomic mass) involved and the neutron energy. The fission process itself generates fast neutrons, which, when suitably slowed down in a process called moderation, are again ready to induce more fission. The ratio of neutrons produced to neutrons absorbed is called the reproduction factor. When that factor exceeds unity, a chain reaction may be started, which is the basis of nuclear-power reactors and other fission devices. The chain is terminated by a combination of adventitious absorption, leakage, and other reactions that do not regenerate a neutron. At the power level at which a reactor operates, the loss rate always balances the generation rate through fission.

Unlike the electron, a neutron loses energy significantly through elastic collisions, because its mass is comparable to.masses of atoms of low atomic number. (According to the laws of mechanics, in elastic collision, on the average, an object loses half its energy to another object of equal mass.)

The average fraction of energy transferred from a neutron per collision, symbolized by $(\Delta E/E)_{av}$, is twice the atomic mass number (A) of the struck atom divided by the square of the mass number plus one; *i.e.*,

$$(-\Delta E/E)_{av} = 2A/(A + 1)^2.$$

Thus, only 18, *25, 42, 90,* and *114* collisions are required to thermalize (reduce the energy of motion to that of the surrounding atoms) a fast neutron in hydrogen, deuterium, helium, beryllium, and carbon, respectively. The effectiveness of a moderator is not determined by moderation considerations alone; the other factor is absorption.

A United States physicist, Eugene Wigner, in the course of consideration of the possible effects of fast neutrons, suggested in 1942 that the process of energy transfer by collision from neutron to atom might result in important

The "knock on" process

physical and chemical changes. The phenomenon, known as the Wigner effect and sometimes as a "knock on" process, was actually discovered in 1943 by Milton Burton and T.J. Neubert, United States chemists, and found to have profound influences on graphite and other materials.

III. Secondary effects of radiation

PURELY PHYSICAL EFFECTS

With respect to radiation effects the terms primary and secondary are used in a relative sense; the usage depends on the situation under study. Thus, ionization and excitation may be considered as primary with respect to some physical and chemical effects. For other chemical effects, production of free radicals (molecular fragments) may be considered as primary even though that process requires a much longer time for its accomplishment. Still longer times are involved in biological processes, in which the end product of an earlier chemical reaction may be considered as primary.

Generally, an atomic solid exhibits little or no permanent chemical change upon irradiation. Important among the atomic solids are such materials as metals and graphite. Production of molecular carbon (C_2) or bigger clusters upon irradiation of carbon and graphite may, in a certain marginal sense, be considered a chemical change. Ionization of a condensed atomic medium followed by recombination regenerates the same atom, but its locale may be affected. For a molecular medium the situation is quite different. Excited electronic states are often dissociative for a molecule and yield chemically reactive radicals. Positive ions, similarly produced, can experience a variety of reactions even before neutralization occurs. Such an ion may fragment all by itself, or it may react with a neutral molecule in what is called an ion–molecule reaction. In either case new chemical species are created. These transformed ions and radicals, as well as the electrons, parent ions, and excited states, are capable of reacting with themselves and with molecules of the medium, as well as with a solute (a dissolved substance) that may be present in homogeneous distribution. The end products of the reactions can be, on the one hand, new stable compounds or, on the other, regenerated molecules of the original species, as in the case of water irradiation.

Physical changes caused by irradiation

A variety of purely physical effects have been observed in different substances under irradiation. They may be broadly classified as: (1) structural change in the crystal, sometimes accompanied by change in the structural dimensions; (2) change in static mechanical properties, such as elasticity and hardness; (3) change in dynamic mechanical properties, such as internal friction and strain; and (4) changes in transport properties, such as heat conductivity and electrical resistivity. Such changes are considered in the section below on Tertiary effectsof radiation.

MOLECULAR ACTIVATION

A molecule is considered activated when it absorbs energy by interaction with radiation. In this energy-rich state it may undergo a variety of unusual chemical reactions that are normally not available to it in thermal equilibrium. Of especial importance is electronic activation; *i.e.*, production of an electronically excited state of the molecule (see Figure 1). This state can be reached (1) by direct excitation by photon absorption; (2) by impact of charged particles, either directly or indirectly through charge neutralization, or by excitation transfer from excited positive ions; and (3) by charge transfer in collision with (relatively) slow incident positive ions. Among the variety of ensuing processes is light emission, or luminescence.

Luminescence. The language of luminescence is clouded by history. Originally, fast luminescence was called fluorescence and slow (*i.e.*, delayed or protracted) luminescence was called phosphorescence. Present scientific practice is to define the terms on the basis of so-called quantum-mechanical selection rules: fluorescence is an allowed transition (*e.g.*, singlet–singlet) and occurs in a typical time of about 10^{-9} second; phosphorescence is a forbidden transition (*e.g.*, triplet–singlet) and may require 10^{-6} second or longer.

In the gas phase (gaseous state), an excited molecule either luminesces, undergoes a process called internal conversion, or undergoes dissociation. Luminescence is the rule for anthracene, whereas for water it is dissociation into hydrogen (H) and hydroxide (OH). As a rule, luminescence processes occur by default—*i.e.*, only if dissociation is energetically impossible or involves a complicated energy-transfer process or if internal conversion to a nonluminescing state is inefficient.

Fluorescence

Fluorescence usually takes place from the lowest electronically excited state (see Figure 1); if higher states are excited they either dissociate or energetically cascade to the lowest excited state by one of several possible internal transition mechanisms before emission occurs. (A notable exception to this rule is afforded by azulene.)

A similar situation exists for triplet excited molecules. The rate of emission, however, is even slower, for in this case it is forbidden by selection rules. If the triplet excitation energy is insufficient for molecular-bond breakage (dissociation), the molecule may remain in a metastable state (*i.e.*, a state of apparent, not real, stability) for a long time until it either phosphoresces, undergoes internal conversion, or combines with other triplets. Such a combination produces a highly excited state, which has enough energy for dissociation. Some of the latter excited states are formed as singlets capable of light emission. This discussion relates to the more common, general features. There are also special cases, not discussed, that do not follow the general pattern.

Ionization phenomena. Ionization (see Figure 1) is that extreme form of excitation in which an electron is ejected leaving behind a positive molecular ion. The minimum energy required for this process is called the ionization potential (IP). The actual energetics are described by the Franck–Condon principle, which simply recognizes that, during the extremely short time of an electronic transition, the nuclear configuration of a molecule experiences no significant change. As a consequence of this principle, in an optical process the ion is almost invariably formed in some kind of excited state by input of energy greater than the IP. Also, because of Franck–Condon restrictions, excitation of an inner electron may result in initial production of nonionized, superexcited molecules (suggested by R.L. Platzman, a United States physicist) with energy exceeding the ionization potential. A superexcited molecule is short-lived and usually converts rapidly (in a time as short as 10^{-14} second) either to neutral products or to an ion plus a free electron with marked excess energy. The ion itself may fragment to give other species with excess kinetic or internal vibrational and rotational energy.

Excitation states. All the various kinds of excitation that occur in the gas phase may occur also in the condensed states of matter (liquid, glass, or solid), but their relative contributions may be affected. In addition, special activated states are produced for which there is no analogue in the gaseous state. They owe their existence to the collective behaviour of atoms and molecules in close proximity. The more important of them are the exciton state, the polaron state, the charge-transfer, or charge-separated, state, and the plasmon state.

Exciton and polaron states

The exciton state is a cooperative state of molecules in which the excitation energy belongs simultaneously to all.

In a polaron state, an electron belongs to the association of molecules, but its motion is relatively slow so that it carries with it its own polarization field, which is described as "a cloud of virtual phonons." A solvated electron (*i.e.*, an electron associated with a particular molecule or group of molecules) is an example of this.

The charge-transfer, or charge-separated, state is an excited state. In a certain sense, electronic excitation involves motion of an electron from a lower orbit to a higher one. Quantum mechanics notes that the electron does not revolve around an atomic nucleus in a precise classical orbit but rather that it occupies an orbital in which it is to be found with maximum probability in the location of the classical orbit. When a molecule in a condensed system is excited, the resulting electronic orbital may overlay one or more adjacent molecules, and, in that sense, the electron belongs to the group because its excitation level does not correspond to the electronic properties of a single, isolated molecule.

The plasmon state is a highly delocalized state formed collectively through coulombic (electrostatic) interaction of weakly bound electrons. Energy losses, approximating 10–20 electron volts in most materials, resulting from formation of plasmon states are seen in the impact of electrons of a few tens of kilovolts' energy on thin films. Both metals and nonmetals, including plastics, show plasma energy losses. The lost energy may reappear in the form of ultraviolet or visible radiation (Ferrell radiation, 1960); no chemical effect is known to have occurred from such losses.

Energy transfer. In general, a small, simple molecule luminesces in the ultraviolet, and a more complex one emits near the blue-violet end of the visible spectrum. Dye molecules, on the other hand, may emit throughout the visible region, including the red end. The ground electronic state of most molecules is a singlet state. Usually, therefore, the optically allowed (*i.e.*, permitted) emission, or fluorescence, is from the lowest excited singlet state to the ground state. The lowest triplet state of the molecule lies somewhat below the excited singlet. Light emission from this triplet state is forbidden by the quantum-mechanical selection rules, but it does occur by default when other processes are even less probable. Such emission is called phosphorescence. It is relatively weak, slow, and shifted toward longer wavelength. Triplet states may be produced from higher singlets by processes called internal conversion and intersystem crossing. The states may also be produced in excitation from the ground state by impact of relatively slow charged particles, such as electrons.

Much of the effect of optical radiation in a condensed system is not on the molecule in which the energy is initially absorbed but on a more remote molecule to which the energy is transferred in a variety of possible processes. They include excitation transfer either directly between adjacent molecules, by a direct quantum-mechanical interaction of an excited molecule with a remote one at a distance of 40 angstroms (*i.e.*, 4×10^{-7} centimetre) or less, or by the so-called trivial process of fluorescence emission from one molecule and reabsorption by one at any distance. These processes are studied mostly in regard to fluorescence and phosphorescence phenomena.

Production and reactions of ions in solution

With high-energy radiation (such as that of electrons, X-rays, and gamma rays), an additional mechanism involving ions is also available. In the case of a solute M in a solvent S, for example, a simplified description of some possible effects of radiation is represented by the following expressions, in which the symbol $+$ is read, "—is acted upon by high-energy radiation, to give—" and *e* represents an ejected electron:

$$S \rightsquigarrow S^{+} + e \qquad \text{(ionization of solvent)}$$

$$e + M \longrightarrow M^{-} \qquad \text{(electron attachment to solute)}$$

$$M^{-} + S^{+} \longrightarrow \begin{cases} M^{*} + S \\ M + S^{*} \end{cases} \qquad \text{(neutralization)}$$

$$M^{*} \longrightarrow M + h\nu \qquad \text{(light emission)}$$

or

$$M \longrightarrow \mathrm{A} + B \qquad \text{(product formation)}$$

Any actual process is considerably more complicated and involves a larger number of species.

Photographic process. One of the most important effects of radiation on matter is seen in photographic action. Apart from its various uses in art, commerce, and industry, photography is an invaluable scientific tool. It is used extensively in spectroscopy, in photometry, and in X-ray examinations. Also, photographic emulsion techniques have been widely used in the detection and characterization of high-energy charged particles. It is important to note that all speculation regarding the primary phenomena involves the notion that, in an energy absorption process, either direct or sensitized, a chloride (or other halide) ion in a silver halide lattice loses an electron. That electron is thereafter captured by a silver ion located at such a point in the lattice that under suitable conditions of exposure and development a silver grain grows to a size representative of the duration and intensity of the light exposure.

IONIZATION AND CHEMICAL CHANGE

Earlier in this section, the ionization phenomenon was briefly discussed as a special case of molecular activation. The ionization process, however, does have certain characteristic features. Most notably, the probabilities (or cross sections) for ionization by light (*i.e.*, photo-ionization) and for ionization by charged-particle impact are different in magnitude and in lowest energy of occurrence (*i.e.*, threshold behaviour) for the same atom or molecule. The photo-ionization cross section shows abrupt onset (*i.e.*, a step behaviour) to a high value at threshold, falling thereafter only gradually with increase of photon energy. Electron-impact ionization in simple atoms (such as hydrogen and helium) begins at the ionization potential, increases in direct proportion to the energy near the threshold, and shows a peak at about 100–200 electron volts' incident energy. With molecules the behaviour is similar except that the peak is broad and much less pronounced. When the incident energy is high and the ejected electron has kinetic energy (energy of motion) largely in excess of its binding energy, the cross section for the process approaches a limit called the classical Rutherford value, after the British physicist Ernest Rutherford.

Electron-impact ionization

In general, the initial processes resulting from the action of high-energy radiation on matter involve the intermediate production and participation of positive ions (both stable and unstable), electrons, negative ions, excited species, and free radicals and atoms, which in turn may enter into the processes of classical reaction kinetics.

Ordinary low-energy (or optical) processes usually involve only excited species and free radicals and atoms.

The important feature that characterizes the chemistry both of optical processes (photochemistry) and of high-energy radiation (radiation chemistry) is that they are conveniently employed and their kinetics studied at room temperature and lower.

Grotthus–Draper and Stark–Einstein laws

Photochemistry. There are two "laws" of photochemistry. The first, the Grotthus–Draper law (for the German chemist C.J.D.T. von Grotthus and the United States chemist J.W. Draper), is simply: for light to produce an effect upon matter it must be absorbed. The second, or Stark–Einstein law (for the German physicist J. Stark and Albert Einstein), in its most modern form is: one resultant primary physical or chemical act occurs per photon (*i.e.*, quantum of light) absorbed. The quantum yield of a particular species of product is the number of moles of that product divided by the number of einsteins of light (*i.e.*, units of 6.02×10^{23} photons)---or the number of molecules of product per photon—absorbed. In the ideal case the quantum yield, frequently denoted by the Greek letters gamma, y, or phi, Φ, is unity. In real cases, Φ may approach zero on the one hand—particularly if a back reaction is involved---or it may be of the order of 1,000,000, in which case the primary product may start a chain reaction, as in a clean, dry mixture of hydrogen (H) and chlorine (Cl). In the following chemical equations each symbol for an element stands for one atom, and the number of atoms bonded into a molecule is given as subscripts following the symbol, while the number of molecules precedes the formula; the arrow indicates the course of the reaction:

$$\mathrm{Cl_2} \xrightarrow{h\nu} \mathrm{2Cl} \qquad (1)$$

$$\mathrm{Cl} + \mathrm{H_2} \longrightarrow \mathrm{HCl} + \mathrm{H} \qquad (2)$$

$$\mathrm{H} + \mathrm{Cl_2} \longrightarrow \mathrm{HCl} + \mathrm{Cl} \qquad (3)$$

etc.,

in which reactions 2 and 3 reoccur repeatedly in a chain reaction. The symbol $h\nu$ may be read "when a photon of light frequency, symbolized by the Greek letter nu, ν (which is always stipulated), is absorbed, gives." Because h is Planck's constant of action (approximately 6.6 × 10 ^{27} erg second) and ν is expressed in reciprocal seconds (*i.e.*, second^{-1}), the product $h\nu$ indicates the energy absorbed per photon. Some reactions may give two primary products; *e.g.*,

$$CH_3CHO \xrightarrow{h\nu} CH_3 + HCO \text{ or } CH_4 + CO.$$

In that case there are different quantum yields for each of the primary reactions, and the ratio of those yields varies with the frequency, ν (or the wavelength, A, or the wave number, $\bar{\nu}$), of the light absorbed.

Radiation chemistry. Although radiation chemistry embraces the effects of the whole range of ionizing particles and radiations, they really cannot be considered as a simple unit.

Nature of positive-ion tracks

When the initial bombardment is by a positive ion such as the hydrogen ion H^+ or the deuterium ion D^+ from a cyclotron, or the alpha particle $^4He^{2+}$ from nuclear decay, or indeed any high-energy heavy positive ion, the initial effects differ significantly from those of a high-energy electron. This situation results from the fact that, for the same kinetic energy, $\frac{1}{2}mv^2$, a particle of greater mass, m, travels with smaller velocity, v. The smaller the velocity of a particle of a particular charge, the greater is its probability of interaction with the medium traversed; *i.e.*, the greater is the linear energy transfer. Thus, positive ions produce their initial effects close together in the ionization track in a condensed medium such as water (perhaps one or two angstroms, 1 or 2 × 10^{-8} centimetre, apart), whereas equally energetic electrons travelling through the same medium deposit energy in small collections called spurs, which may be 1,000 angstroms (10-5 centimetre) or so apart. The appearance of the excitation and ionization track has been likened to a rope (in the case of positive-ion bombardment), on the one hand, as compared with isolated beads on a string (in tbe case of electron bombardment), on the other. The dense track, as well as the isolated spurs, contains ions, excited molecules, and electrons; but the distributions in the two essentially different types of tracks are so different that the ensuing chemical reactions (*i.e.*, the track effects) may be quite dissimilar. As an example, alpha-particle irradiation of pure water produces substantial yields of hydrogen and hydrogen peroxide (H_2O_2), whereas irradiation with beta particles, X-rays, or gammas is essentially without effect. One of the reaction sequences suggested in overall considerations of the radiation chemistry of water is

$$H_2O \rightsquigarrow H + OH \qquad (1)$$

$$2H \longrightarrow H_2 \qquad (2)$$

$$2OH \longrightarrow H_2O_2;\ \textit{i.e.}.\ HOOH, \qquad (3)$$

in which reaction (1) summarizes the early chemical consequences both of ionization and of excitation. It has been suggested that reactions (2) and (3) occur with high probability in dense tracks (*e.g.*, of alpha particles), but that, in isolated spurs (as in fast-particle tracks), such reactions may occur only with low probability. In such a case, the hydrogen atoms and OH radicals enter with somewhat greater probability into back-reaction chains with any $H_2 + H_2O_2$ already produced and existent in the body of the liquid:

$$H + HOOH \longrightarrow H_2O + OH \qquad (4)$$

$$OH + H_2 \longrightarrow H_2O + H \qquad (5)$$

The H atom produced in reaction (5) thereupon enters into reaction (4), so that whatever small amounts of H_2 and H_2O_2 are actually produced in reactions (2) and (3) are consumed in reactions (5) and (4), respectively, and remain essentially undetectable no matter how long the reaction is run.

Mechanism of radiation chemical reactions

Radiation chemical reactions. In more detailed discussions of the mechanism of radiation chemical reactions, the roles of both excitation and ionization are considered. Information regarding the former is available from the extensive data of photochemistry; frequently, the initial excitation process leads to no significant chemical effect. By contrast, ionization may result in a large variety of chemical changes involving tbe positive ion, the outgoing electron, and the excited states resultant from charge neutralization, as well as (parent) positive-ion fragmentation and ion-molecule reactions. Some such consequences are summarized for a few cases.

Different channels of fragmentation from the same parent ion (*e.g.*, the propane ion $C_3H_8^+$), such as

$$C_3H_8^+ \longrightarrow C_3H_7^+ + H \qquad \text{atom elimination}$$

$$\left.\begin{array}{l} C_3H_8^+ \longrightarrow C_2H_5^+ + CH_3 \\ C_3H_8^+ \longrightarrow C_2H_5 + CH_3^+ \end{array}\right\} \text{radical elimination}$$

$$C_3H_8^+ \longrightarrow C_3H_6^+ + H_2 \qquad \text{molecule elimination}$$

compete unless barred by energetic considerations. Because ionization potentials of various possible fragments may differ greatly, charge localization may occur on only one of them. On the other hand. because the initial ionization rarely leads to the ground state of the positive ion, the energy is usually adequate for bond breakage.

Ion–molecule reactions such as that between a water ion and a molecule,

$$H_2O^+ + H_2O \longrightarrow H_3O^+ + OH$$

are more important in the condensed phase, and fragmentation is more important in the gas phase. The parent ion in liquid water almost invariably undergoes ion–molecule reaction as indicated above. Many ion–molecule reactions have high cross sections. The same ion may undergo fragmentation or ion–molecule reaction, depending on circumstances. Thus, methane (CH_4), acted upon by high-energy gamma radiation, producing an electron, symbolized by

$$CH, \rightsquigarrow CH_4^+ + e^-,$$

may be followed by fragmentation

$$CH_4^+ \longrightarrow CH_3^+ + H \text{ or } CH_2^+ + H_2,$$

as well as an ion–molecule reaction

CH_4^+ t CH, [redacted] $CH_5^+ + CH_3$

Electrons in chemical transformations

The electron ejected in an initial ionization process may further ionize and excite other molecules in its path, thus causing other chemical transformations. Additionally, it may produce chemical changes of its own by dissociative attachment as in carbon tetrachloride (CCl_4) and nitrous oxide (N_2O)

e- + CCl_4 [redacted] $CCl_3 + Cl^-$

$$e^- + N_2O \longrightarrow N_2 + O$$

and by formation of negative ions of either permanent or virtual (*i.e.*, very short-lived) nature. Many of the negative ions produced in a dissociation process are chemically reactive (H-, O-, etc.) as well. Virtual negative ions are almost invariably in a high vibrational state; *i.e.*, they are vibrationally hot.

The important point to note from this limited discussion of primary physical effects and their consequences in radiation chemistry is that, in general, each such effect is the progenitor of many ionizations and excitations, the distribution of which in space depends on the energy of the particle involved as well as on the system traversed. There is no single resultant primary process corresponding to the result of absorption of a single optical photon and thus no analogue to the concept of quantum yield in photochemistry.

In radiation chemistry, yields are conventionally reported on the purely empirical basis of the number of molecules of a particular kind produced (or destroyed) per 100 electron volts' input of a particular type of radiation. In the radiolysis (radiation-induced decomposition) of cyclohexane, for example, by cobalt-60 gamma radiation or by electrons of about 2,000,000 electron volts' energy, the overall yield of hydrogen per 100 electron volts' input is frequently given as approximately 5.6 or G $(H_2) \simeq 5.6$, in which the symbol G is read as "the 100-electron-volt yield of." Sometimes, a small *g* is used to denote the 100 electron-volt yield of a postulated intermediate, not directly determinable by measurement. Table I summarizes some typical G values.

Table 1: Examples of Yield Studies at Room Temperature

state	system	radiation	yields
Gas	hydrogen and chlorine	alpha (radium)	$G(HCl) \sim 10^6$
	perfluoroethane	gamma (cobalt-60)	$G(CF_4) = 1.6$ $G(c\text{-}C_3F_6) = 0.30$ $G(C_3F_8) = 0.21$ $G(C_4F_{10}) = 0.14$
Liquid	ferrous sulfate and sulfuric acid in aerated water (Fricke solution)	gamma (cobalt-60)	$G(Fe^{3+}) = 15.6$
		alpha (from neutron irradiation of boron)	$G(Fe^{3+}) = 4.2$
		alpha (from neutron irradiation of lithium)	$G(Fe^{3+}) = 5.7$
	cyclohexaue	1.8 MeV electrons	$G(H_2) = 5.6$ $G(C_6H_{10}) = 2.86$ $G[(C_6H_{11})_2] = 1.55$ $G(C_6H_{11} . C_6H_9?) \sim 0.068$
		1.8 MeV or gamma (cobalt-60)	$G(C_6H_{10})/G[(C_6H_{11})_2] = 1.85$
		alpha (polonium-210)	$G(C_6H_{10})/G[(C_6H_{11})_2] = 3.0$
	benzene	1.5 MeV electrons	$G(H_2) = 0.022$ $G(C_2H_2) = 0.017$ $G(C_6H_6 \rightarrow \text{"polymer"}) \sim 0.8$
Solid	ferrocene	1 MeV electrons	$G(H_2) = 0.0101$ $G(C_5 \text{ products}) = 0.0036$ $G(\text{inorganic Fe}) = 0.042$

Source: Chemical and Engineering *News*.

Symbolism of *radiation chemistry.* The symbolism of radiation chemistry differs from that of photochemistry; the chemistry is somewhat more complicated, and the establishment of the variety of initial chemical processes is somewhat more of a chore. For the action of high-energy radiation on water, the variety of early products is typically indicated by the relation

$$H_2O \rightsquigarrow H, OH, H_2, H_2O_2, e^-_{aq}, \text{etc.},$$

in which $\rightsquigarrow$ is read "acted upon by high-energy radiation, gives" and e_{aq}^- is the symbol for the hydrated electron. Particular note is addressed to the species e_{aq}^- (*i.e.*, an electron solvated by water) indicated in the same reaction. For many years there was an awareness in the radiation chemistry of water of the anomalous behaviour of the hydrogen atom, H, as compared with the same atom produced in established chemical processes. The anomaly was resolved on the one hand by J.W. Boag, a British biophysicist, and E.J. Hart, a U.S. chemist, who spectroscopically observed the species e_{aq}^- in the spectral region predicted by R.L. Platzman, and on the other hand by H.A. Schwarz, a U.S. chemist, and Gideon Czapski, an Israeli chemist, who showed the existence of the ionic reducing species with charge of minus unity.

Time scales in radiation chemistry. The time scale characteristic of radiation chemistry ranges from the extraordinarily short time (about 10^{-18} second) required for a fast electron to traverse a molecule to the time (about three hours) required for essential completion of some neutralization processes in very viscous media. In between, there can be a variety of reactions involving intermediate formation and disappearance of the collections of the various species already discussed. The time-scale spread is so great that a pt scale (in which pt is defined as minus the logarithm, to the base 10, of the time t [*i.e.*, $-\log_{10}t$], and t is the time in seconds) is conveniently employed. Actual observances in the long-time-scale region follow fairly well-established chemical practice. The short-time region, on the other hand, presents interesting challenges. The Van de Graaff generator and the linear accelerator both made possible irradiations by electrons and X-rays in the microsecond (10^{-6} second) region, and spectroscopic devices were quickly devised to make observations in that region. Improvements in irradiation technique (with X-rays by Herbert Dreeskamp, a German physicist, and Milton Burton, a U.S. chemist, and with ultraviolet by P.K. Ludwig, a German chemist, and Juan d'Alessio, an Argentine physicist) and in observation techniques in study of luminescence extended precision of observation to 5×10^{-10} second in the work of W.P. Helman, a U.S. chemist. J.K. Thomas, a British chemist, combined use of a fast linac (linear accelerator) with Cherenkov radiation as a marker to extend chemical studies into the same region. Use of the same radiation as light source for spectroscopic observation of the chemistry produced by a travelling electron front (from a linac) enabled J.W. Hunt, a Canadian chemist, to make actual observations in the time range of (2 to 4) $\times 10^{-11}$ second. Table II summarizes an approximate time scale of events.

Chemical studies with linac

Table 2: Approximate Time Scale of Events in a Condensed System

-log *t* (seconds) $\equiv pt$	events	Stage		
18	fast electron traverses molecule	↑ p ys cal		
17	1-MeV proton traverses molecule; time for energy loss to fast secondary electrons			
16	time for energy loss to electronic states (Franck–Condon process)			
14	fast ion–molecule reactions involving hydrogen-atom transfer; molecular vibration; fast dissociation		↑ physicochemical	
12	electron thermalizes; self-diffusion time scales for liquids of low molecular weight			
II	dielectric relaxes in water; neutralization time for polar media	↓		
10 9 8 7 6	spur* formed spur* reactions intratrack reactions completed neutralization times in media of low viscosity and low dielectric constant		↓	↑ chemical
4 3 0 −2 −4	escape time for electrons in media of low viscosity and low dielectric constant radiative lifetime of triplet excited states neutralization times for media of high viscosity and low dielectric constant			↓

*A spur is a region about 2×10^{-7} centimetre in diameter in a liquid and of high ionic and excitation density.
Source: Advances *in* Radiation Chemistry. Wiley–Interscience.

IV. Tertiary effects of radiation

The electrons liberated by high-energy irradiation that have sufficient energy cause further ionizations in which additional electrons are produced. Some of such second-generation electrons also cause additional ionizations, and this process continues until their remaining energy becomes inadequate. Even though this process goes through several generations of events, it actually takes little time, and therefore it appears as an impact phenomenon as far as radiation-induced chemical changes are concerned. For this purpose, then, they may be considered as primary. Fast chemical changes induced by radiation may take time of the order of nanoseconds (a nanosecond is 10^{-9} second) or less to complete; slower reactions involving (relatively) less reactive scavengers (reagents that eliminate residues) in dilute concentrations may need time of about a microsecond ($\frac{1}{1,000,000}$ second).

This section is concerned with radiation effects on much longer time scales, arbitrarily greater than about one minute. Attention is here addressed to physical changes in the solid state, of which there is a wealth of experimental information. It should again be emphasized that little chemical change is expected in an atomic medium in which the absorption of ionizing radiation also results ultimately in structural changes and induced imperfections. With neutron irradiation, except for specific nuclear interactions, one gets "knocked off" atoms or ions (note the discussion of Wigner effect in *Neutrons* above). These ions quickly capture electrons and then travel as neutral atoms. Even though a small effect occurring in ionization and electronic excitation attributable to knocked off ions cannot be denied, it is believed that this effect is small compared with that brought about by the neutral knock offs in the form of structural changes.

HEATING EFFECTS

The simplest ultimate effect of absorption of radiation is heating. It can be argued that, for low-linear-energy-transfer (LET) ionizing radiation, the heating effect is negligible. A spur created by a low-LET radiation is a small spherical region in which the energy deposit is localized in isolation. The temperature rise, AT, of the spur above the surrounding temperature has a space–time dependence that, by hypothesis, has a statistical distribution, called Gaussian, because of random superposition of events leading to the heating process. The temperature rise at a point located a distance r away from the spur centre at time t is given by the equation

Temperature rise in spur

$$\Delta T(r,\ t)=\Delta T_{max}(1+\frac{4\gamma t}{a^2})^{-3/2}\exp\ (-\frac{r^2}{a^2+4\gamma t}),$$

in which a is the initial spur-size parameter, the Greek letter gamma, y, is the thermal diffusivity of the medium (equal to heat conductivity divided by the product of density and specific heat at constant volume). and ΔT_{max} is the initial maximum temperature rise at the centre of the spur. Taking reasonable values for energy deposition (30 electron volts) and spur size a (20 angstroms, or 2 × 10^{-7} centimetre) and using (for water) density equal to one gram per cubic centimetre and specific heat 4 × 10^7 ergs per gram-degree, AT,,,, may be estimated to be 30" ***C*** (54" F). The time required for the central temperature to drop to half its initial value (*i.e.*, $t_{1/2}$) is given by $(1 + 4\gamma t_{1/2}/a^2)^{3/2} = 2$. With the thermal diffusivity, symbolized by the Greek letter gamma, y, equal to 10^{-3} centimetre squared per second for water, $t_{1/2} = 6 \times 10^{-12}$ second. The conclusion is that for low-LETradiation the local temperature rise is too small and drops too fast to have any appreciable chemical (or physical) effect. It is particularly notable that the actual temperature rise is smaller than that estimated here because part of the deposited energy is invariably utilized in ionization, dissociation, and similar processes. This part of the energy resides in the potential form and is not completely available for heating. With high-LETionizing radiations (such as fission fragments, stripped nuclei, α-particles) the situation is somewhat different. In such a case, a large amount of energy is deposited per unit path length, resulting in cylindrical tracks (rather than spherical spurs). The equation for temperature rise in this case is written in a form much like the equation for spur geometry; that is,

$$\Delta T(r,\ t)=\Delta T_{max}(1+\frac{\gamma t}{a^2})^{-1}\ \exp\ (\frac{-r^2}{a^2+4\gamma t}),$$

except that, in this case, a is the initial size parameter for the track cylinder and AT,,, is the maximum initial temperature rise on its axis. For fission fragments with LET of 500 electron volts per angstrom (*i.e.*, 10^{-8} centimetre) and a equal to 20 angstroms, AT,,, for water is 1.6 × 10^{4}° K. Admittedly, this figure is an overestimate for reasons similar to those that apply for low-LETradiations, but it is believed that the temperature rise is high and may approach 10^4° K. The time for this temperature to drop to half the initial value is given by $t_{1/2} = a^2/4\gamma$, which in this case is estimated to be about 10^{-11} second. This time is not much larger than the corresponding time for survival of an isolated spur. Because of the high local temperature, however, the reaction time of the radiation-produced intermediates is also very small. In an intermediate ·+ substrate (*i.e.*, solvent) reaction, for example, with an activation energy of approximately eight kilocalories per mole, the rate constant, k, for such a typical pseudo-first-order reaction may be written. according to the simple usage of chemical kinetics, in the form

$$k \simeq 10^{-11} \exp(-8{,}000/RT)\ \text{cubic centimetre per molecule per second,}$$

in which R is the gas constant in units of two calories per degree and T is the absolute temperature. For a substrate concentration of 10^{22} molecules per cubic centimetre and a temperature of 10^4° K, $-d(\ln \nu)/dt = 6.6 \times 10^{10}$ reciprocal seconds, in which the Greek letter nu, ν, here denotes the concentration of intermediates still existent at time t. Therefore, the time required for the intermediate concentration to drop to an "e-folding" value (*i.e.*, to fraction $1/e$) is approximately 1.5×10^{-11} second, a time that compares favourably with the duration of the temperature pulse. The conclusion is that for very high LET radiation there is indeed a high degree of local heating, and, even though the heat pulse survives only a short time, that time is enough for acceleration of substrate reactions.

Temperature rise to 10,000° K

CRYSTAL-LATTICE EFFECTS

In neutron irradiation of a solid, atoms are dislodged from normal lattice positions and set in motion (the Wigner effect). The fractional amount of energy transfer depends, as in any elastic collision, on the mass ratio of the neutron to that of the recoil atom. Thus, in graphite a carbon atom, on first collision with a 1,000,000-electron-volt neutron (produced, say, in a fission process), receives a kinetic energy of approximately 10^5 electron volts, which is large compared with its binding energy in the lattice (about 10 electron volts). It is estimated that the 1,000,000-electron-volt neutron strikes about 60 carbon atoms before it is thermalized or its speed is so much reduced that it cannot knock off other carbon atoms. Much of the structural damage caused by radiation is attributable to these (relatively heavy) carbon atoms rather than to the original neutron. In this sense radiation damage by fast neutrons may be viewed as an indirect action. Slowing of the fast carbon (or other dislodged) atoms is basically governed by interaction time. This fact means that the stopping is light in the beginning of the journey of a dislodged atom and results only in occasional displacement of atoms. Toward the end of its career **a** large number of atoms are displaced in quick succession along a row, and finally a large amount of residual energy is dumped locally into a relatively small group of atoms. This process generates the displacement (or thermal) spike; the local temperature rise is estimated to be about 10^3°. Even though the temperature rise lasts only about 10^{-11} second before the track is cooled down, this duration is enough for permanent structural damage. **At** least a part of the swelling of graphite under reactor-neutron irradiation is the result of this local heating; another part originates in change in lattice dimension under irradiation.

Cause of structural damage

The high temperature rise in a thermal spike probably results in local melting of the solid. Evidence in that direction has been obtained from the study of copper–zinc alloy under neutron bombardment at low temperature. Before irradiation the structure is ordered; each copper atom lies next to zinc atoms only as nearest neighbours and vice versa. After irradiation, a general random rearrangement of atoms is seen, presumably the result of melting and refreezing.

Long-term effects of radiation on crystals are numerous, and the magnitudes of these effects depend on the crystal structure and previous history. Only some general features of these effects are recounted here.

Long-term effects. 1. Radiation damage may be considered as attributable to the production of Frenkel defects, the result of ejection of a struck atom into an interstitial position with the associated generation of **a** vacancy.

2. In a certain sense radiation-induced damage to the crystal structure is qualitatively similar to that produced

by cold-working (for example, by hammering). Neutron irradiation of pure copper, which is naturally soft at room temperature, makes it so hard that it can be made to sing like a tuning fork.

Neutron effect on pure copper

3. A solid has a tendency to recover spontaneously from radiation damage. If it were not for this property it would indeed be extremely difficult to operate nuclear reactors (which are permitted to heat up periodically to remove the effect in the graphite core). The healing (or so-called annealing) is presumably attributable to the recombination of interstitial atoms and vacancies, thereby removing Frenkel defects. It is not necessary that an interstitial atom always recombine with its corresponding vacancy. Often it may recombine with a vacancy that resembles the one that it left; the result is approximate restoration of the original properties of the crystal. Such annealing is facilitated by the increased mobility of the vacancies and interstitials at higher temperature. At a particular temperature called the annealing temperature, the healing becomes fast and essentially complete. The same substance may have somewhat different annealing temperatures depending on the particular property under study. Many experiments on radiation damage must be done at low temperatures to freeze in the defects produced.

4. So far as radiation sensitivity is concerned, pure metals are the most resistant and the most easily annealed. Annealing temperatures in such cases are relatively low. Thus, the annealing temperature for the increase of electrical resistance in pure copper is only around 40° K. On the other hand, changes in elastic modulus and hardness, such as are required to produce tuning-fork characteristics, persist up to room temperature; *i.e.*, 293" K (20" C or 68° F). Quick annealing in pure metals is directly attributable to the high mobility of atoms in perfectly ordered structures. At the other extreme are organic solids, particularly polymers, composed of large molecules. Here the damage originates in breaking of bonds that usually do not rejoin in the original manner but instead produce chemically different material. Inorganic materials are between metals and organics in their radiation sensitivity. Semiconductors such as silicon and germanium and crystals such as diamond resemble metals in this behaviour.

5. In simple metals irradiation decreases conductivity for both heat and electricity. Conduction of both in metallic crystals is attributable to their ordered structure. The more perfect the structure, the better is the conduction. Frenkel defects, generated by irradiation, therefore, decrease both conductivities. In extreme cases conductivity decrease of orders of magnitude has been observed. With moderate irradiation, however, both thermal and electrical conductivities decrease usually by about half.

Change in conductivities

6. Hardness and ductility depend on perfection of the crystal structure. It is therefore found, as is to be expected, that irradiation results in loss of ductility and increase in hardness. Such effects are attributed to glide-plane obstruction in the crystal. Most structured materials become harder, less ductile, and sometimes more brittle as the result of neutron irradiation. Similarly, most polymers (substances composed of giant molecules) also lose ductility on irradiation.

Properties of irradiated crystals. A few of the properties of irradiated crystals are the following.

1. Expansion. Expansion and lattice distortion produced by neutron irradiation are generally more pronounced than those produced by gamma irradiation. Other general features are: (a) the more disordered the starting crystal, the less is the expansion and the more is the tendency for volume contraction; (b) the expansion is not linear with dose—with continued exposure, a material expands much less for the same dose when it is administered later in its history; and (c) the expansion is anisotropic (*i.e.*, not the same in all directions)—the irradiated material looks more isotropic than the unirradiated control. A perfect crystal of graphite consists of planes of carbon atoms, layer upon layer. Upon neutron irradiation, graphite expands perpendicular to the base plane and contracts slightly parallel to the base plane. After moderate exposure in a nuclear reactor, the expansion is about **1** percent for a flux of 10^{20} neutrons per square centimetre. The actual amount of expansion, of course, depends on the fabrication history and operating temperature of the graphite. Beryllium oxide expands about **1** percent over the first day of usage in a nuclear reactor; the subsequent expansion is much less. Fast-neutron irradiation of natural zircon crystal gives a density loss of about 1 percent per total flux of 10^{20} neutrons per square centimetre. In barium titanate the unirradiated tetragonal crystal has lattice separations that are different in two directions. Upon fast-neutron irradiation, both lattice separations increase, but in differing degrees. Eventually, at about 10^{20} neutrons per square centimetre total flux, the lattice separations become almost equal. Expansion of moderator materials such as graphite is of great importance in the design of nuclear reactors. Even a small percentage change in dimension can result in large total change in the reactor structure; if this change is not allowed for in the engineering design of the reactor, it may create strained operating conditions eventually leading to failure.

Influence on nuclear-reactor design

2. Mechanical properties. Graphite experiences increase in strength and hardness upon irradiation. Annealing is faster at elevated temperatures; also, damage is less when the irradiation is at a higher temperature. A similar effect is seen for the compressive stress–strain curve. Studies of dynamic properties in ceramics indicate a saturation effect at large doses.

3. Conductivity. Thermal conductivity of graphite falls to about half the unirradiated value with an exposure of **3** $\times$ 10^{20} neutrons per square centimetre at room temperature. Like other property changes, this effect can also be annealed at elevated temperatures with concomitant release of stored energy. Energy storage in graphite amounts to about 200 calories per gram per 10^{20} neutrons per square centimetre total flux. Interstitial carbon atoms produced in the irradiation scatter electrons and thus decrease electrical conductivity. The pattern of conductivity decrease and increase depends on the nature of the graphite and the duration of exposure in a reactor. With ceramic materials, loss of thermal conductivity by a factor of about **3** to 5 may be observed under conditions in which the decrease is about one-half in graphite. In mica, on the other hand, the change is somewhat less than in graphite.

BIOLOGICAL EFFECTS

The discussion here is of a cursory nature. For details, the reader is referred to the articles PHOTOSYNTHESIS; RADIATION, BIOLOGICAL EFFECTS OF. Effects of radiation on biological systems are usually investigated from two extremes, viz., with whole-animal experiments and with experiments on individual cells or unicellular organisms, such as bacteria. Closely associated with these investigations are experiments on biological material such as enzymes, amino acids, and DNA (deoxyribonucleic acid).

Action of light on living organisms is a subject that has been studied in the past in great detail. The study still continues because of its obvious importance to photosynthesis and vision. Indeed, it is through photosynthesis that oxygen is replenished in the atmosphere and carbon dioxide is fixed in plants for support of animal life. It has been estimated that the entire oxygen supply of the atmosphere is replenished by plants in a cycle of **2,000** years. The similar cycle time for carbon dioxide fixation is **200** years. Some general biological implications of the effects of light follow.

Oxygen and carbon dioxide replenishment cycles

Effects of light. 1. It has been known over a long period of time that bacteria such as Escherichia coli and Staphylococcus *aureus* absorb light in the near-ultraviolet; *i.e.*, approximately in the region **300–240** nanometres. This absorption results directly in bactericidal action.

2. Ultraviolet also does damage to DNA and interferes with the replication process. Indeed, uninterrupted replication of nucleic acids and proteins would not have been possible on the Earth's surface were it not for the protection afforded by the absorption of ultraviolet by the ozone in the upper atmosphere.

The role of chlorophyll

3. The absorption spectrum of chlorophyll, the chemical substance responsible for the initiation of the photosynthetic process, matches the broad peak of solar radiation observed on Earth in the visible range at 680 nanometres corrected for absorption by ozone and water vapour in the atmosphere. Thus, photosynthesis makes good use of the sunlight available on the surface of the Earth. In this process six water molecules combine with six carbon dioxide molecules utilizing 29 electron volts of energy from the sunlight to produce one molecule of carbohydrate and six molecules of oxygen. This "balance sheet" refers to a series of complex chemical reactions, not all of which are completely understood; many of them, however, are known in detail at present. Chlorophyll-a, present uniformly in all green plants, is the prime material for photosynthesis. Other chlorophyll molecules, such as b, *c*, and *d*, which may be present in algae and also in land plants, absorb energy and transfer to chlorophyll-a before photosynthesis occurs. Similar mechanisms, but utilizing sunlight in a different spectral region, presumably are responsible for photosynthesis in red and blue-green algae.

4. It is apparent that there are many metabolic, behavioral, and physiological processes in plants and animals that are controlled periodically by sunlight. This photoperiodic phenomenon is believed to be regulated by various internal (biological) clocks, which are in turn (probably) governed by one master clock. Examples are well-known in sleep, body temperature, and pathological and physiological functions. Indeed, flowering in seasonal plants is connected with photoperiodicity (flowering of poinsettia at the Christmas season is a good example). Detailed experimentation has shown that it is the night duration rather than the day duration that controls flowering.

Biological effects of ionizing radiations are classified as acute or chronic depending on whether the effect is observable immediately or after a delay. Chronic effects include inherited abnormality and malignancy; in these effects, damage of DNA is probably involved. DNA in the cell is not easily replaced or repaired. Any radiation-induced change in DNA that results in faulty duplication is a potential source of danger of cellular malfunction or of abnormal inheritance or both.

Other radiation effects. Most radiation biological damage results from ionization, though excitation effects can also be significant. Repair of radiation biological damage at an early stage results from the return of a small fragment to a complex molecule, from which it was torn away as a result of the effect of ionizing radiation. Often, that fragment is a hydrogen atom. A relatively heavy fragment recombines more easily. Sometimes it is seen that complete recovery follows acute radiation damage in whole-animal irradiation. It is not that damage has not occurred but simply that the damage has been quickly repaired. Thus, in many cases, there is an apparent threshold dose below which the animal recovers completely. Above the threshold, the recovery is partial or may not occur at all.

A radiation biological effect is primarily cellular in origin. Extracellular material is not immune to radiation; the cellular material is just more sensitive, and its changes have profound influence on the gross biological behaviour of the living organism.

Radiation sensitivity of cells

Radiation sensitivity of a cell depends on a variety of factors, including the conditions of its environment. Cells of higher organisms are generally more radiation sensitive than cells of lower organisms. Thus, such unicellular organisms as bacteria are most resistant to irradiation; it should be realized, however, that bacteria differ greatly among themselves in respect to radiation sensitivity. For the same organism, usually but not always, radiation sensitivity increases with the linear energy transfer of radiation; sometimes a saturation effect may be seen, or, on the other hand, a falloff of radiation sensitivity occurs, at very high LET. If damage requires multiple ionizing events in order to occur (such as a chromosome injury), it is easy to understand why high-LET radiation is more effective in production of damage. If only one ionization is sufficient to bring about damage (as in bacteriophage, a body-waste virus), however, multiple ionization is relatively inefficient; it merely overkills. Radiation sensitivity of a cell depends also on the presence or absence of other molecules, particularly oxygen. Most cells are two or three times as sensitive in the presence of oxygen as in its absence. The cause of the oxygen effect is probably the intermediate production of the free perhydroxyl radical HO_2. Malignant tissues are usually in an anoxic state (*i.e.*, in an oxygen-free environment). To be rendered ineffective they therefore need a greater radiation dose than does similar healthy tissue.

Two kinds of survival curves

Quantitative radiation sensitivity of a cell (mammalian, bacteria, bacteriophage) is expressed by a survival curve in which the surviving fraction is plotted as a function of radiation dose. There are basically two types of survival curves. The first type shows exponential survival; *i.e.*, the type of survival that results if a single hit (by radiation) in a cell is sufficient to produce the biological damage. The second type represents sigmoidal survival; *i.e.*, the survival curve shows a threshold, or a substantial dose is required for perceptible radiation damage. Theoretically, such survival results if multiple hits are required in a cell for its deactivation.

V. Significance *of* radiation-induced changes

BIOLOGICAL PROCESSES

Protophotosynthesis and *photosynthesis.* The most important chemical process to life is botanical photosynthesis; *i.e.*, the process of production of amino acids, carbohydrates, and their derivatives in living plants as the consequence of the absorption of light. In the history of the world, however, there was a time when no living thing existed. Indeed, there was a period in which high temperature itself precluded formation or survival of any of the complicated organic compounds necessary for life. At that time, the atmosphere of the Earth consisted of water (H_2O), ammonia (NH_3), and methane (CH,). It was only as the planet continued to cool that water could condense. Meanwhile, as a result of the absorption of far-ultraviolet light (see Figure 1) and of the action of X-rays and gamma rays and of corpuscular radiations, chemical reactions began to occur in the atmosphere. These processes, when the atmosphere was still quite hot, resulted principally in molecular decomposition and the loss of some (low-molecular-weight) hydrogen from the gravitational field of the Earth. Later, in the cooling process, liquid water began to form, and the rains that thereupon periodically occurred were accompanied by electrical discharges (lightning) and electron bombardment of the components of the atmosphere. The result—as the cooling process continued—was protophotosynthesis of the first compounds (fatty acids, aldehydes, ketones, amino acids, ring compounds, heterocyclics) essential for the existence of even the most primitive forms of life.

Consequences of accidental absorption

Photochemistry and radiation chemistry have contributed in most essential ways to both the very existence and the perpetuation of life processes. By comparison with such immensely important processes, the man-determined applications of photochemistry (such as photography in all its complicated variations, improvements, and subtleties) and of radiation chemistry are trivial indeed. The accidental consequences of the absorption of radiation (in the sense that man has little control over them) still remain the most important of the processes that affect all living creatures, including man.

Both electromagnetic (as represented particularly by ultraviolet light) and corpuscular radiations can cause chemical reactions and mutations in living matter. In the smaller living units, the effects can be either immediately lethal or terminative for production of progeny. In the larger biological units, there can be immediate destruction of some of their component parts or production of mutations, most of which are regressive and ultimately lethal. Only a small fraction of the mutations so produced are viable and persistent in modified progeny and only a minor few of the progeny are improvements on the original organism. Such occasional improvements, in the sense of their increased ability to survive and to interact favourably with their environment, may be an important contribution to the evolutionary process. In contrast, the destructive effects of the absorption of radiation can be important indeed; one suggestion is that they are important contributors to the aging process. The reason for such speculation is simple enough but some definitions are first required.

The rern unit. The rem is a quantity of biologically effective radiation that was originally written as an abbreviation for "roentgen equivalent man"; actually it is the dose in rads multiplied by a term called RBE, or "relative biological effectiveness." The rad is a unit of exposure to high-energy radiation. One rad is the energy deposited by radiation in units of 100 ergs per gram. (One erg equals 2.389×10^{-8} calorie.) RBE is itself a characteristic of the type of radiation employed and is typically about 0.7 for high-energy gamma rays.

Natural background radiation

According to U.S. federal radiation standards, 500 millirem (*i.e.*, one-half rem) of high-energy irradiation per year is set as the threshold of danger for human beings. The natural background dose from cosmic rays, radioactivity in concrete, radioactive contaminants in the smoke of burning fossil fuel, and even radiation from the soil and equally "safe" sources, however, is estimated to average 90 to 200 millirem per year. Permitted contributions from known sources are considerably smaller. Human tissue is regularly insulted (injured), in spite of any precautions, by radiation-induced, physically undetectable chemical changes.

Mutations. The most immediately apparent and dramatic evidences of the effects of high-energy radiation in humans were in the accidental causing of cancers (among many of the scientists who worked with radium in the early days after its discovery) and are in the deliberate destruction of cancers (as in the radioactive iodine treatment of thyroid cancers). Careful regulation of the many sources of radiation has reduced accidental exposure to a minimum. Nevertheless, experimentation continues to present hazards. Particularly notable is an incident that occurred at the Boris Kidrić "zero power" nuclear reactor at Vinča, Yugoslavia, on October 15, 1958. On that occasion, a girl at the control desk, who was using the quiet hours for study, accidentally permitted the reactor to run above zero power for a protracted period. She died 32 days later, and six other people at the reactor suffered near-fatal injury. No even mildly similar incident, however, has ever occurred at a power reactor; operations of nuclear-power plants are rigorously regulated and highly automated.

TECHNOLOGICAL FIELDS

The principal applications of photochemistry (including photography) are in the initiation of reactions by light that can pass through glass or quartz windows. Such light has a wavelength not less than about 185 nanometres. Light of shorter wavelength is also effective, but the windows required (sapphire, lithium fluoride, or extraordinarily thin aluminum) and the associated mechanical difficulties seriously limit application of photochemical methods in the range from 185 nanometres down to a conceivable lower limit of about 85 nanometres. Photochemical techniques are particularly applicable when a specific initial process (*e.g.*, the breakage of a particular bond in a molecule of a particular substance) is required. For such purposes, high-intensity ultraviolet lamps are usually employed, the window is glass or quartz, and the initiation reaction is limited to the relatively thin layers in which the light is absorbed. The processes include photochlorination of aromatic compounds (such as benzene, toluene, and xylene), sulfhydration of olefins, production of cyclohexanone oxime, photopolymerization (principally in surface-curing processes), sulfoxidation, and vitamin D synthesis.

High-energy electrons in industrial applications

High-energy radiation has a much greater range of potential applicability. Electrons penetrate about half a centimetre per 1,000,000 electron volts in a medium of unit density (*e.g.*, water), proportionally less in mediums of higher density, and proportionally more in mediums of lower density (*e.g.*, hydrocarbons). Low-energy electrons (including low-energy beta rays) have proved effective in the rapid curing of paints. Electrons produced (by Van de Graaff generators, as an example) at energies up to 3,000,000 electron volts have been used advantageously in many industrial processes. Gamma rays from radioactive decay have the advantage of deep penetration through the walls of pressure vessels and through the irradiated media.

An advantage of using high-energy radiation as a reaction initiator is the fact that it may be employed to start and control the rates of reactions that must be maintained, for example, at very low temperature. In the case of high-energy radiation like X-rays or gamma rays a window is not required for the admittance of the rays. Consequently, such radiation can be conveniently employed for initiation of high-pressure reactions. Principal industrial applications of high-energy radiation have been in the synthesis of ethyl bromide, synthesis of graft copolymers, curing of specialized surface coatings and paints, synthesis of ion-exchange membranes, preparation of polyethylene oxide (of carefully regulated molecular weight) for textile finishing, and preparation of polymer-impregnated woods and concrete–polymer composites. Such impregnated woods include water-resistant structural materials, parquet flooring, knife handles, and a variety of other products in which impact strength and resistance to cracking are required. Polymer-impregnated concrete is potentially important in both building and road construction. Use of high-energy radiation in the textile field has been promoted principally in Japan, but there are numerous other places throughout the world where techniques of (high-energy) radiation chemistry are promoted vigorously and employed effectively on a regular large-scale industrial basis. These include sterilization of surgical materials, sterilization of rabbit hair (principally in Australia), prevention of germination of potatoes, and production of thermoshrinkable, cross-linked polymer. The last (in diameter ranging from a fraction of an inch to a few feet) may be used, on the one hand, as insulating tubing (insulating spaghetti) for electrical wiring and, on the other, as a protective coating tubing for joints on long-distance oil and gas pipelines. Radiation treatment of foodstuffs for improved storage is actively investigated and promoted.

In the United States, use of high-energy radiation for any purpose, including manufacturing, is strictly regulated both by the federal government and by a variety of state agencies.

BIBLIOGRAPHY. M.K.E. PLANCK, *Introduction to Theoretical Physics*, vol. 4, *Theory of Light* (1932), a classic work on the subject of light and quanta; R.W. DITCHBURN, *Light*, 3rd ed., 2 vol. (1976), a well-presented text on physical optics that, although not too mathematical, does require understanding of the use of differential equations; K.Z. MORGAN and J.E. TURNER, *Principles of Radiation Protection* (1967), an outstanding and authoritative text on high-energy-radiation protection; *Current Topics in Radiation Research* (quarterly), a major collection related principally to biological effects of radiation with brief reviews confined to limited aspects of the subjects; G.K. ROLLEFSON and M. BURTON, *Photochemistry and the Mechanism of Chemical Reactions* (1939): W.A. NOYES and P.A. LEIGHTON, *The Photochemistry of Gases* (1941, reprinted 1966), classic works that include material on internal conversion and predissociation; they predate the present language of intersystem crossing which, however, is discussed in detail in the more comprehensive text by J.G. CALVERT and J.N. PITTS, *Photochemistry* (1966). *Advances in Photochemistry* (irregular), is a series concerned mainly with brief surveys of current advances in that field. G.J. DIENES and G.H. VINEYARD, *Radiation Effects in Solids* (1957); and D.S. BILLINGTON and J.H. CRAWFORD, *Radiation Damage in Solids* (1961), taken together, summarize the actual effects of radiation on solids rather thoroughly. For a survey of radiation effects on aqueous solutions and organic compounds, see J.W.T. SPINKS and R.J. WOODS, *An Introduction to Radiation Chemistry*, 2nd ed. (1976); *Actions chimiques et biologiques des radiations* (annual, 1955–71), the first survey on a large variety of subjects in radiation chemistry written by scientists largely about their own work, with the reviews in either French or English; and M.S. MATHESON and L.M. DORFMAN, *Pulse Radiolysis* (1969), an excellent book on modern techniques in radiation chemistry.

(M.Bu/A.Moz.)

Radiation Injury

The severity of radiation damage to living tissue depends upon the type of radiation, the size of the exposure, and the rate at which it is experienced. The two main categories of radiations are non-ionizing and ionizing; the latter category comprises those radiations of sufficient energy to cause electrons to be added to, or removed from. atoms or molecules previously electrically neutral (an electron is an elementary particle with a negative electrical charge and an extremely small mass). Non-ionizing radiations include visible and ultraviolet light, radio wave, and microwaves. Of these, ultraviolet radiations are a common source of injury.

ULTRAVIOLET RADIATION

Types of ultraviolet rays

Sources and terminology. Ultraviolet radiation comes from the sun and from many man-made devices such as arc welding equipment and ultraviolet lamps. This form of electromagnetic radiation is called ultraviolet because it is slightly more energetic than violet light. Three groups of wavelengths are commonly described: the "far" ultraviolet region, with wavelengths of 160 to 280 millimicrons; the "middle," from 280 to 320 millimicrons; and the "near"—closest to visible light—of 320 to 400 millimicrons. Wavelengths even shorter than 160 millimicrons can be produced under special conditions. Another way of expressing wavelength is in terms of the Angstrom (symbol Å), which is equal to 0.1 millimicron or $^1/_{100,000,000}$ centimetre. (There are 2.54 centimetres in an inch.) The longer the wavelength the more penetrating the radiation; those in the near region go through significant thicknesses of glass, water, and tissue; those in the middle region are almost completely absorbed by window glass; and those in the far region are poorly transmitted even through air. Most of the radiations reaching man from the sun have wavelengths in the middle and near classes; the earth's atmosphere shields out those of less than 295 millimicrons quite effectively.

Depending on the source of the ultraviolet radiations and the amount of shielding interposed, a particular distribution of wavelengths, or spectrum, may be encountered. Certain wavelengths are believed to be especially effective in producing specific biological effects; thus both the spectrum and the total amount of energy deposited are important in determining the result.

Even the longer ultraviolet wavelengths do not penetrate very far in tissues, and therefore the effects ordinarily seen are limited to the skin and eyes. Under laboratory conditions it is possible to study the changes produced in other types of living cell systems; important alterations, qualitatively different from those produced by ionizing radiations, are seen.

Sunburn and tanning

Injuries produced. ***Skin.*** Sunburn and tanning, well-known effects of ultravioiet radiations from the sun, are produced by mechanisms not entirely understood. The redness, swelling, and pain of sunburn are associated with denaturation—a change in the molecular structure with consequent change in characteristics—and coagulation of some of the substance of the epidermal cells, and enlargement of the small vessels beneath the outer layer (epidermis) of the skin. In severe cases, blisters and even ulcers may form, and the victim may experience a serious generalized illness with fever. Unfortunately, there is a latent period of several hours between exposure and the onset of symptoms, so that a person may receive serious injury without knowing it. Sun bathers often fail to realize that the hazard is not limited to perfectly clear days but can occur in the presence of a hazy atmosphere.

Tanning, a later manifestation, or the result of repeated exposures, is due to the darkening of existing melanin granules in the epidermis and the stimulated production of new granules of the pigment. With prolonged exposure to ultraviolet radiation, as experienced by farmers, sailors, and other outdoor workers, a definite set of chronic changes are seen in the skin. The epidermis becomes irregularly pigmented, wrinkled, and somewhat thin, and loses much of its elasticity. The connective tissues beneath the epidermis undergo undesirable changes, including alterations in the intercellular substance and damage to small blood vessels. In the general population chronic ultraviolet exposure contributes greatly to the skin changes ordinarily attributed to aging.

It is unfortunate that sunburn is often considered a trivial and healthful ailment and that deep tanning is looked upon as a manifestation of health and beauty, when, in fact, excessive exposure to ultraviolet radiation is distinctly harmful. In addition to hastening the aging process, sunlight contributes to the development of keratoses—elevated, scaling, pigmented areas—and to the development of cancer of the skin. Evidence for the role of sunlight in skin cancer is that this disease most commonly develops on exposed rather than clothed areas of the body, that it is more prevalent in people living near the Equator, and that it is more frequent in outdoor workers.

It is well known that persons with pale skin and blonde or red hair are more susceptible to most of the undesirable effects of ultraviolet light than are people with dark skin and hair; they have less melanin to protect the deeper tissues. Certain disease states specifically predispose to sensitivity to light, and some reactions to drugs are associated with severe skin manifestations at the sites reached by ultraviolet radiation. Sunlight does, however, play a part in the synthesis of vitamin D in the skin, a function useful to those with inadequate dietary intake.

Flash burn and snow blindness

Eye. Ultraviolet light can produce keratitis—an injury to the cornea of the eye with an associated inflammatory response. When produced by arc welding this may be known as a flash burn; a similar injury, known as snow blindness, may be caused by long exposure to sunlight reflected by snow, especially at high altitudes (where there is less atmosphere to absorb the radiation). After either type of injury there is usually a latent period of several hours before the onset of symptoms, which consist of intense pain and visual impairment. Some of the injured cells of the cornea may swell and die while others proliferate. The small blood vessels become swollen with blood, and there is leakage of serum and white blood cells at the site of injury. Usually the damage is repaired spontaneously within several days, leaving no significant residual abnormality, but abnormal sensitivity to light may persist for some time. These types of injury can be prevented by suitable protective lenses.

IONIZING RADIATION

Sources and terminology. The term ionizing is applied to all forms of radiation that directly or indirectly ionize the atoms with which they interact. X-rays and gamma rays are forms of electromagnetic radiation, much more energetic than ultraviolet, with wavelengths of less than ten millimicrons. Gamma and X-ray differ in the way they are produced but have identical effects on tissues. Particulate radiations, also considered as ionizing, include beta particles, which usually penetrate tissues a few millimetres or less, and alpha particles, which, in general, have even less penetrating power. Neutrons are particles without electrical charge; they may travel for some distance in tissue and have special biological effects.

The amount of ionization produced can be measured and the result used to describe the amount of energy transferred and thus to express the dose. The roentgen (R) is a unit of exposure stated in terms of an effect in air. The rad is the unit of absorbed dose in tissue; one rad is equivalent to the absorption of 100 ergs of energy per gram of absorbing material. (The rad is a small unit; when one pound receives one rad the energy deposited is equivalent to about $^1/_{300}$ of that required to lift the pound one foot.)

Injuries produced. The types of injury that ionizing radiation can cause have been extensively studied in experimental animals and in human beings who have been accidentally exposed. The damage produced depends on the type of radiation, its penetrating ability, the portion of the body exposed, the duration of exposure, and the total dose. In describing the effects it is convenient to compare a single large absorbed dose received all at once and an identical total dose that is administered in small increments over a long period. It is also convenient to discuss separately the observed results from an exposure caused by an external source of radiation, such as an X-ray machine, and one caused by radioactive material located within the body.

Acute total body irradiation. If the whole body is exposed to penetrating ionizing radiation, a rather characteristic pattern of injury occurs. Figure 1 shows, within the limits of information now available for man, what the results would be if a large number of normal persons were exposed to a single short burst of radiation and absorbed doses from a few rads up to several thousand rads. Each fine horizontal line symbolizes a dose group, and the termination of that line in the heavy black line represents the average time of death.

When therapy may be successful

In the known examples of human exposure, many more people have received relatively low doses, from which recovery is possible, than have received the higher, inevitably fatal doses. Thus the outlook for successful therapy of acute radiation injury may be good—a fact often not recognized. The change in the survival curve that can be achieved by therapy is shown in Figure 1 as a broken line.

From G.A. Andrews and R.J. Cloutier, Archives of Environmental Health, vol. 10 (March 1965); © 1965 American Medical Association

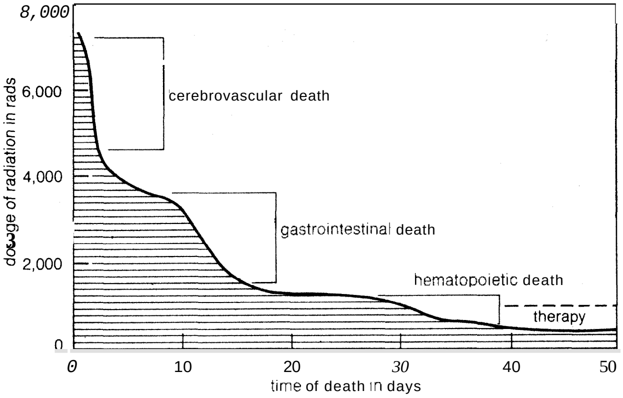


Figure 1: Expected period of survival after the entire body has been exposed to ionizing radiation (see text).

It can be seen that the survival for those receiving more than about 4,000 rads is two or three days or less and that they die with manifestations related chiefly to the nervous system and to the heart and blood vessels. Within less than an hour of exposure they experience vomiting and diarrhea. Mental changes and a fall in blood pressure soon occur, and death is preceded by convulsions and a period of unconsciousness.

The gastrointestinal syndrome claims those who receive a somewhat lower dose; the commonest time of death for persons receiving this dose is about ten days after exposure. Nausea, vomiting, and diarrhea develop a little less promptly than in persons who receive higher doses. The symptoms are very persistent, and a progressively more severe illness develops; this is believed due mainly to death of cells lining the intestinal tract. Fever, loss of fluid and of soluble salts from the blood, early damage to the bone marrow, and infection toward the close of the illness are also features of the clinical picture.

In the fraction of the population that absorbs below about 1,000 rads (but over 150) the syndrome involving the hematopoietic (blood-forming) system ensues; it is the syndrome of greatest importance because it is the one most commonly seen and the only one for which effective treatment is available. For example, a man receives an average absorbed dose of 300 rads in a radiation accident. Figure 2 shows the effects of this dose. It might correspond to an air exposure of 450 roentgens or so and is in a range that might prove lethal for some persons. The alterations produced in the numbers of circulating blood cells are caused mainly by irradiation damage to precursor cells (stem cells) in bone marrow and lymphatic tissues. Nausea and vomiting, early effects that last a day or two (and are not to be confused with the gastrointestinal symptoms produced at much higher doses), are followed by a virtually symptom-free latent interval, a hazardous period of marrow depression, and a recovery phase. The dangerous marrow depression phase is characterized by bleeding, due mainly to a lack of platelets, and by susceptibility to infection, due mainly to a lack of certain white cells (neutrophils) in the blood. Death, if it occurs, is most likely to come approximately 30 days after exposure. Effective treatment is directed toward keeping the affected person alive for about five weeks, until natural recovery processes have time to come into play. Specific measures include protection from outside infection, use of antibiotics and other antimicrobial drugs, and administration of donor platelets. Other, less established forms of treatment may be tried when recovery is doubtful with standard treatment.

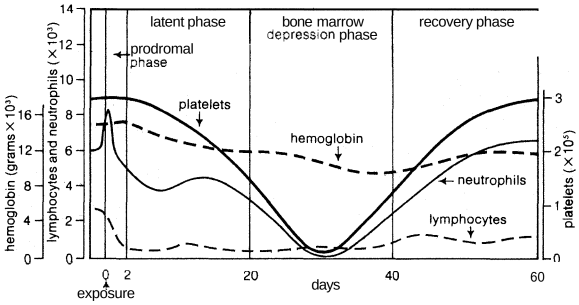


Figure 2: Changes in blood composition during successive stages of radiation illness caused by a dose of 300 rads.

From G.A. Andrews and R.J. Cloutier, Archives of Environmental Health, vol. 10, p. 501 (March 1965); © 1965 American Medical Association

In considering the three characteristic sets of effects (syndromes) caused by total body irradiation, one should not make the assumption that high doses affect only the cerebrovascular tissues, lower ones only the gastrointestinal tract, and still lower ones the hematopoietic system. Of these systems the hematopoietic is most sensitive, the gastrointestinal next, and the cerebrovascular system least. The order of promptness of manifestation is just the opposite, however. Thus it follows that a person who receives 500 rads will not show the cerebrovascular syndrome because he has not received a dose high enough to cause it. On the other hand, a person who receives 5,000 rads dies before he can show the full picture of the gastrointestinal and hematopoietic syndromes, but he would show them in severe form if somehow he could be kept alive long enough to do so. It is striking that, while the injury may all be incurred instantaneously, the manifestations of acute damage, especially the hematopoietic syndrome, develop inexorably over a period of days or weeks, in a time pattern that is rather rigidly maintained for a given species.

A person who has been irradiated by an external source usually has not been made radioactive and therefore cannot expose anyone else. In special circumstances, such as neutron irradiation, atoms naturally present in the body can be made radioactive by an external beam, but usually when this happens the amount of radioactivity produced is too small to be hazardous to other people. The only way that a person can become dangerously radioactive is by taking in a large amount of radioactive material or by having it spilled upon him.

Acute local radiation. When small parts of the body are exposed to ionizing radiation, the total dose tolerated is usually much higher than for total body exposure. Since partial body irradiation almost always leaves some of the most sensitive organ, the bone marrow, unexposed, the protected portion can save the injured person from the effects of complete marrow depletion. After local irradiation the effects are often limited to the part of the body exposed; but irradiation to the abdominal area can cause nausea and vomiting.

Irradiation for cancer

When local radiation therapy is given for cancer, the treatments are usually extended over a period of weeks, but the effects produced are more those of acute than of chronic exposure. At the doses commonly used, some degree of reddening of the skin may occur, sometimes followed by darkening of the colour of the skin. In the deeper tissues early inflammatory changes are followed by some scarring. The cancer tissues, which in cases favourable for radiation therapy have greater sensitivity to radiation than the surrounding tissues, and on which the radiation is concentrated, may undergo partial or total cell death, with prevention or curtailment of future growth.

When acute partial body exposure is received in a radiation accident, the dose may be enormously high, with the death of much tissue. Damage to blood vessels is of great significance in these cases. The blood vessels, like the bone marrow, fail to manifest the full degree of their injury at the outset; unlike the marrow, they do not have a great capacity for recovery but show progressively greater degrees of impairment. Initial swelling of the cells lining the vessels may be followed by the death of many of them, with subsequent thickening of vessel walls and eventual blocking of the vessel channels, leading to the death of the tissues that they normally supply. Early evaluation of damage in accidents involving partial body radiation often underestimates the extent of the injury. After the exposure has occurred there is no known treatment that will interrupt the progression of the manifestations of injury. The lesions need protection from infection, and treatment is required for infections that do occur. Surgical removal of dead tissue may be useful, and for some parts of the body repair by plastic surgery may be pcssible; this often tends to be limited by the fact that surrounding tissues may also have received enough radiation injury to make such repair especially difficult.

There may be acute exposures to total body irradiation combined with much higher doses to local areas of the body. In these cases, if the total body injury is not severe enough to be fatal, the affected person will be recovering from the systemic total body effects by two months after exposure but may still face many months of progression of the local injury, often without a satisfactory final healing.

Prolonged exposure. When irradiation is experienced continuously or intermittently over a period of weeks, months, or years, as might happen in an uncontrolled occupational exposure, the effects are quite different from those caused by a single acute exposure. In general the protraction allows a much larger total dose to be tolerated, because biological reparative processes tend to heal part of the damage while the exposure is still going on. Once a dose is reached that is high enough to produce clinical effects, however, the outlook for recovery is not nearly as good as it is for the single, acute exposure. If recovery does occur it is slow. The degree and types of injury are less predictable than those from an acute exposure and may include profound damage to the bone marrow as well as the manifestations discussed below in the section on delayed effects.

Isotopes and their half-lives

Internal radioisotopes. Radioisotopes are varieties of atoms that spontaneously emit radiation because of transformation of their nuclei; this process is also called decay. Each particular isotope has a half-life, which is the time required for its disintegration rate to fall by one-half; with succeeding half-lives, there is left %, %, %, 1/16, 1/32, and so on, of the original activity. For different isotopes, half-lives may vary from tiny fractions of a second to billions of years. The unit of measurement is the curie, which is the amount of a radioisotope that at the time of reference is undergoing 3.7×10^{10} (37,000,000,000) disintegrations per second. It follows that over an extended period a curie of an isotope of long half-life will have many more disintegrations than a curie of one of short half-life. Disintegration produces radiations (alpha and beta particles, gamma rays) of various energies, with a specific pattern for each radioisotope. Depending upon the type and energy of these radiations and the nature of the matter that they encounter, they may interact with the tissues close to their site of origin or at a great distance from it. Because many factors are involved, there is no simple overall relationship between the curie (the unit of measurement for radiation at its origin) and the rad (the unit of measurement for the energy deposited at the site of interaction of radiation with tissue).

Some radioisotopes when absorbed internally spread diffusely throughout the body (*e.g.*, radioactive sodium) and give what amounts to total body irradiation. Others, because of their biochemical properties, concentrate in certain types of tissue and irradiate mainly the areas where they are deposited (*e.g.*, radioactive iodine). Still others are not absorbed, being insoluble, but can become lodged internally; for example, certain inhaled particles, tend to stay in the lung, and inert radioactive substances may be deposited in penetrating wounds.

Radiations of short range, particularly alpha particles, can be especially dangerous internally, even though they are not able to travel as far as a thickness of skin. When alpha emitters are lodged internally they often cannot be detected by means of external radiation measurements.

Delayed radiation effects

It is rare for a person to take in enough radioactive material to have nausea and vomiting or other signs of acute radiation sickness. There is much more likelihood of chronic or delayed radiation effects; these occur particularly as a result of radioisotopes of long half-life that are not excreted promptly. For each radioactive material the effects produced depend upon the sites in which it localizes, its half-life, the nature of its radiations, and other factors. General types of response include degeneration of the irradiated tissue and the induction of cancer. These responses may not become manifest for many years, but once they appear there is little chance of the spontaneous improvement seen after acute radiation injury.

Characteristics of Four Radioisotopes and Their Internal Effects

radioisotope and half-life	radiations	characteristics
Radium-226 1,620 years	alpha and gamma	a naturally occurring radioactive element, discovered by Pierre and Marie Curie in 1898; when injected internally it seeks bone and is a potent cause of cancer; for modern medical use it is kept encapsulated and used as a source of gamma radiation
Iodine-131 8.1 days	beta and gamma	a product of nuclear fission and artificially produced in other ways; when absorbed internally most of it concentrates in the thyroid gland (or in certain kinds of thyroid cancer); when taken in accidentally it can damage the thyroid and may produce tumours in it; very useful medically for diagnosis and treatment of thyroid disorders
Plutonium-239 24,000 years	alpha, small amount of weak gamma	artificially produced, usually by bombarding uranium in a reactor; used, because it undergoes fission, for nuclear power and weapons; hazardous when inhaled as particles, or accidentally deposited in wounds; it is difficult to detect from outside the body
Strontium-90 28 years	strong beta	a product of nuclear fission and an important component of fallout; main source of intake is with the diet; in the body it behaves like calcium, depositing in bone, where, if present in large enough amounts, it may produce cancer

The table summarizes information on four radioisotopes that have important effects when deposited in the internal organs.

Effects on specific tissues. Immediate effects. The hematopoietic tissues of the bone marrow, which is made up of cells that undergo division frequently, constitute the most radiosensitive of all vital systems; this has been discussed in the section on acute total body irradiation.

Male and female reproductive organs are especially sensitive to ionizing radiation. Doses of several hundred rads can produce sterility that lasts several months. With uniform total body irradiation a nonlethal dose usually will not produce permanent sterility. Doses not high enough to kill the germ cells can cause damage to their chromosomes and thus lead to genetic defects.

The lining cells of the gastrointestinal tract are highly radiosensitive; they have a rapid rate of cell replacement by division, a feature generally associated with high radiosensitivity. Through death of these cells radiation thus causes denudation of the surface of the tract.

Blood vessels are easily damaged by irradiation, and damage to local tissues seen after irradiation is often a result of blood-supply impairment arising from injury to the blood vessels.

Effects of radiation on a developing embryo, which is highly radiosensitive, are dependent upon the embryo's exact stage of development at the time the radiation is received. Here again, sensitivity is associated with cell division. When a particular organ system is in its most active stage of development it is most sensitive, and radiation sustained at that time will produce abnormalities at birth that are easily identified as having occurred during the period of growth of the organ system. Many of the critical stages in human embryologic development occur very early, even before the mother is aware of her own pregnancy.

Delayed effects. The delayed effects of ionizing radiation are much less predictable and less clearly related to dose than are the acute ones. Degenerative changes can be produced in many organs, leading to impaired function; probably damage to blood vessels is an important factor in many of these changes. Organs typically damaged include the bone marrow, the bones, the kidneys, and the lungs. The lenses of the eyes may become opaque months or even years after the eyes have been irradiated. It is believed that a total dose of at least 200 rads of gamma radiation is required, and many persons tolerate more than this without any injury to the lenses (*i.e.*, getting cataracts). Neutron irradiation is especially dangerous in this respect.

Cancer. Perhaps the most severe latent effect of radiation is the predisposition to cancer in some exposed persons. The causes of most naturally occurring cancers are not known, but it is clear that the incidence of at least some types can be increased by radiation, and that the increase in the cancer rate can be significant if the doses of radiation are high. The interval between exposure and onset of the cancer is usually several years. In the early days of diagnostic radiology physicians were not as careful as they are today, and they exposed their hands while X-raying their patients. Many suffered radiation injury to their hands that later developed into skin cancer. Another type of cancer, developing in bone or in tissues adjacent to bone, is seen in persons irradiated there by radium that has been taken into the body, as occurred in workers who painted radium watch dials. Still another example is the increased incidence in cancer of the lung that occurs among uranium miners, whose respiratory systems are exposed to radiation from radon. In those miners who are cigarette smokers, the smoking appears to combine with the occupational exposure to accentuate the hazard greatly.

Cancer of the skin, bone, and lung

Leukemia, a malignant disease involving the blood-forming tissues, is clearly influenced by radiation. In the atomic bomb casualties of Japan a distinct increase in this disease began a few years after the exposure and continued for many years. Confirmatory evidence is seen in certain other groups. It is estimated that one or two added cases of leukemia per year will occur if one rad is received by each person in a population of 1,000,000. Leukemia is a rare disease, and, even when its incidence is increased in a population exposed to radiation, it occurs in a very small percent of the people. Other blood disorders, such as those characterized by bone marrow failure, fibrosis (formulation of scar tissue) of the bone marrow, or other abnormalities of blood cell formation, may sometimes result from irradiation.

In experimental animals it has been shown that radiation exposure can hasten aging processes and thus can shorten life. No such effect has been clearly shown in humans.

Genetic effects. The genetic effects of radiation are a potentially serious problem in a society that is using an increasing amount of radiation for diagnostic medical purposes and that is expanding its use of industrial processes involving radiation. To determine the hazard there has been a great deal of intense, well-supported effort by highly qualified scientists, but the problem is inherently very difficult to answer. The genetic effects of large doses of radiation in small animals are clear enough; the difficulty is in determining the genetic effects of exceedingly small doses given over a long time period to a human population. Experiments involving hundreds of thousands of animals over periods of many years are necessary to obtain statistically valid information for answering these questions. Even when answers are obtained for one species, such as the mouse, and for one set of circumstances, it is not certain that the results that have been obtained apply to humans.

Gene mutations

Two major types of effects can be produced: true gene mutations and chromosome breakage; the latter involves a number of genes. The true gene mutations might be produced by a "single hit" of a charged particle. Their probability has been stated to be directly proportional to total radiation dose and was formerly believed independent of the rate at which the radiation dose is delivered. Even though mutations are the basis of evolution, almost all of them are obviously undesirable. Since a single radiation event could cause a mutation, one would not expect a threshold—that is, a radiation dose below which no effect could be produced. Many of the mutations would be recessive and thus would not be apparent until chance pairing with a matching defect. Harmful gene mutations tend eventually to be eliminated from the population at a rate proportional to their harmfulness. It would be possible for undesirable recessive mutations to persist for many generations, however, and for the human gene pool to build up a dangerous level of them—and this has been a clearly expressed apprehension.

The other type of genetic defect—breakage of chromosomes, involving multiple genes—is believed to require more than a "single hit." These defects would increase with the square of the radiation dose, would produce many fetal deaths, but usually would not persist for generations as a hidden hazard in the gene pool of the population.

Extensive studies of the mouse suggest that the inexorable piling up of harmful recessive mutations may not be as great a hazard as has been thought by some. It is now known that when the dose is protracted, rather than given all at once, fewer persistent single-hit mutations are produced; there is apparently a repair process that is effective when the dose is not given all at once. Furthermore, at least for the female germ cell, an elapse of time between the radiation exposure and conception decreases the number of mutations that appear. In addition, the mouse studies show that the incidence of true gene mutations per rad is lower for small radiation doses than for large radiation doses,.

Without any unusual radiation exposure, about 1 percent of newborn human beings show some visible defect in development that might be caused by a mutation. Little is known about the role of existing radiation levels in producing these abnormalities, but there is good reason to believe that a great majority of them are caused by factors not related to radiation.

Some of the most direct information available regarding the genetic effects of radiation on human beings has come from the offspring of the atomic bomb casualties in Japan. It appeared at first that there was an alteration in the sex ratio in births from previously irradiated parents, suggesting selective fetal mortality, but this finding has not been substantiated. The pregnancies from this group, when compared with those of nonirradiated controls, have been reported as failing to show any detectable increase in stillbirths, neonatal deaths, gross malformations, or mortality during childhood.

BIBLIOGRAPHY. AL. LORINCZ, "Physiological and Pathological Changes in Skin from Sunburn and Suntan," ***J.A.M.A.*, 173:1227–1231 (1960)**, a good, general, not too technical summary of the effects of ultraviolet irradiation; I. MATELSKY, "The Non-Ionizing Radiations," ***Industrial Hygiene Highlights*, 1:140–178 (1968)**, information on the nature of radiations and their effects, including their role as industrial hazards; W. HARM, ***Biological Effects of Ultraviolet Radiation* (1980)**, a comprehensive work; V.P. BOND, T.M. FLIEDNER, and J.O. ARCHAMBEAU, ***Mammalian Radiation Lethality* (1965)**, an excellent detailed book emphasizing cytokinetic aspects of radiation injury; T. ALPER, ***Cellular Radiobiology* (1979)**, fundamentals of lethal effects of radiation on many types of cells; I. ASIMOV and T. DOBZHANSKY, ***The Genetic Effects of Radiation* (AEC, 1966)**, a lucid presentation for the lay reader, giving excellent explanations of genetic concepts, but not dealing with all controversial aspects of injury produced by

ionizing radiation; A.A. AWA *et al.*, "Cytogenetic Study of the Offspring of Atom Bomb Survivors," ***Nature,*** 218:367–368 (1968), an interesting, brief scientific report on the follow-up of offspring born to Japanese who were previously heavily irradiated (includes references on earlier reports of effects of the atomic bomb); R.W. MILLER, "Delayed Radiation Effects in Atomic-Bomb Survivors," ***Science,*** 166:569–574 (1969), a clear statement about the radiation injuries suffered by the Japanese after World War II; W.D. NORWOOD, ***Health Protection of Radiation Workers*** (1975), includes chapters on short- and long-term effects, and hereditary effects, of radiation exposure, as well as diagnosis and treatment of radiation injury.

(G.A.A.)

Radicals, Free

Most molecules contain even numbers of electrons, and the covalent chemical bonds holding the atoms together within a molecule normally consist of pairs of electrons jointly shared by the atoms linked by the bond. Under certain circumstances, however, fragments of such molecules may be formed. Among these are groups of atoms containing single or unpaired electrons. Such fragments may be considered to have arisen by cleavage of normal electron-pair bonds, every cleavage having produced two separate entities, each of which contains a single, unpaired electron from the broken bond (in addition, of course, to all the rest of the normal, paired electrons of the atoms). Molecular fragments with unpaired electrons are called free radicals or often, when the meaning is unambiguous, simple radicals.

Although free radicals contain unpaired electrons, they may be electrically neutral. Charged fragments, such as carbonium ions and carbanions, which also are formed by breaking molecular bonds, on the other hand, need not contain unpaired electrons. Because of their odd electrons, free radicals are usually highly reactive. They combine with one another, or with single atoms that also carry free electrons, to give ordinary molecules, all of whose electrons are paired; or they react with intact molecules, abstracting parts of the molecules to complete their own electron pairs and generating new free radicals in the process. In all these reactions, each simple free radical, because of its single unpaired electron, is able to combine with one other radical or atom containing a single unpaired electron. Under special circumstances diradicals can be formed with unpaired electrons on each of two atoms (giving an overall even number of electrons), and these diradicals have a combining power of two.

Certain free radicals are stabilized by their peculiar structures; they exist for appreciable lengths of time, given the right conditions, and can be studied by relatively straightforward techniques. Most free radicals, however, including such simple ones as the methyl ($\cdot CH_3$) and ethyl ($\cdot C_2H_5$) radicals, are capable of only the most fleeting independent existence, and very special techniques have had to be developed to study them. In spite of the difficulties, research on short-lived free radicals is actively pursued, in part because these substances are known to be active intermediates in many chemical reactions, including industrially important polymerizations. It is also believed that free radicals play a role in vital biological oxidation-reduction processes, such as photosynthesis.

An earlier usage

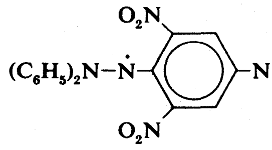
The term radical was first used in 19th-century chemical nomenclature to designate a group of atoms that comprised part of a molecule and that remained together through a number of chemical transformations. One of the first such radicals to be described was the benzoyl radical (C_6H_5CO-). Each radical of this type had an unfulfilled valence—*i.e.*, an atom whose valence number was one less than normal, such as a trivalent carbon or a bivalent nitrogen atom. In combination with another radical, or with a single atom, any of these hypothetical groupings of atoms could, in principle, be converted into a normal molecule. The interactions and transformations of radicals were presumed to make up all observed chemical reactions. When, somewhat later, the first independently existing substances corresponding in structure to one of the theoretical radicals was actually produced in the laboratory, it was designated, naturally enough, a "free" (meaning "uncombined") radical. Since that time the term radical has been used less and less frequently in the original sense; but the odd-electron molecular fragments studied today are still usually referred to as free radicals to avoid confusion with the earlier usage, which remains in the older literature.

Stable radicals. Types. The first relatively stable free radical, triphenylmethyl (structure I), was discovered by Moses Gomberg in 1900. In this compound the central carbon

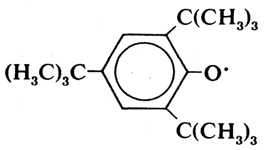
$$\underset{\text{I}}{(C_6H_5)_3C}\ .\ \underset{\text{II}}{(C_6H_5)_3C{-}C(C_6H_5)_3}$$

is trivalent since it is combined with three substituents instead of four, and its unshared electron is represented by a dot. By withdrawing chlorine from chlorotriphenylmethane, $(C_6H_5)_3CCl$, Gomberg expected to obtain hexaphenylethane (structure II) as a result of the combination of two triphenylmethyl radicals. The product obtained (in solution) was, however, a mixture of I and another substance recently shown to be an isomer hexaphenylethane which remains unknown. Actually, in solution, I and this dimer are in rapid equilibrium with each other, the dimer dissociating into two triphenylmethyl radicals, and the radicals continually recombining. Free radicals of the triphenylmethyl type are stable only in certain organic solvents; they are rapidly destroyed by irreversible reactions in the presence of air, water, or strong acids. The degrees of dissociation of such radical dimers vary considerably when the phenyl groups are replaced by other aromatic groups. Thus, when each phenyl, C_6H_5, is replaced by diphenyl, $C_6H_5-C_6H_4$, dissociation into free radicals is practically complete.

In a manner analogous to the above, free radicals are formed to a small extent by the breaking of the nitrogen–nitrogen bond in aromatic hydrazines of the general structure R_2N-NR_2, or to a greater extent, by the breaking of the central nitrogen–nitrogen bond in aromatic tetrazanes, $RN-RN-NR-NR_2$. Thus, the radical 1,1-diphenyl-2-picrylhydrazyl (structure III) exists as a stable violet solid, showing no tendency to dimerize to a tetrazane even in the solid state. Similar examples of free radicals, in which, however, the odd electron is on oxygen, are also known; *e.g.*, the 2,4,6-tri-*tert*-butyl-phenoxy radical (structure IV), a blue substance that is also known in the dimeric form.

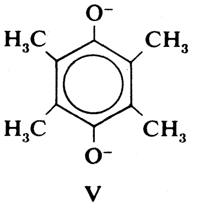


III

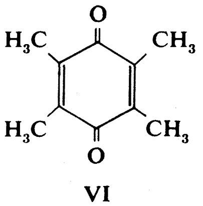


IV

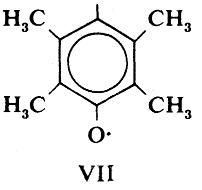
Another variety of free radical with a degree of stability is represented by the class of compounds known as semiquinones. If the dianion of the durohydroquinone (structure V) that exists in alkaline solutions of durohydroquinone is treated with oxidizing agents, it loses two electrons and is converted to duroquinone (structure VI). An intermediate stage of oxidation is possible, however, in which a single electron has been lost to yield the free



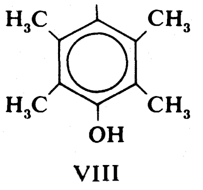
V



VI



VII



VIII

Radical ions

radical smiquinone ion (structure VII). Actually, in partially oxidized systems all three species are in mobile equilibrium, and mixing colourless solutions of V and VI immediately gives rise to detectable quantities of the yellow semiquinone. Semiquinones of a variety of types are known and in some cases may be obtained in pure form in the solid crystalline state. Most semiquinones, however, are usually in rather unfavourable equilibrium with the completely oxidized and reduced forms of the molecule, so that for many years their existence was overlooked.

Still another type of stable radical ion, a metal ketyl, forms when a substance such as benzophenone,

$(C_6H_5)_2C{=}O,$

is treated with metallic sodium to give the coloured substance $(C_6H_5)_2C{-}O^-$. The reaction involved is essentially a reduction, the sodium being oxidized to sodium ion, Na^+. Rather similarly in certain solvents such as tetrahydrofuran, sodium reacts with complex aromatic hydrocarbons such as naphthalene, converting them to the highly coloured radical ions.

A final class of relatively stable organic free radicals are those containing the group > NO. Such a group necessarily contains an odd electron (apparently distributed between the nitrogen and oxygen), and the resulting molecule is thus a free radical. An example is diphenylnitrogen oxide, $(C_6H_5)_2NO$, which is obtained by the oxidation of diphenylhydroxylamine, $(C_6H_5)_2NOH$, and which shows no tendency to dimerize or to disproportionate to products of higher and lower oxidation state. A somewhat similar behaviour of nitrogen is found in nitric oxide, NO, and nitrogen dioxide, NO_2, which are also molecules with odd numbers of electrons.

Structural requirements. Certain structural features appear to be required for the existence of stable free radicals. One condition of particular importance is shown by the semiquinone radical ion VII. As depicted, the upper oxygen atom has a negative charge and the lower one an odd electron. This assignment is arbitrary, however, and the same molecule would be represented if the charge and the odd electron were interchanged. When a situation of this type is encountered, the actual average distribution of electrons within the molecule is presumed not to be that of either of the structures just described but to be intermediate between the two. This circumstance is called delocalization or resonance; according to quantum mechanics, the resonance considerably increases the stability of the substance and, as in this case, the probability of its existence. When the oxidation of duroquinone is conducted in an acid medium, so that the semiquinone anion would be converted to the neutral radical VIII, the opportunity for delocalization disappears, and under such conditions no semiquinone can be detected.

Delocalization

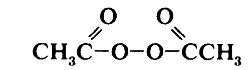
Similar arguments account for the stability of the other free radicals discussed above. In the case of triphenylmethyl I, it is considered that the odd electron is in effect distributed over the three phenyl groups as well as the central carbon atom. In addition, the dissociation of hexaphenylethane may be considerably facilitated by repulsive forces between the three bulky phenyl groups on one carbon atom and those on the other. As a measure of the magnitude of these two phenomena, dissociation of the carbon-carbon bond in the simple molecule ethane, $H_3C{-}CH_3$, requires an energy input seven times that needed to break the comparable bond in the molecule hexaphenylethane.

Unstable radicals. ***Preparation and detection.*** Simple free radicals such as methyl, $\cdot CH_3$, also are capable of existence and play important roles as transient intermediates in many chemical reactions. The existence of the methyl radical was first demonstrated by Friedrich Adolf Paneth and W. Hofeditz in 1929 by the following experiment. The vapours of tetramethyllead, $Pb(CH_3)_4$, mixed with gaseous hydrogen, H_2, were passed through a silica tube at low pressure. When a portion of the tube was heated to about 800" C, the tetramethyllead was decomposed and a mirror of metallic lead deposited on the internal surface of the tube. The gaseous products of the decomposition were found capable of causing the disappearance of a second lead mirror, deposited at a more distant cool point in the tube. Since none of the recognized stable products of the decomposition was able similarly to dissolve a lead mirror, the inference was drawn that methyl radicals formed in the high-temperature decomposition reacted with lead at the cool mirror to regenerate tetramethyllead. Methyl radicals obtained in this way proved to be highly reactive and short-lived. They not only reacted with lead and other metals but disappeared rapidly and spontaneously, largely by dimerization to ethane, $H_3C{-}CH_3$. In hydrogen at a total pressure of two millimetres, for example, the concentration of methyl radicals was estimated to decrease to half its original value in about 0.006 seconds. In subsequent years, techniques for producing reactive free radicals in the gas phase have been greatly extended by Paneth and his coworkers and many other investigator~It has been found that a variety of unstable species, such as ethyl, $(\cdot C_2H_5)$ propyl, $(\cdot C_3H_7)$, and hydroxyl, (OH) can be obtained by several methods including: (1) thermal or photochemical decomposition of a wide variety of organic and inorganic materials, (2) reaction between sodium vapour and an alkyl halide, and (3) discharge of electricity through a gas at low pressure. Atoms that arise from dissociation of a diatomic molecule (such as the chlorine atom, .Cl, from the dissociation of the chlorine molecule Cl_2) can also be obtained and have the properties of short-lived radicals of this type.

Disappearance of metal mirrors

The existence of the various known unstable free radicals, like that of methyl itself, is most commonly demonstrated by the reactions that they undergo. Thus, ethyl radicals, formed from tetraethyllead, $Pb(C_2H_5)_4$, dissolve zinc and antimony mirrors. The resulting ethyl derivatives of zinc and antimony, $Zn(C_2H_5)_2$ and $Sb(C_2H_5)_3$, have been isolated and chemically identified. In a few instances, unstable radicals also have been identified spectroscopically. Here the important technique of flash photolysis, the use of a single, intense flash of light to produce a momentary high concentration of free radicals, is used.

Transient, unstable free radicals also may be produced in solution by several means. A number of molecules, of which organic peroxides are typical, possess such weak chemical bonds that they decompose irreversibly into free radicals on warming in solution. Diacetyl peroxide, for example,



is considered to decompose, at least in large part, into carbon dioxide, CO_2, and methyl radicals. These, in turn, rapidly attack most organic solvents, often by abstracting hydrogen to given methane, CH_4, together with other products. Irradiation of solutions of many organic substances with ultraviolet light leads to the absorption of sufficient energy to disrupt chemical bonds and produce free radicals, and, in fact, most photochemical processes are at present thought to involve free radical (sometimes diradical) intermediates. The chemical changes that occur when solutions (and also gases) are exposed to high-energy radiation (α-β-γ-rays from nuclear reactions, X-rays, electrons beams, etc.) also frequently appear to involve the transient formation of free radicals. Thus, a major consequence of passing high-energy radiation into water, H_2O, is believed to be its dissociation into hydroxyl radicals, $\cdot OH$, and hydrogen atoms, $H\cdot$, together with solvated electrons. Short-lived radicals also are formed in some oxidation-reduction processes. Organic hydroperoxides of the general formula R–OOH react rapidly with ferrous ion, Fe^{2}+, the first step being the formation of ferric ion, Fe^{3}+, hydroxide ion, OH–, and an alkoxy radical, $\cdot RO$.

Radicals from radiation

An interesting method for prolonging the lifetime of highly reactive radicals is to produce them in a glassy or solid medium where they are unable to diffuse together and disappear by dimerization and similar processes. The technique was first developed by Gilbert Newton Lewis, who irradiated various organic molecules, chiefly dyes, in glassy solvents at liquid air temperatures and was able to observe the colours of the resulting unstable radicals, which in some cases could be preserved for days. Similarly, the irradiation of organic plastics (and many other solid organic materials) with high-energy radiation at ordinary temperatures can produce high concentrations of trapped radicals that may be detected and studied via their electron paramagnetic resonance spectra (see below).

Short-lived radicals as intermediates. It is generally considered that free radicals are transient intermediates in many high-temperature reactions (such as combustion and the thermal cracking of hydrocarbons), in many photochemical processes, and in a number of other important reactions in organic chemistry, although the concentrations of the free radical intermediates are in general too low for direct detection. One class of free radical reaction is of particular importance and is illustrated by the following example. Methane, CH_4, reacts with chlorine, Cl_2, by an overall process that gives chloromethane, CH_3Cl, and hydrogen chloride, HCl. The reaction is enormously accelerated by light and apparently involves the following steps:

$$\begin{aligned} Cl_2 &\rightarrow {}_2Cl\cdot \\ Cl\cdot + CH_4 &\rightarrow \cdot CH_3 + HCl \\ \cdot CH_3 + Cl_2 &\rightarrow CH_3Cl + Cl\cdot \\ {}_2Cl\cdot &\rightarrow Clz \end{aligned}$$

Radical chain reaction

Chlorine atoms are produced in (1) and destroyed in (4), while the products that are actually isolated arise from (2) and (3). Since chlorine atoms consumed in (2) are regenerated in (3), a single atom of chlorine can lead to the production of many (actually many thousand) molecules of chloromethane. Such processes, in which an intermediate is continually regenerated, are known as chain reactions, and their study constitutes an important branch of chemical kinetics. Similar chains involving transient free radicals are involved in the halogenation of many other organic molecules, in many of the polymerization reactions employed in the manufacture of plastics and synthetic rubber, and in the reaction of molecular oxygen, O_2, with a great number of organic molecules. In general these chain processes occur with great rapidity—a typical lifetime of a chain may be of the order of a second, during which many thousands of individual steps such as (2) and (3) may occur—although at any one time perhaps only one molecule out of 10^9 in the system exists as a free radical. Since a relatively small number of free radicals can produce significant chemical reaction in this manner, the initiation of typical radical chain processes (such as polymerization of certain unsaturated molecules) provides one of the best demonstrations of the existence of transient free radicals in chemical systems.

While most reactions of molecular oxygen with organic molecules appear to involve free radical intermediates, the situation in other oxidations and reductions depends much upon the particular system involved. Thus, the oxidation of isopropyl alcohol, $(CH_3)_2CHOH$, to acetone, $(CH_3)_2C = O$, by ferrous ion and hydrogen peroxide apparently involves the intermediate radical $(CH_3)_2COH$. On the other hand, there is good evidence that when chromic acid, H_2CrO_4, is the oxidizing agent, the oxidation occurs in one step without the intervention of any intermediate radical. Basically, the presence or absence of intermediate free radicals in oxidations and reductions depends upon whether the reactions involve the transfer of single electrons (which necessarily produces an odd molecule as an intermediate) or involve two-electron transfers (which do not produce odd molecules). In many reactions, including most of the important oxidation-reductions occurring in biochemical systems, which path is followed is not yet known.

Magnetic properties of free radicals. The magnetic properties of free radicals provide a powerful tool for their detection and study. Molecules with even numbers of paired electrons are diamagnetic; *i.e.*, they are slightly repelled by a magnet. Free radicals, however, are paramagnetic (attracted by a magnet) because of the spin of the odd electron, the spins of the remaining paired electrons effectively cancelling each other. The magnetic property of a substance most commonly studied is its magnetic susceptibility, effectively its behaviour in an inhomogeneous magnetic field, and the extent of paramagnetism of the substance is described in terms of its magnetic dipole moment. The magnitude of this dipole moment, which is the same for all free radicals containing single electrons, can be calculated, and the value obtained (1.73 Bohr magnetons) has been confirmed experimentally with free radicals in the solid state or at known concentrations in solution. Magnetic susceptibility measurements may be used to demonstrate the existence of free radicals and to measure the position of equilibrium between radicals and their dimers or disproportionation products. Diradicals, with even numbers of electrons, two of which, however, are not paired, are also paramagnetic, the oxygen molecule, O_2, being probably the simplest example of the kind.

Paramagnetic resonance spectra

The electron paramagnetic resonance spectra of free radicals provide another technique for their detection and study. According to quantum mechanics, the spin of the odd electron of a free radical, when placed in a magnetic field, may have two, and only two, orientations, one with and the other against the field. These two orientations differ slightly in energy by an amount proportional to the strength of the magnetic field, and the majority of the electrons have the orientation of lower energy. If a system containing free radicals is placed in a magnetic field and exposed to electromagnetic radiation (usually in the microwave region of very short radio waves), molecules with the lower energy orientation absorb radiation of a frequency corresponding to an energy just sufficient to flip the odd electron into its higher energy state. This phenomenon gives rise in the simplest case to a paramagnetic resonance absorption spectrum consisting of a single sharp absorption line. The technique is sensitive and will detect extremely small concentrations of free radicals, as little as one part in 10^7 having been reported detected. In many organic free radicals, interaction of the odd electron with the magnetic moments of the nuclei of different atoms in the molecule (most commonly with the nuclei of hydrogen atoms) gives rise to a more complicated system of energy levels and an absorption spectrum consisting of a series of lines. The nature of the spectrum permits the identification of particular free radicals and also gives information about their electronic structure. As an example, the simple semiquinone radical ion, $\cdot OC_6H_4O^-$ (analogous to VII, with the CH_3 groups replaced by hydrogen atoms), exhibits a spectrum of five equally spaced lines, indicating that the odd electron spends part of its time near the hydrogen nuclei. Accordingly, it is not only associated with the two oxygen atoms but is effectively delocalized over the entire molecule. Similarly, the triphenylmethyl radical I shows a complex spectrum, consistent with the idea that its stability arises from the spreading of the odd electron over all three phenyl groups.

BIBLIOGRAPHY. Short accounts of free radicals and their reactions appear in most general texts on organic chemistry. The following specialized works are especially recommended: C. WALLING, *Free Radicals in Solution* (1957), a comprehensive account at the time it was written; W.A. PRYOR, *Free Radicals* (1966), less detailed but covers more recent data; A. FORRESTER, J. HAY, and R. THOMSON, *Organic Chemistry of Stable Free Radicals* (1968), a comprehensive discussion of free radicals sufficiently stable for direct observation; and P.B. AYSCOUGH, *Electron Spin Resonance in Chemistry* (1967), a discussion of the principles and application of electron spin resonance to the study of free radicals.

(C.T.W.)

Radio

The term radio covers the radiation and detection of signals propagated through space as electromagnetic waves to convey information. One of the chief branches of telecommunication, radio embraces wireless telegraphy, telephony, and television. Specialized communications applications include radar and certain navigation aids, and an industrial application is radio-frequency heating.

GENERAL PRINCIPLES

Elecromagnetic radiation (*q.v.*) includes light as well as radio waves, and the two have many properties in common. Both are propagated through space in approximately straight lines at a velocity of about 300,000,000 metres (186,000 miles) per second and have amplitudes that vary cyclically with time; that is, they oscillate from zero amplitude to a maximum and back again. The number of times the cycle is repeated in one second is called the frequency (symbolized as f) in cycles per second, and the time taken to complete one cycle is l/f seconds, sometimes called the period. To commemorate the German pioneer Heinrich Hertz, who carried out some of the early radio experiments, the cycle per second is now called a hertz so that a frequency of one cycle per second is written as one hertz (abbreviated Hz). Higher frequencies are abbreviated as shown in Table 1.

Table 1: Frequency Terms and Their Abbreviations

term	cycles per second	abbreviation	equivalent
1 hertz	1	1 Hz	
1 kilohertz	1,000	1 kHz	1,000 Hz
1 megahertz	1,000,000 (10^6)	1 MHz	1,000 kHz
1 gigahertz	1,000,000,000 (10^9)	1 GHz	1,000 MHz

A radio wave being propagated through space will at any given instant have an amplitude variation along its direction of travel similar to that of its time variation, much like a wave travelling on a body of water. The distance from one wave crest to the next is known as the wavelength.

There is a definite relationship between wavelength and frequency. Dividing the speed of the electromagnetic wave (*c*) by the wavelength (designated by the Greek letter lambda, A) gives the frequency: $f = c/\lambda$. Thus a wavelength of ten metres has a frequency of 300,000,000 divided by ten, or 30,000,000 hertz (30 megahertz). The wavelength of light is much shorter than that of a radio wave. At the centre of the light spectrum the wavelength is about 0.5 micron (0.0000005 metre), or a frequency of 6×10^{14} hertz or 600,000 gigahertz (one gigahertz equals 1,000,000,000 hertz). The maximum frequency in the radio spectrum is usually taken to be about 45 gigahertz, corresponding to a wavelength of about 6.7 millimetres. Radio waves can be generated and used at frequencies lower than ten kilohertz (A = 30,000 metres).

Mechanism of wave propagation. Like all other electromagnetic waves, a radio wave is made up of electric and magnetic fields vibrating mutually at right angles to each other in space. When these two fields are operating synchronously in time, they are said to be in time phase; *i.e.*, both reach their maxima and minima together and both go through zero together. As the distance from the source of energy increases, the area over which the electric and magnetic energy is spread is increased, so that the available energy per unit area is decreased. Radio-signal intensity, like light intensity, decreases as the distance from the source increases.

Function of transmitting antenna

A transmitting antenna is a device that projects as much as possible of the radio-frequency energy generated by a transmitter into space. The antenna can be designed to concentrate the radio energy into a beam like a searchlight and so increase its effectiveness in a given direction (see ANTENNAS AND WAVE GUIDES).

Frequency bands. The radio-frequency spectrum is divided arbitrarily into a number of bands from very low frequencies to superhigh frequencies. The bands are designated in Table 2. Sections of the radio-frequency spectrum have been allocated by international agreement to the various users (see Table 3), such as telegraph, telephonic speech, telemetry, and radio and television broadcasting.

Table 2: Frequency Band Designations

frequency designation	frequency range	wavelength range
Very low frequencies (vlf)	3–30 kilohertz	100,000–10,000 m
Low frequencies (lf)	30–300 kilohertz	10,000–1,000 m
Medium frequencies (mf)	300–3,000 kilohertz	1,000–100 m
High frequencies* (hf)	3–30 megahertz	100-10 m
Very high frequencies (vhf)	30–300 megahertz	10–1 m
Ultrahigh frequencies (uhf)	300–3,000 megahertz	1 m–10 cm
Superhigh frequencies (shf)	3–30 gigahertz	10–1 cm

'Also called shortwaves.

The radio-frequency bandwidth is the range of frequencies covered by the modulated radio-frequency signal. The information carried by the radio-frequency signal has a certain bandwidth associated with it, and the radio-frequency carrier must have a channel width at least as great as the information bandwidth. For regular amplitude-modulated (AM) broadcasting (see below) the radio-frequency bandwidth must be twice the information-frequency bandwidth. When other techniques of modulation are employed, bandwidth may have to be much greater. Teleprinter and telex operation requires only a small bandwidth, on the order of 200 hertz, depending on the maximum speed of the pulses forming the information code. Telephonic speech must have high intelligibility, but naturalness (high fidelity) is not of great importance. Tests have shown that the main components of speech lie between about 300 and 3,500 hertz, and telephonic channels carried by radio are therefore normally confined to a bandwidth of about four kilohertz. The smaller the information bandwidth employed the more speech channels can be carried in a given carrier bandwidth and the more economical the system will be.

Minimum acceptable bandwidths

Young people can hear audio frequencies ranging from about 30 hertz to 18 kilohertz, but as humans grow older, hearing ranges from about 100 hertz to 10 kilohertz. For high-quality (high-fidelity) reproduction of voice or speech (as in radio sound broadcasting), the range should be not less than about 30 hertz (the lowest frequency of a large organ pipe) to 15 kilohertz (piccolo, cymbal, triangle). Acceptable audio quality under certain circumstances may be achieved with a bandwidth as small as five kilohertz, as in AM (amplitude-modulated) radio; a much larger bandwidth is needed for transmitting a moving picture, as in television broadcasting, because it is necessary to convey the overall average light content of a picture as well as the picture detail. The average light content requires frequencies as low as 20 hertz to be transmitted, and picture detail demands frequencies up to five megahertz for a standard television picture.

Radar apparatus is used to determine the distance to an object or to provide a visual image of the position of objects in a given area. For this purpose, very short pulses of radio energy are transmitted to and reflected back by an object. The reflected pulses are in turn picked up by receiving apparatus, and distance is determined by the time delay between outgoing and returning pulses. The pulse must not be distorted upon its return or distance determination will be inaccurate. Thus radar requires a bandwidth on the order of five megahertz, similar to that for television broadcasting.

Modulators and demodulators. A carrier wave is a radio-frequency wave that carries information. The information is attached to the carrier wave by means of a modulation process that involves the variation of one of the carrier-frequency characteristics, such as its amplitude, its frequency, or its duration. (All of these processes are discussed in greater detail in the article TELEPHONE AND TELECOMMUNICATIONS SYSTEMS.)

In amplitude modulation the information signal varies the amplitude of the carrier wave, a process that produces a band of frequencies known as sidebands on each side of the carrier frequency. These sidebands (a pair to each modulation frequency) cover a range of frequencies equal to the sum and difference between the information signal and the carrier frequency.

Frequency modulation involves varying the frequency (the number of times the wave passes through a complete cycle in a given period of time, measured as cycles per second) of the carrier in accordance with the amplitude of the information signal. The amplitude of the carrier wave is unaffected; only its frequency changes. Frequency modulation produces more (often many more) than one pair of side frequencies for each modulation frequency.

The variation of carrier frequency is known as the fre-

Table 3: Utilization of Radio-Frequency Spectrum

very low frequencies (vlf)		timc signals, standard frequencies
	30 kHz	
low frequencies (If)		fixed, maritime mobile, navigational, radio broadcasting
	300 kHz	
medium frequencies (mf)		land, maritime mobile, radio broadcasting
	3 MHz	
high frequencies (hf)		fixed, mobile, maritime and aeronautical mobile, radio broadcasting, amateur
	30 MHz	
very high frequencies (vhf)		fixed, mobile, maritime and aeronautical mobile, amateur, radio and television broadcasting, radio navigation
	30Q MHz	
ultrahigh frequencies (uhf)		fixed, mobile, maritime and aeronautical mobile, amateur, television broadcasting, radio location and navigation, meterological, space communication
	3 GHz	
superhigh frequencies (shf)		fixed, mobile, radio location and navigation, space and satellite communication
	30 GHz	

quency deviation, and for very-high-frequency broadcasting it can reach ± 75 kilohertz. The greater the frequency deviation the greater is the effective modulation. Though theoretically its maximum value need not be limited to 75 kilohertz, any increase beyond this value requires a wider channel, which adds to the cost of reception and reduces the number of transmitters that can be accommodated in the band. The total channel width is approximately twice the sum of the maximum deviation frequency and modulating frequency. If channel width is restricted in either transmitter or receiver circuits, distortion of the information signal occurs.

A radio broadcast normally consists of only one information signal. The listener hears what he would hear at the microphone position if only one of his ears was functioning; *i.e.*, it is a monaural system. In such a system it is not possible to gain any impression of the position of the instrument groupings in an orchestra, nor can lateral movement be indicated, though movement toward or away from the microphone is conveyed by a change in sound volume.

Stereophonic broadcasting requires two microphones, one to collect sounds from the left and one from the right; the two sets of information must be separable in the receiver and be fed to loudspeakers on the left and on the right at the listening position. For high-fidelity reproduction the full audio range up to 15 kilohertz is transmitted; this can only be achieved satisfactorily at very high frequencies with frequency modulation. The broadcast signal is received on monaural receivers by making one set of information the sum of left and right signals (L + R). The other set of information is the difference of left and right (L − R). Summation of the two sets of information at the receiver output recovers the left (L) signal and subtraction recovers the right (R) signal.

Pulse code modulation

Another system of modulation switches the carrier on and off in pulses, the duration or position of the pulse being determined by the information signal. This system of pulse-coded modulation can provide better protection from noise, and a number of separate speech channels can be combined by allocating specified groups of pulses for each information channel and then interleaving these pulses in a process called time division multiplex. A comparatively wide transmission channel is needed, and the carrier must be an ultrahigh or superhigh frequency.

The ionosphere. An English mathematician, Oliver Heaviside, and a U.S. electrical engineer, Arthur Edwin Kennelly, almost simultaneously suggested about 1901 that radio waves, which normally travel in straight lines, are returned to Earth when projected skyward because electrified (ionized) layers of air above the Earth (the ionosphere [*q.v.*]) reflect or refract (bend) them back to Earth, thus extending the range of a transmitter far beyond line of sight. In 1923 the suggestion was proved to be accurate when pulses of radio energy were transmitted vertically upward and returning pulses were received back from the reflecting layer. By measuring the time between the outgoing and returning pulses, it was possible to estimate the height and number of layers. Three layers can normally be distinguished at distances from 50 to about 400 kilometres (30 to 250 miles) above the earth's surface. The layers result from a breakdown of gas atoms into positively charged ions and free electrons caused by energy radiated from the sun. The electrons maintain a separate existence in the lower layers for as long as the Sun's energy is being received, and in the upper layers some can remain free throughout the hours of darkness.

The three layers are designated D, E, and F. The D layer is approximately 80 kilometres (50 miles) high and exists only during daylight hours. Because it absorbs medium frequencies and the lower frequencies of the shortwave bands, it limits the range of such stations during daylight. The E layer, about 110 kilometres (68 miles) high, maintains its reflectivity for four or five hours after the Sun sets and so extends the range of such stations to as much as 1,000 kilometres (620 miles). This layer also serves as a good reflector of shortwaves during the day and into the night, until its reflectivity drops.

Most important of the three layers is the F layer, which has considerable power to reflect the higher frequencies. During the day it often splits into two layers (F_1 and F_2) at about 200 and 400 kilometres (125 and 250 miles), but at night only one layer is generally present at a height of about 300 kilometres (190 miles).

Radio noise, fading, and interference. Any sudden discharge of electrical energy, like that of lightning, produces transient (short-duration) radio-frequency waves, which are picked up by antennas in the same way as normal signals. These packets of radio-frequency energy produce the crackle heard on an amplitude-modulated radio receiver when an electrical storm is nearby and may be classed as natural noise.

Switching of high-voltage power lines can produce similar effects; the lines help to carry the noise-producing signals over long distances. Local switching of lights and electrical machinery can also produce the familiar crackle when the receiver is close to the noise-producing source. These sources are classed as man-made noise.

Generally noise of both types decreases as the frequency is increased. An exception is automobile ignition noise, which produces maximum effect in the very-high-frequency range, causing a sound in nearby loudspeakers every time a spark plug fires. Many countries have legislation requiring the suppression of man-made noise by means of filters that reduce the amount of radio-frequency energy released at the source. Metallic shielding of leads to and from the noise source curtails the radiated interference. It is also possible to install various noise-reducing devices at the input to radio receivers.

Filters in noise production

Noise is also caused by irregularities in the flow of electrons in metals, transistors, and electron tubes. This source of noise ultimately limits the maximum useful signal amplification that can be provided by a receiver. Noise due to the random movement of electrons causes a hiss in the loudspeaker. Radio noise can also be picked up from outer space as a hiss similar to random electron noise.

Fading of a signal, on the other hand, is due to variation in the propagation characteristics of the signal path or paths. This is particularly true when propagation depends on reflection from the ionosphere as it does for shortwaves. Propagation of waves in the very-high-frequency range and above, which penetrate the ionosphere, can be affected by temperature changes in the stratosphere, that part of the atmosphere up to about 15 kilometres (nine miles) from the Earth's surface. The fading effect can be greatly reduced at the receiver loudspeaker by various electronic controls, such as automatic volume control.

The phenomenon of interference occurs when an undesired signal overlaps the channel reserved for the desired signal. By interaction with the desired carrier, the undesired information may cause speech to become unintelligible. Countermeasures include narrowing the desired channel, thus losing some information but preventing overlap, and using a directional antenna to discriminate against the undesired transmission.

EARLY HISTORY

Maxwell's prediction. Early in the 19th century, Michael Faraday, an English physicist, demonstrated that an electric current can produce a local magnetic field and that the energy in this field will return to the circuit when the current is stopped or changed. James Clerk Maxwell, professor of experimental physics at Cambridge, in 1864 proved mathematically that any electrical disturbance could produce an effect at a considerable distance from the point at which it occurred and predicted that electromagnetic energy could travel outward from a source as waves moving at the speed of light.

Hertz: radio-wave experiments. At the time of Maxwell's prediction there were no known means of propagating or detecting the presence of electromagnetic waves in space. It was not until about 1888 that Maxwell's theory was tested by Heinrich Hertz, who demonstrated that Maxwell's predictions were true at least over short distances by installing a spark gap (two conductors separated by a short gap) at the centre of a parabolic metal mirror. A wire ring connected to another spark gap was placed about five feet (1.5 metres) away at the focus of another parabolic collector in line with the first. A spark jumping across the first gap caused a smaller spark to jump across the gap in the ring five feet away. Hertz showed that the waves travelled in straight lines and that they could be reflected by a metal sheet just as light waves are reflected by a mirror.

Maxwell's prediction verified

Marconi's development of wireless telegraphy. The Italian physicist Guglielmo Marconi, whose main genius was in his perseverance and refusal to accept expert opinion, repeated Hertz's experiments and eventually succeeded in getting secondary sparks over a distance of 30 feet (nine metres). In his experiment he attached one side of the primary spark gap to an elevated wire (in effect, an antenna) and the other to Earth, with a similar arrangement for the secondary gap at the receiving point. The distance between transmitter and receiver was gradually increased first to 300 yards (275 metres), then to two miles (three kilometres), then across the English Channel. Finally, in 1901, Marconi bridged the Atlantic when the letter *s* in Morse code travelled from Poldhu, Cornwall, to St. John's, Newfoundland, a distance of nearly 2,000 miles (3,200 kilometres). For this distance, Marconi replaced the secondary-spark detector with a device known as a coherer, which had been invented by a French electrical engineer, Edouard Branly, in 1890. Branly's detector consisted of a tube filled with iron filings that coalesced, or "cohered," when a radio-frequency voltage was applied to the ends of the tube. The cohesion of the iron filings allowed the passage of current from an auxiliary power supply to operate a relay that reproduced the Morse signals. The coherer had to be regularly tapped to separate the filings and prepare them to react to the next radio-frequency signal.

The Fleming diode and De Forest audion. The next major event was the discovery that an electrode operating at a positive voltage inside the evacuated envelope of a heated filament lamp would carry a current. The American inventor Thomas A. Edison had noted that the bulb of such a lamp blackened near the positive electrode, but it was Sir John Ambrose Fleming, professor of electrical engineering at Imperial College, London, who explored the phenomenon and in 1904 discovered the one-directional current effect between a positively biassed electrode, which he called the anode, and the heated filament; the electrons flowed from filament to anode only. Fleming called the device a diode because it contained two electrodes, the anode and the heated filament; he noted that when an alternating current was applied, only the positive halves of the waves were passed—that is, the wave was rectified (changed from alternating to direct current). The diode could also be used to detect radio-frequency signals since it suppressed half the radio-frequency wave and produced a pulsed direct current corresponding to the on and off of the Morse code transmitted signals. Fleming's discovery was the first step to the amplifier tube that in the early part of the 20th century revolutionized radio communication.

Fleming failed to appreciate the possibilities he had opened up and it was the American inventor Lee De Forest who in 1906 conceived the idea of interposing an open-meshed grid between the heated filament and positively biassed anode, or plate, to control the flow of electrons. De Forest called his invention an Audion. With it he could obtain a large voltage change at the plate for a small voltage change on the grid electrode. This was a discovery of major importance because it made it possible to amplify the radio-frequency signal picked up by the antenna before application to the receiver detector; thus, much weaker signals could be utilized than had previously been possible.

Invention of the Audion

Early research by commercial companies. The first commercial company to be incorporated for the manufacture of radio apparatus was the Wireless Telegraph and Signal Company, Ltd. (England) in July 1897 (later changed to Marconi's Wireless Telegraph Company, Ltd.); other countries soon showed an interest in the commercial exploitation of radio.

Among the major developments of the first two decades of the 20th century was De Forest's discovery in 1912 of the oscillating properties of his Audion tube, a discovery that led to the replacement of the spark transmitter by an electronic tube oscillator that could generate much purer radio waves of relatively stable frequency. By 1910, radio messages between land stations and ships had become commonplace, and in that year the first air-to-ground radio contact was established from an aircraft. In 1918 a radiotelegraph message from the Marconi long-wave station at Caemarvon, in Wales, was received in Australia, over a distance of 11,000 miles (17,700 kilometres).

Though early experiments had shown that speech could be transmitted by radio, the first significant demonstration was not made until 1915 when the American Telephone & Telegraph Company successfully transmitted speech signals from west to east across the Atlantic between Arlington, Virginia, and Paris. A year later, a radiotelephone message was conveyed to an aircraft flying near Brooklands (England) airfield. In 1919 a Marconi engineer spoke across the Atlantic in the reverse direction from Ballybunion, Ireland, to the U.S.

From 1920 onward radio made phenomenal progress through research activities in Europe, America, and Asia. The invention of the electron tube and later the transistor (1948) made possible remarkable developments in fields such as control and computing as well as telecommunications.

Sarnoff's proposals

Beginning of broadcast service. In 1916 David Sarnoff, then contracts manager to the American Marconi company, recommended that transmitting stations be built for the purpose of broadcasting speech and music and that "a radio music box" should be manufactured for general sale. Sarnoff wrote,

> a radio music box . . . this device must be arranged to receive on several wavelengths with the throw of a switch or the pressing of a button. The radio music box can be supplied with amplifying tubes and a loudspeaking telephone, all of which can be neatly mounted in a box.

Sarnoff's proposals were not implemented at once, partly because of America's entry into World War I in 1917, but they proved highly practical. Technically, the first entertainment broadcasting was done by the German Army in 1917, but the first regular broadcasting station was KDKA in Pittsburgh, which started operations in November 1920. The service gained instant popularity, and the idea spread at once around the world (see BROADCASTING).

The medium-wave band of frequencies proved inadequate in both Europe and America to provide sufficient broadcast channels. As a result, there was considerable development of the very-high-frequency band after World War II. The greater availability of channel space in the very-high-frequency band allowed the use of frequency-modulation techniques (see below) for broadcasting stereo (two-channel) programs so coded that the one transmission can provide monaural listening from the normal type of frequency-modulated receiver and binaural (stereo) listening from a receiver fitted with a special decoder.

Shortwaves are also used for broadcasting, and the lower frequencies (about four megahertz) are used in the tropics for local broadcasting. These have some advantage over medium waves during the tropical rainy season, when atmospheric noise due to thunderstorms is serious; but the signal must be obtained by almost vertical reflection from the ionosphere, and fading and distortion are generally greater than with medium waves.

Long-distance broadcasting (unless a satellite is employed) is only possible by means of shortwaves, because the shortwaves are reflected back to Earth from the ionosphere. Groups of channels are allocated throughout the shortwave band for the purposes of external broadcasting beyond the national boundaries of the originating authority. The number of channels available is now insufficient for the countries wanting to broadcast, and, in spite of a degree of international control, interference is often a serious problem, particularly on the lower frequency channels at night when more channels are in use because higher frequency channels tend to be unusable due to reduced ionospheric reflection.

Broadcasting from a synchronous satellite (one that maintains a relatively fixed position with respect to Earth) is a possibility, but power-supply and antenna-size limitations on the satellite require the carrier frequency to be in the ultrahigh-frequency range. The distance of the satellite from Earth, about 22,300 miles (36,000 kilometres), results in the signal at the Earth's surface being very weak. A sensitive receiver and directional antenna system is therefore essential. Another disadvantage is that over the area of Earth covered by the radio beam from the satellite antenna, the particular carrier frequency cannot be used by any other transmitters.

RADIO CIRCUITRY

Components. The basic operating principles of the major circuitry and passive components used in radio are described in the article ELECTRONICS. Active devices such as tubes and transistors are treated in the articles ELECTRON TUBE; and SEMICONDUCTOR DEVICES. In this article, only enough description is included to permit the reader to understand the applications to radio circuitry.

Primary function of the electron tube

Active devices: *vacuum* tubes and transistors. An electron tube or transistor, designated an active element, functions basically as an amplifier, and its output is essentially an amplified copy of the original input signal. The simplest amplifying electron tube is the triode, consisting of a cathode coated with material that provides a copious supply of electrons when heated, an open-mesh grid allowing electrons to pass through but controlling their flow, and a plate (anode) to collect the electrons. The plate is maintained at a positive voltage with respect to the cathode in order to attract the electrons; the grid usually has a small negative voltage so that it does not collect electrons but does control their flow to the plate. The output voltage is usually many times greater than the input voltage to the grid. The tube must be pumped to a high degree of vacuum, or the plate current flow is very erratic.

Other electrodes, also in the form of open-mesh grids, may be included in the tube to perform various special functions. An example is the four-electrode tube known as the tetrode, in which an open-mesh grid (screen grid) maintained at a positive voltage is placed between plate and control grid. This reduces the effect of plate voltage on electron flow and increases the amplifying property of the tube. Introduction of a third grid, known as a suppressor grid, produces the pentode (five-electrode tube), which can provide even greater amplification.

The transistor, which has largely replaced the electron tube as the active element in low-voltage electronic circuits, is made from semiconductor materials—that is, substances that are neither good conductors nor good insulators. Two common semiconductor materials are germanium and silicon, to which small amounts of impurities such as indium, gallium, arsenic, or phosphorus are added to impart electrical charges to them. Arsenic and phosphorus, for example, provide extra negative charges, giving n-type (signifying excess negative charges) material; indium or gallium yield a shortage of electrons or an excess of positive charges or holes, giving p-type (signifying excess positive charges) material (see also SEMICONDUCTOR DEVICES).

A transistor is a sandwich of semiconductor materials with the same impurity in the two outer layers and a different impurity in the centre layer providing current carriers of opposite charge to those produced by the outer layers.

If the outer layers are reservoirs of positively charged current carriers (p type) and the centre layer provides an excess of electrons (n type), the transistor is known as a *p–n–p* (positive–negative–positive carriers) type. If the p and *n* layers are reversed, the transistor is an n–p–n type. The two outer layers are termed the emitter and collector, and the centre layer is known as the base.

A transistor is an amplifier of current; the vacuum tube, in contrast, is an amplifier of voltage. The transistor produces a very adequate supply of current carriers (electrons and holes) at room temperature and does not require a heated cathode like the vacuum tube. Thus the power required from the power supply is much reduced, much less heat is produced, and the transistors and their circuitry can be packed into a smaller space. Transistors are also physically much smaller than comparable electron tubes. Thus the transistorized portable radio can fit in a pocket in contrast to the cumbersome tube radio it has replaced.

In its early form the transistor was capable of amplifying only comparatively low frequencies because the exchange of electrons and positive charges across the sandwich was slow. Modem techniques however, have overcome this difficulty so that amplification up to frequencies over 1,000 megahertz is commonplace.

Modulation

Tuned circuits and the superheterodyne principle. For information (voice, music, television) to be transmitted, it must be attached to a radio-frequency carrier wave, which is then transmitted in a given frequency channel. The carrier wave and information can be picked up by a receiver tuned to this channel. The process by which the information is attached to the carrier wave is known as modulation. Modulated carriers are isolated in their separate slots or channels; if transmitters are geographically close to each other, they must not use the same channel or overlap each other's channels. If such overlap occurs, serious interference results—two radio programs may be heard simultaneously or one may form a distorted background to the other.

In most modern radio receivers, reception is based on what is known as the superheterodyne principle. The incoming radio frequency is mixed (heterodyned) with the output of an oscillator the frequency of which is adjusted so that the difference between it and the incoming signal is constant; the result is called the intermediate frequency. Amplification is thereafter carried out at this intermediate frequency. Both preliminary selection of the incoming frequency and adjustment of the local oscillator frequency are accomplished by variable tuned circuits consisting of inductance and capacitance. Tuning may be accomplished by varying the capacitance, which consists of interleaved metal plates separated by air spaces with one set of plates movable. Another method of tuning involves varying the inductance by insertion or withdrawal of an iron dust or ferrite (*q.v.*) core in a cylindrical coil of copper wire. To simplify the tuning procedure, the variable elements of all stages requiring tuning are ganged together and coupled to a tuning knob. The intermediate-frequency amplifier stages always operate at the same frequency and so require tuning only when the receiver is manufactured or serviced. The intermediate frequency is usually around 455 kilohertz for AM receivers, 10.7 megahertz for FM receivers, and 38 megahertz for television receivers. Most of the amplification in a radio receiver is carried out in the intermediate-frequency stages, and most of the selectivity (ability to separate adjacent stations) is obtained in these stages.

Oscillators. A self-oscillating circuit consists of a vacuum tube or transistor, a tuned circuit, and some form of positive feedback (energy fed from the output back to the input in such a way as to increase the input). Since both tubes and transistors can function as amplifiers, they can also function as oscillators. For receiver circuits, adequate oscillator stability can be obtained with conventional tuned circuits, but the transmitter oscillator must be highly stable, and a circuit made up of inductance and capacitance, tuned to the desired frequency, is not sufficiently stable. A piezoelectric crystal oscillator (a device that vibrates or oscillates at a given frequency emitting radio waves when voltage is applied to it) or its equivalent is ordinarily used.

Amplifiers. Amplifiers may be classified in a number of different ways; according to bandwidth (narrow or wide); frequency range (audio, intermediate or radio frequency); or output parameter requirement (voltage or power).

Wide-band radio-frequency amplifiers are not needed for audio signals unless a frequency-modulated system is used. Amplitude-modulated signals for sound broadcasting should have a radio-frequency bandwidth of ±10 kilohertz though on medium waves it is often limited to about ±5 kilohertz (total bandwidth of 10 kilohertz). High-quality frequency-modulated audio needs a bandwidth of about ±100 kilohertz.

Use of negative feedback to minimize distortion

Audio-frequency amplifiers present few design problems, and negative feedback of the output into the input can overcome distortion problems. Radio-frequency amplifiers, which can be tuned, suffer from variation of selectivity (ability to separate adjacent stations) and gain (amplification) over the tuning range. Selectivity tends to broaden and gain to increase as capacitance is decreased, and instability can be troublesome at the highest tuning frequency. Intermediate-frequency amplifiers do not suffer from these defects since the tuning frequency is fixed.

The main problem with radio-frequency amplifiers. in receivers is the possibility of cross modulation—that is, the mixing of two information channels, which can occur if an undesired modulated signal enters the radio-frequency input together with the desired signal.

High-power amplifiers are not required for radio reception but are needed for radio transmission. Semiconductors are not as yet able to fill the high-power role. Vacuum tubes are essential for providing the required kilowatts of radio-frequency power to the antenna.

Antennas. The antenna is an essential part of a radio transmission and reception system (see ANTENNAS AND WAVE GUIDES). Its purpose at the transmitter is to project electromagnetic energy into space and at the receiver to extract energy from the travelling electromagnetic wave produced from the transmitter antenna.

The size of the antenna relative to the wavelength of the electromagnetic radiation is important. The wavelength of medium waves is about 300 metres (1,000 feet), and a vertical transmitting mast or self-supporting tower 150 to 210 metres (490 to 690 feet) high may be used with a high-power transmitter (200 kilowatts or more). An equally tall receiving antenna would be ideal but impractical. A vertical rod or suspended wire about six metres (20 feet) long is a workable solution. If the transmitting antenna is vertical, the receiving antenna must also be vertical; if the former is horizontal the receiving antenna must be horizontal. This rule applies at all radio frequencies except shortwaves because the plane of their electromagnetic field can be twisted in its passage through the ionosphere, and a vertical shortwave antenna may pick up a good signal from a horizontal transmitting antenna. The antenna system becomes progressively smaller as the transmitting frequency is increased, and at ultrahigh frequencies (300 megahertz or more) the individual antenna may be only about 50 centimetres (20 inches) long.

For normal amplitude-modulated broadcasting, the receiver antenna may be composed of a fairly short coil of wire wound on a powdered iron or ferrite core. This type of design permits adequate signal pickup with a very small antenna which may be located in a small space, a necessity for small, battery-operated portable receivers.

Directional characteristics

Antennas may have their directional characteristics modified by employing multiple elements. Thus an antenna may be omnidirectional (transmitting in all directions) horizontally but narrowly beamed vertically, or it may be bidirectional (transmitting in two directions) in a figure eight pattern with two main directions of energy projection at 180". It may be unidirectional, having energy projected to one side, or the energy may be concentrated in a relatively narrow beam both horizontally and vertically.

In point-to-point communication, as from one network centre to another, highly directional antennas are used. Local broadcasting uses an omnidirectional antenna, radiating equally in all directions, except in such special cases as a coastal location or proximity to a neighbouring transmitter.

Broadcasting planned to serve distant areas, employing shortwaves and depending on reflection from the ionosphere, normally uses a relatively narrow beam of energy projected skyward at an angle from 5° to 10° to the horizontal. A reflecting curtain is placed behind the antenna to prevent loss of energy in the reverse direction. The beam is divergent (spreads out) so that after two or three reflections between ionosphere and Earth it covers a relatively large area.

Transmission lines. The lines that carry radio waves from the radio transmitter to the antenna are known as transmission lines; their purpose is to convey radio-frequency energy with minimum heating and radiation loss. Heating losses are reduced by conductors of adequate size. Only the outer layers of the conductor carry radio-frequency current.

Concepts of selectivity and sensitivity. Radio-frequency communication requires the receiver to reject all but the desired signal. Were the number of frequency channels equal to the demand, each channel could be given its correct width in the tuning stages of a receiver. Thus, for audio broadcasting each carrier channel should be 20 kilohertz wide to accommodate both side bands, and each transmission carrier should be 20 kilohertz, separated from those on either side. In much of the world, the medium-wave and shortwave bands are in such demand that transmitters must share the same channel and channels thus must overlap. Though efforts have been made to arrange sharing by geographically separated transmitters, the congestion has forced receiver manufacturers to reduce the receiver bandwidth to about eight kilohertz (±four kilohertz).

Very-high-frequency transmissions can generally be received at full bandwidth because their signals are confined to line of sight and are, in effect, local-station signals to the receiver. Frequency-modulated transmissions must be received on full bandwidth (about 200 kilohertz) if serious distortion is to be avoided on loud programs.

Receiver sensitivity is the ability of a receiver to pick up weak signals. Though a communication receiver should always have a high sensitivity, there is a maximum determined by the noise generated inside the receiver itself. Little value is gained by increasing sensitivity if noise at the receiver output is already considerable and comparable with desired signal output. Normally, radio broadcasting systems operate with the signal voltage at least 10 to 50 times greater than the noise. To take full advantage of high sensitivity, receiving antennas for communications links are usually located in an area where there is little man-made noise. A receiver intended only for local-station reception can have a much lower sensitivity than a shortwave receiver intended for picking up signals from the other side of the world.

Effect of noise on sensitivity

APPLICATIONS

Applications of radiotelephony. The major use of radio is in the communications industry (see also TELEPHONE AND TELECOMMUNICATIONS SYSTEMS). Telephone, telegraph, and television signals are transmitted by microwave-radio (above 1,000 megahertz) systems, either between land-based towers or via satellite, and shortwave radio is used for overseas circuits. Radio broadcasting (AM, FM, TV) is a worldwide industry of tremendous scope, and mobile radiotelephone service to moving vehicles such as automobiles, boats, and aircraft has become a major world industry.

Development of commercial radiotelephone. Until the early 1920s all commercial radiotelephone developments had concentrated on the use of long-wave carriers because these ensured worldwide communication. Medium-wave transmissions were known to have a comparatively short service range during the day, but even at night when their range was greatly increased their signal strength decreased at a much greater rate than that of long-wave carriers. It was consequently assumed that the longer the wavelength the better the propagation. A short wavelength of 120 metres (2.5 megahertz) had been employed for communicating between ships from 1901 to 1909, and it had been noted that reception over distances of 1,000 miles (1,600 kilometres) or more could be obtained, but this was thought to be caused by freak conditions.

Marconi experimented with shortwaves in 1916 because the shorter wavelength permitted directional antennas to be used to concentrate the transmitted energy into a narrow beam. Only the direct wave launched parallel to the Earth was utilized, however.

Radio had already attracted numerous amateur enthusiasts and experimenters, and at the end of World War I they were granted permission to communicate by radio with each other at the upper end of the medium-wave band and in the shortwave band, frequencies then thought to be useless for long-distance links. Nevertheless, amateurs in Britain and the United States hopefully arranged a series of test transmissions on 230 metres (1.3 megahertz) during December 1921, and contact was established on several occasions. Other amateurs maintained contact across the Atlantic during daylight throughout most of February 1925 with the aid of shorter wavelength transmissions in the order of 100 metres.

Tests by amateurs

Meantime Marconi had also been experimenting with shortwaves and had been able to pick up signals from Cornwall, England, at the Caribbean island of St. Vincent, 2,300 miles (3,700 kilometres) away, using a one-kilowatt transmitter of about 100 metres (three megahertz) carrier wavelength. It was found that short wavelengths could be used to provide a satisfactory service at great distances, and daylight contact was maintained on 32 metres (9.8 megahertz) with Australia. On the basis of these tests, the Marconi company constructed links between Britain and Canada, Australia, India, and South Africa.

The Canadian circuit was opened in October 1926; the others followed at regular intervals during 1927. Marconi's faith in the successful commercial operation of the system was more than justified, and radio engineers elsewhere were quick to change from skepticism to enthusiasm. Shortly after the start of the radiotelegraph transmissions, the Marconi company developed a method (called multiplex) of transmitting two telegraph channels and one telephone channel on the same carrier wave, but this was not taken up by the British government, which had decided to preserve a monopoly in telephone communication. Meantime, the Radio Corporation of America, which in September 1924 had tried out a 100-metre service to South America as an auxiliary to long-wave transmission, began to test the efficacy of shortwaves for commercial speech circuits. The U.S. Navy and Bell Telephone Laboratories also began research into shortwave propagation. In Germany the Telefunken Company set up stations early in 1925 to transmit to South America.

Shortwaves are now a well-established means of commercial telephonic communication, and sufficient information has been gained over the years to develop frequency schedules allowing a regular service to be maintained at any time of the day between any two populated points on Earth, except on the comparatively rare occasions of severe ionospheric storms.

Attention was subsequently paid to even shorter wavelengths, and in 1931 Standard Telephones and Cables set up a commercial circuit across the Straits of Dover using an ultrahigh-frequency wavelength of 18 centimetres (1,700 megahertz).

Exploitation of all radio-frequency ranges has steadily continued since the 1930s, and parts of the very-high-, ultrahigh-, and superhigh (microwave)-frequency spectrum are allocated to commercial communication. As a general rule, these frequency ranges can provide a satisfactory service only over a line-of-sight path. There is some bending of the transmitted wave as it travels over the Earth, however, and transmission is sometimes carried to distances beyond the actual optical range. The longer wavelengths of the very-high-frequency range can be used for longer distance communications (up to 300 miles [480 kilometres]) by employing a technique called "tropospheric scatter," in which high-powered radio beams are scattered from the upper layers of the troposphere.

Line-of-sight path

Satellite communication has greatly extended facilities because microwave transmission may be used, and these high frequencies allow a very large number of conversations to be transmitted on a single carrier. Microwave frequencies in the range three to 20 gigahertz suffer little loss of power in penetrating the ionosphere and their very short wavelengths allow the construction of highly directional small-sized antennas, which concentrate the radio energy into a very narrow beam (see also SATELLITE COMMUNICATION).

Long-distance intercontinental radiotelephone service. Until Marconi had proved the worth of the shortwave-beam system, communication engineers had continued to develop radiotelephony on long-wave transmissions, and there were a number of such systems operating between 1920 and 1930. The limited number of channels available in the long-wave band and the large antenna systems and high power output needed from the transmitters made it clear that development could not go far. The discovery of shortwave propagation created new possibilities, and the 1930s saw the rapid growth of a worldwide system of shortwave telephonic communication. In October 1929 the International Telephone and Telegraph Corporation set up a circuit from Madrid to Buenos Aires, and in 1930 further shortwave telephonic circuits were established among London, Sydney, Buenos Aires, and ships at sea and from Buenos Aires to Paris, Berlin, and New York. The years 1932 and 1933 brought the radiotelephone to South Africa and India.

The need for economic operation, conservation of available frequency channels, and better rejection of noise led to the development of single-side-band pilot-carrier systems. The transmission of a carrier as well as side bands simplifies the design of the receiver but represents a considerable expenditure of power at the transmitter. When the transmission is modulated to its maximum extent, for example, the speech content is only one-third of the total power supplied to the antenna. The economics of point-to-point communication justify the installation of a complicated and expensive receiver if the saving on the transmitter is greater than the additional receiver cost; this occurs if the carrier is generated in the receiver itself. The frequency of the generated carrier must be very accurate if distortionless detection is to be achieved; a small degree of carrier signal is transmitted and used to synchronize the receiver-generated carrier frequency. Hence the term pilot carrier.

All the speech-signal information on such a system is present in one of the two side bands; elimination of one side band halves the frequency channel width as well as improving the strength of the signal with respect to noise. It is possible to use the eliminated side-band frequency space for another speech channel so that one transmission carries two different conversations simultaneously. This is an independent-side-band pilot-carrier circuit.

Greatly improved performance as regards signal-to-noise ratio, fading, and distortion on shortwave telephone links resulted from a British Post Office invention called Lincompex, in which speech variations are compressed electronically into a narrow volume range to maintain modulation at a high level; a pilot signal containing information on the degree of compression is also transmitted with the speech. At the receiver an expander, the gain of which is controlled by the pilot signal, restores the original speech volume variations.

Earth satellites

The successful placing of satellites into prescribed orbits around the Earth has made possible the realization of a revolutionary communications system. A satellite placed in an equatorial orbit of the Earth at a height of approximately 22,300 miles (36,000 kilometres) above the Earth's surface maintains a fixed position relative to the Earth. In any other orbit there is relative motion between Earth and satellite, and ground-station antenna systems must keep the satellite in sight and follow its movement by means of a complicated tracking system. Frequencies above those affected by the ionospheric layers (*i.e.*, very high frequencies and above) can be used for satellite communication; superhigh frequencies at about six gigahertz (to the satellite) and four gigahertz (from the satellite) are employed because small, highly directional antennas can be constructed on the satellite at these very short wavelengths (7.5 to 5 centimetres). The satellite must be accurately stabilized to prevent "swinging" of the antenna system since this would lead to appreciable variations of signal strength at satellite and ground station.

Ship-to-shore service. The first recorded radiotelephone contact from ship to shore was made in 1916 by the United States Navy. Tests of long-wave low-power transmissions took place in 1920, and Marconi installed the first commercial equipment in 1922. During the 1920s, progress continued and expanded to include lightships and harbour craft; the British lifeboat service was fitted with radiotelephone apparatus beginning in 1926.

In 1927 a shortwave telephone link on the British liner "Carinthia" established contact with Britain from the Pacific, halfway around the world, and all passenger ships were soon equipped with radio for passenger service. The radiotelephone was adopted by fishing fleets during the 1930s because it had the advantage of not requiring a skilled operator with knowledge of Morse code. After 1934 many small coastal craft were equipped for radio contact with the shore. After World War II radiotelephone systems for ship-to-shore communication multiplied, communication over distances up to about 70 miles (112 kilometres) being provided by very high frequencies and over greater distances by shortwaves.

Passenger ships equipped with radio

Mobile radiotelephone. Communication with moving vehicles is an important application of radiotelephone. Frequency allocations for this use have been made at the top end of the medium-wave band, at a number of points in the shortwave band from two to five and from about 23 to 27 megahertz, and in the very-high-frequency band. The latter band has the advantage that wavelength is small so that antennas need not be large, and the attenuation (loss of power) with increasing distance from the transmitter is very much less than for shortwaves with a given height of antenna from ground, but it is more susceptible to loss of signal due to intervening obstacles and to multiple signals by reflection from tall buildings.

Broad-band microwave radio relay. A large number of speech channels of four kilohertz (the normal channel width for commercial speech) can be impressed on a single radio carrier by proper modulation and multiplexing techniques. These techniques are used in broad-band cable and also microwave transmission systems operating with from 12 to 960 channels; successful transmissions have been carried out with as many as 2,700 channels.

Microwave frequencies from three to 12 gigahertz are employed for line-of-sight transmission; their antennas must be well clear of the ground on towers not less than 150 feet (46 metres) high, and the direct radio path must clear the summit of any intervening hill by at least 50 feet (15 metres). Spacing of the towers depends on intervening terrain but averages around 30 miles (48 kilometres).

Dish antennas, generally parabolic in shape, or horn antennas are used at each end of a link. Amplification may be provided in a so-called head amplifier very close to the centre of the dish, though to reduce maintenance problems at very exposed sites it is customary to provide the amplification at the base of the mast and a connection made to the antenna by cable or wave guide. In some instances a 45° reflector is used at the top of the mast to reflect the energy downward to an antenna dish placed at the foot of the mast.

Emergency radio service. Before radio, ships' distress signals were limited to optical distances. One of the first important applications of radio was for broadcasting the maritime distress signal, the morse coded SOS. For the signal to be effective, all ships' receiving equipment had to be manned continuously 24 hours a day, and in 1927 it was made obligatory for all ships not maintaining a permanent radio watch to carry an automatic alarm receiver tuned to 500 kilohertz, the internationally agreed marine distress frequency.

A similar provision is made for aircraft to call on a distress frequency of 121.5 megahertz, but it is more usual to call "Mayday" (French *m'aidez*, "help me") on the frequency already being used for communication with the airport controllers.

Distress signals

Applications to marine and air communications. ***Early developments in the airline and maritime field.*** Maritime radio communication has now become the main method of shipping control by owners and port authorities, **a** vital disabled-ship warning device, and an aid in dealing with injury or sudden illness. Radio is also used for weather warnings.

The development of passenger air traffic stimulated the need for radio communication between plane and ground station, for landing and takeoff instructions, and for exchanging positional data.

Radio direction finding was the earliest method of using two known sites to determine the position of an unknown transmitter, a principle used by the British during World War I for keeping track of German fleet activities. A ship can determine its position by direction finding when two or more transmissions are available from separated sites; this facility is especially valuable in conditions of poor visibility. In World War II direction-finding equipment was standard in bombing planes.

Frequencies used. Frequencies have been allocated in the various radio transmission bands for marine and air communication as well as navigation. Maritime communication has been given allocations in parts of the long-wave, medium-wave, and shortwave bands as well as in the very-high- and ultrahigh-frequency bands. Aeronautical communication has been allotted a small section at the top end of the long-wave band, is well distributed throughout the shortwave band, and has two allocations in the very-high-frequency band. No allocations are made above 144 megahertz.

Radio navigation is allocated a large number of frequencies in the long-wave band, a very small section of the medium-wave and shortwave bands (at the lower frequency end of each), and some parts of the very-high-, ultrahigh-, and superhigh-frequency bands. Radar functions only in the ultrahigh- and superhigh-frequency bands.

Radio aids to navigation. Radio-navigation systems can operate from a moving vehicle, aircraft, or ship, giving the pilot or captain a sound or visual indication of his position relative to a fixed object; or they may operate from a given site, the information being transmitted to the moving vehicle. Both systems may be necessary for safe navigation.

Navigation aids are classified into those that provide direction but not distance (direction-finding equipment) and those that give continuous readings of position through laying out over the surface of the Earth a radio-frequency grid of identifiable electrical parameters by combination and analysis of signals, generally from three fixed slave stations, all controlled from one master station. The loran (long-range navigation) and Decca systems are examples of this type of navigational aid. Finally there is radar, which can give distance, azimuth, and elevation (see also NAVIGATION; RADAR).

Use of loop antenna

The radio direction finder consists of a highly directional loop antenna, which is either rotatable or has circuits which simulate rotation, together with a sensitive receiver. Transmissions from two geographically separated points and a knowledge of true north are needed in order to fix position. The directional-antenna system is rotated until the transmitted signal is minimum or zero (this is called the null position), and the angle of the antenna relative to true north is noted. The bearing of the transmitting site is on a line at right angles to the null position. Ambiguity results, however, because there are two null positions at 180" to each other. The rotatable antenna has a figure-eight pattern of reception; to determine on which side the transmitter is located, it is necessary to add the signal from an omnidirectional vertical wire antenna. The correct addition of the two produces a cardioid (heart-shaped) pattern with a null pointing to the true transmitter bearing.

Although the line of direction to the transmitting station can be drawn on a map, a bearing on another known transmitting station must be taken as the intersection of the two bearings determines true position.

Direction finding, which dates from 1911, is the simplest method of determining position, and a radio transmitter at a site near a seaport or airport can serve as a beacon line by which to steer. There are two disadvantages: time is needed to obtain position from the bearings, and reflections from terrestrial objects and the ionosphere as well as refraction (bending of radio signals) at the seacoast can cause bearing errors.

The loran system of long-range navigation by radio provides positional information of high accuracy from a measurement of the time difference between two pairs of pulse-modulated signals transmitted on different carrier frequencies from specially selected sites. The carrier frequencies are at the top end of the medium-wave band from approximately 1,700 to 2,000 kilohertz, and coverage is about 700 miles (1,100 kilometres) by day and 1,400 miles (2,300 kilometres) at night.

The so-called Decca system employs a master and three slave stations, all sending continuous waves and placed at the corners of an equilateral triangle of about 200-mile (320-kilometre) sides. The master is located near the centre of the triangle. Only two slaves are used at one time for information on bearing, and Decca charts tell the user which ones to tune to. The slave transmitters receive their basic frequency from the master. The frequency is adjusted so as to have the same time phase at all slave stations. Interaction between frequencies from each pair of slaves produces a relationship dependent on the difference in distance. A special circuit and an instrument called a decometer accurately determine the relationship and register it with the aid of rotating pointers. The reading on the pointers can then be translated to position by consulting the Decca chart. The decometer pointers can be reset if for any reason the apparatus has been switched off or there has been a transmission fault.

The transmissions are in the long-wave band, and satisfactory operation is obtainable to a distance of about 240 miles (380 kilometres) from the master station.

The great advantage radar possesses over direction finding and loran is that it requires no cooperation from shore, other ship, or ground station.

Radar apparatus measures the time elapsing between the sending and reception of a pulse of radio energy reflected from an object. The pulse duration is longer or shorter for longer or shorter ranges, typical values being 0.1 to 0.2 microsecond or millionths of a second, and repetition frequency is in the range of 500 to 4,000 pulses per second, being higher for short ranges. Microwave frequencies are employed so that very narrow beam antennas can be constructed.

Radio altimeter

An aircraft needs to know its height above ground, and this can be obtained from a radio altimeter that measures the time taken for an outgoing radio pulse projected vertically downward to be reflected from the ground back to the plane.

Applications of radiotelegraphy. ***Early commercial history.*** Radiotelegraphic communication represents one of the simplest methods of conveying information because it is an on–off system. When undertaken manually it is much slower than speech, about 20 words per minute being the maximum speed. Faster communication can be obtained by machine operation; a teleprinter can deal with *75* words per minute and a tape puncher with about 400 words per minute. An advantage of automated transmission and reception is that a record of the message is available (see TELEGRAPH).

The first radiotelegram for which payment is recorded was transmitted in June 1898 on the occasion of a visit by Lord Kelvin to Marconi's experimental station at Alum Bay, Isle of Wight (England). Kelvin was so impressed by the transmission that he insisted on paying for the telegrams even though he knew that a financial transaction was illegal by the terms of the British Post Office's monopoly.

By March 1915 there were 706 coast stations and 4,866 ships throughout the world fitted for radiotelegraphy. Radio was soon made compulsory on ships carrying 50 or more passengers, and research followed by new design improved accuracy and speed of telegraphic messages.

Trend of radio transmission since 1900. Early radiotelegraph transmission was by manual operation of a Morse key and by aural and visual monitoring and translation of the message. Though the teleprinter speeded, and automated, the operation, present mechanical limitations prevent speeds beyond about 100 words per minute. Paper tape and magnetic tape allow records to be made of much higher speed signals.

Effect of distortion on speed

Transmission and reception speeds of telegraph signals are limited primarily by distortion occurring over the propagation path and only secondarily by the bandwidth that can be made available. In this respect, microwave transmission is superior to shortwave radio or even cable links since the carrier frequency is very high and a wide bandwidth is still only a small proportion of the carrier frequency. Shortwave radio suffers from selective fading of the signal-frequency components (worse for high-speed operation requiring a greater transmission bandwidth), from comparatively low carrier frequencies, and from an overloaded frequency spectrum due to too many users. Cable transmission over very long distances sets a limit on bandwidth and can distort the shape of the telegraphic signals.

Radio transmission, particularly on shortwaves, tends to be more susceptible to distortion, noise, and interference than land lines, and an important development for ensuring message accuracy was an automatic error-correction system pioneered during World War II.

Spark transmitters were originally employed for telegraph signalling and continued in service long after the electron tube had taken over for speech transmission. Spark signals caused considerable interference outside the channel carrying the telegraph code because of the broad spectrum occupied by their frequency components, and in 1941 their further use was banned by international agreement.

Modern radiotelegraph practice. Modem radiotelegraph practice employs a seven-unit code to transmit telegraph characters by frequency–shift keying (in which the frequency of the carrier is shifted for mark or space transmission) with single-side-band, independent-side-band, or double-side-band transmitters. It is possible to send several separate telegraph signals in the same frequency spectrum as one speech channel. Thus a telegraph signal having 50 intervals per second (one interval equals one on–off period and is known as a baud) requires a bandwidth of 100 hertz; and 24 of these channels can be transmitted in the same bandwidth as a speech channel. If channels at higher speeds must be accommodated, the number of channels must be proportionally reduced.

Amateur and professional radio operation. ***History of amateur radio.*** From its beginnings, radio communication has attracted armies of amateur enthusiasts all over the world. By 1912 these were numerous enough to require control measures to avoid interference with commercial signals. Control was first instituted in 1911; during World War I there was a ban on amateur activity in belligerent countries, and receiving sets were sealed.

After the war, amateur activities in the medium-wave band were permitted only at the top end (above about 1.3 megahertz) but were given complete freedom in the shortwave band. Experimentation by amateurs soon revealed the unexpected favourable propagational qualities of shortwaves. An American amateur, in association with British amateurs in 1921, installed a receiver in Scotland with which transatlantic contact was made on 1.3 megahertz from a one-kilowatt transmitter 3,200 miles (5,100 kilometres) away. Radiotelegraph signals were exchanged over considerable distances using low-powered transmitters operating at 180 metres (1.7 megahertz) and later at 100 metres (3 megahertz). In 1924 shortwave transmissions were sent from England to New Zealand.

Amateurs are allocated frequencies at the extreme high-frequency end of the medium-wave band, five groups of frequencies in the shortwave band, two groups in the very-high-frequency band, three in the ultrahigh-frequency band, and seven in the superhigh-frequency band for telegraphic and telephonic communication using amplitude and frequency modulation. There are restrictions on the power of the transmitters, and certain of the frequencies must be shared with due regard for the needs of other users. Experimental transmissions still being made seem unlikely to yield such spectacular results as those of the 1920s.

Military use of radio. During World War I relatively little use was made of radio except in naval conflict, though some artillery spotting aircraft were fitted with transmitters late in the war. World War II saw a vast expansion of radio communication, with constant contact maintained between forward units, including mobile units such as tanks, and headquarters. All messages likely to be intercepted are normally coded, though in certain combat situations voice messages are sent in clear.

All radio-frequency bands may be brought into service for military purposes. Long-distance communication in World War II depended entirely on shortwaves, but the development of satellites .has made possible, microwave communication with greater reliability and less message error because of the elimination of multipath difficulties.

One of the main problems with military radio communications is the possibility of enemy jamming and interference. Very highly directive antenna systems, the use of meteor trails for reflecting signals, and various detection techniques are helpful in reducing the effects.

Radio also plays an important part in guiding aircraft, in artillery aiming, in remote control of missiles, and in surveillance of enemy activity by radar and by satellite observation.

Military equipment, especially that for combat use, is rugged, simple to operate, and as light in weight as possible. It requires an adequate supply of spare components on hand since proper maintenance can be undertaken only at base workshops.

Government and civil authority radio operation. Radio is an essential instrument of national and, to a lesser extent, local government all over the world. Microwaves play a vital role in national telephone, telegraph, and data links, and shortwaves in communication with embassies abroad; satellite links are becoming more important in this field.

Broadcasting is not only a medium for government announcements but for educational purposes, especially in the underdeveloped countries that are short of qualified teachers. It can be a potent weapon in maintaining or undermining morale and is widely used for propaganda and psychological-warfare purposes. Many countries have resorted to jamming unwelcome foreign broadcasts, even at the cost of obliterating a number of communication channels.

To collect and assess news from all parts of the world, many governments use news-agency-teleprinter-radio services and monitoring posts.

Civil authorities require radio communication for the control of fire, police, ambulance, and traffic services and for dealing with emergencies caused by natural disasters such as floods, hurricanes, and earthquakes. The very-high-frequency band is most useful.

Shortwave listening. Shortwave transmission plays a very important part in external broadcasting beyond the land frontiers of many nations. Developed nations use this method of keeping in touch with their nationals residing abroad as well as for ideological purposes.

The disadvantage of shortwave radio is that the signal is highly variable because of its dependence on reflection from the ionosphere, and it is often distorted. Speech is generally at least intelligible, but music may be distorted out of all aesthetic value. A further disadvantage is that the frequency giving optimum listening conditions varies over the 24 hours, the season of the year, and the sunspot cycle. Experience suggests that listeners commonly prefer to listen to the same frequency, even though the signal is deteriorating, rather than retune to another frequency. This is due partly to natural laziness and partly to the difficulties of tuning and of maintaining correct tuning. The shortwave band itself possesses 1,500 channels 10 kilohertz wide, whereas the medium-wave band has only 100. If both bands are covered by one sweep of the tuning scale, it will clearly be 15 times more difficult to tune a shortwave signal. This problem is reduced by dividing the wave band into three or more sections, or better still by using band spreading, in which a band encompassing about 200 kilohertz is given the full sweep of the tuning scale. Receivers providing this facility are much more expensive.

Band spreading

Special applications of radio. ***Remote-control radio.*** There are many circumstances in which control and monitoring signals cannot be carried by wires; examples are moving objects such as a model boat, an aircraft making a blind landing, or a spacecraft being placed in orbit. Radio control can be used in these circumstances. Coded signals for control are used to modulate a carrier frequency, and the moving object has a receiver permanently tuned to this frequency. The control may be a simple on–off signal operating a relay to switch the motor of a model boat on or off, or it may be a series of coded signals controlling the altitude, azimuth, elevation, and speed of an aircraft making an automatic or a blind landing.

The remote operation of radio apparatus is now well-established practice, and many commercial installations for telephony and telegraphy transmitting and receiving as well as broadcasting transmitters are remotely controlled by signals sent along wires. The operator can select the desired carrier frequency, the telegraph, speech, or music channel, and the antenna and can adjust the tuning of transmitter radio-frequency stages. Receivers can be remotely controlled from a single control desk, which has means for selecting the desired incoming signal and the correct antenna, as well as for tuning the circuits, adjusting output level, and directing output to any required destination.

Broadcasting transmitters may be time-switch operated or switched and monitored from a main station; monitoring may involve checking all program quality conditions and providing warning that these are unsatisfactory or that the transmitter is off. There are also interrogating systems giving the operator information on the situation at specified points in the apparatus. Unattended transmitters generally have their most vulnerable circuits such as the output circuit duplicated so that failure of one leaves the transmission still functioning, though frequently at reduced power.

Radio paging. The efficient employment of many types of personnel besides police is enhanced by an effective means of signalling an emergency or change of plan. A doctor may carry a lightweight portable alarm device operated by induction from wires built into the ceiling or walls of hospital wards. The frequency used may be in the audio or low-radio range and the receiver sharply tuned to accept the selected paging frequency. This system can only be used in confined areas, and coverage of a large area generally requires speech communication using amplitude or frequency modulation of carriers in the very-high- or ultrahigh-frequency bands. If a "return to base" or "call base by the nearest telephone" is all that is needed, a simple alarm can be employed and made selective to any number of individuals.

Citizen's band radio. Citizen's band radio involves the use of radio for business and personal communication or remote-control purposes over 10 to 15 miles (16 to 24 kilometres). Channels are made available for voice communication and remote control, such as opening of garage doors, at frequencies around 27 megahertz and 465 megahertz, and a band centred on 75 megahertz is available for model-aircraft control. There were over 1,000,000 licenses in force in the United States alone in 1967, and the government insisted on stringent conditions to avoid excessive local interference. Maximum transmitter power and antenna height are fixed; personal calls are permissible, for example, from a remote campsite or in case of emergency, only if a satisfactory telephone service is not available, and calls are limited to a maximum of five minutes. Each user is given a call sign that must be announced before a message is sent.

Industrial radio. Radio-frequency energy is of value in a number of industrial heating applications, and radio-frequency electric furnaces, operating at about 10 kilohertz, produce materials of very high purity. Radio-frequency heating (*q.v.*) is employed in removing impurities from the two elements, germanium and silicon, that form the basis of semiconductor technology, as well as from other elements and compounds. The process is known as zone melting (*q.v.*).

Heating of glues bonding nonmetallic surfaces is often best undertaken by placing the material between the plates of a capacitor to which radio-frequency energy is applied. Frequencies at the upper end of the shortwave band are valuable in diathermy—*i.e.,* therapeutic heat treatment of the human body. Recovery from muscular injury is often helped by subjecting the injured limb to diathermy. Microwave energy is now used on a limited scale for cooking; because of the nature of microwaves, such an oven cooks the food from inside to outside.

All radio-frequency heating equipment must be fully shielded to prevent interference with radio communication.

Radio telemetry. Remote reading of instruments taking measurements in conditions dangerous to human life, or in places inaccessible or not easily accessible, can be undertaken over long distances by radio; the technique is known as radio telemetry (see also TELEMETRY). Control of nuclear-power sources and of the manufacture of toxic materials, as well as experiments aimed at exploring conditions in the stratosphere, ionosphere, and outer space may involve radio telemetry. The measurement is converted to a change of an electrical quantity and is transmitted as an information signal superimposed on a radio-frequency carrier, which may be continuous or pulsed. Change of temperature. for example, can change the length of a coil or the spacing of the plates of a capacitor in the tuning circuit of a small transmitter; the transmitter frequency can then be calibrated in terms of temperature. Pressure and humidity changes can be converted in a similar manner; this is the principle of operation of the radiosonde that carries by balloon a small transmitter and instruments for measuring the temperature, pressure, and humidity of the upper atmosphere.

Information about solar radiation, solar wind, magnetic fields, and ionospheric soundings from probes in outer space may have to be stored on magnetic tape for subsequent transmission either on command, or by prearrangement, to the information centre on Earth. Telemetry signals have been received back satisfactorily from space probes at distances greater than 35,000,000 miles (56,000,000 kilometres).

Miscellaneous applications. Radio has found applications in many fields besides communications and has accelerated developments in almost all branches of science; medicine, surgery, astronomy, and mathematics have all benefitted. Diagnosis of disease has been speeded up and made more positive by the adoption of radio methods. The radio-frequency scalpel has proved its worth in delicate brain and eye surgery.

Radio waves from outer space

Radio astronomy, now an important branch of science, was born with the reception (1932) of appreciable 20 megahertz radio emissions from the Milky Way galaxy. In 1942 investigators measured radio emissions from the Sun. Subsequent work has shown that emissions reaching the Earth from outer space cover a range of at least 10 megahertz to 30 gigahertz. Well-defined sources of radiation have been discovered, and some have been associated with visible stars or have led to the discovery of hitherto unknown stars; but many of these quasars, as they are called, emanate from a part of space where no stars have yet been seen. Certain of the radio sources have shown a slow periodic variation in emission, and these have been named pulsars.

Clear resolution of the longer wavelength emission is difficult without very large antenna systems; such systems may be movable or may consist of fixed installations on Earth depending on the Earth's rotation to scan the sky. Such fixed systems are in action in Britain, the United States, The Netherlands, France, Germany, Australia, and the Soviet Union.

RADIO RECEIVER MANUFACTURING

Hand wiring. In the early days of radio and up to the end of World War II, radio receivers consisted of resistors, capacitors, inductors (coils), and electronic tubes joined together by wires with coloured insulation. A colour code, whereby a particular colour was assigned to a particular circuit connection, such as black leads for filaments, green for grid, was adopted throughout the world to facilitate manufacture and the tracing of faults. Later, wires cut to the right length were laced together into a harness to speed assembly. Plugs and sockets were employed for connecting one major part with another. Printed circuit wiring, developed during the 1940s, eliminated much of the hand work and produced important manufacturing economies.

Printed circuits. With printed wiring, the layout of the circuit is planned with component size and position in mind, and connections are made by suitably shaped copper strip or foil bonded to an insulating board or substrate. An extension of this technique was the printed component; resistors, capacitors, and low value inductors became a part of the printing process.

The development of the transistor simplified the exploitation of printed circuitry by eliminating one of the bulkiest components, the vacuum tube. Further development led to the manufacture of the integrated circuit (*q.v.*), which can perform a multiplicity of tasks such as amplification and switching. These circuits are widely used in computers where space is at a premium. Integrated-circuit amplifiers are likely to become more important because of their ability to amplify very high frequencies.

The size of a portable receiver constructed from microminiature circuits is now dictated almost entirely by the loudspeaker and the quality of reproduction required. The smaller the loudspeaker the lower the power it can accept and the less the output of low audio frequencies.

Safety considerations. Voltages involved in a portable radio receiver operated from batteries are so low (less than 20 volts) that it represents no safety hazard, but a receiver that has an alternative power-line connection (that is, it can operate either from batteries or from a wall plug) must be designed with care so that all metal that could come into contact with the user is properly isolated from power-line voltages.

Receivers supplying a large audio output power need adequate ventilation, especially in the neighbourhood of the power transformer and output transistors, because these may be destroyed by overheating. Ventilating openings in the receiver must not be blocked by walls or curtains. The input transformer and the direct current supply to the transistors should be protected by fuses. The danger of fire is much less with a power-line-operated transistor receiver than with a vacuum tube receiver since transistors require no heater supply and so produce much less heat.

BIBLIOGRAPHY. BRITISH STANDARDS INSTITUTION, ***Glossary of Terms Used in Telecommunications,*** 3rd ed. (1960), terms, general and esoteric, are defined, and alternatives listed; J.G. CROWTHER, ***Discoveries and Inventions of the 20th Century,*** 5th ed. rev. (1966), information on the early developments in radio transmission and reception; S.G. STURMEY, ***The Economic Development of Radio*** (1958), a review of the economic forces governing the worldwide development of radio; W.J. BAKER, ***A History of the Marconi Company*** (1970), a study of the development and commercial exploitation of radio in the 20th century, with particular reference to Marconi and the Marconi company; INTERNATIONAL TELEPHONE AND TELEGRAPH CORPORATION, ***Reference Data for Radio Engineers,*** 5th ed. (1968), data of value to those working in radio, and some information of interest to the layman; J.A. BETTS, ***High Frequency Communications*** (1967), a comprehensive survey of the problems of high-frequency communication; K.W. GATLAND (ed.), ***Telecommunication Satellites*** (1964), an examination of the technical and economic problems associated with satellite communication; AMERICAN RADIO RELAY LEAGUE, ***Radio Amateur's Handbook*** (annual), and RADIO SOCIETY OF GREAT BRITAIN, ***Radio Communication Handbook,*** 4th ed. (1968), two guides for the amateur radio operator.

(K.R.S.)

Radioactivity

The dramatic discovery and early studies of radioactivity —*i.e.*, the property exhibited by certain types of matter of emitting energy and subatomic particles spontaneously --at the end of the 19th century ushered in an unprecedented period of exciting advances in man's knowledge, first of the submicroscopic nature of matter and eventually of the enormous energy sources that supply the Sun and stars.

Discovery of radioactivity

Radioactivity was discovered by Henri Becquerel, a French physicist, in 1896, only a few months after the discovery in Germany of X-rays. The fact that X-ray tubes showed visible fluorescence (*i.e.*, re-emitted the energy as light) of the glass during X-ray production first led Becquerel to suppose that there might be some close relation between fluorescence in the form of visible light and the emission of X-rays. Some of the uranium salts are fluorescent; hence, Becquerel concentrated his studies on uranium, exciting fluorescence by sunlight. During a succession of cloudy days in Paris, he left wrapped photographic plates in the dark near the uranium but found them nevertheless blackened when developed. This experiment and subsequent ones convinced him that uranium continuously emitted radiations of great penetrating power, like X-rays, and the emission proceeded independently of chemical form and of external energy sources.

This article is divided into the following sections:

I. General considerations

HISTORY AND NOTABLE MILESTONES IN RADIOACTIVITY

The history of discoveries and advancing knowledge of radioactivity is too long and varied to be treated in detail here. Table 1 presents a chronology of notable events.

NUCLEAR STRUCTURE

To understand the phenomena of radioactivity, a knowledge of the structure of atoms is useful (see ATOMIC STRUCTURE; NUCLEUS, ATOMIC; SPECTROSCOPY; ISOTOPES; PARTICLES, SUBATOMIC; ELECTROMAGNETIC RADIATION; only a brief overview can be given here).

The known elements, over 100 in number, are composed of atoms consisting of a massive nucleus carrying one or more positive charges, surrounded by a cloud of negatively charged electrons equal in number to the charge of the nucleus, called the atomic number (conventionally designated Z). The nucleus consists of two kinds of more massive particles, neutrons without electrical charge and protons each with a single positive charge, the two kinds of particles having about the same mass, respectively 1,836.1 and 1,838.63 times the mass of an electron. The sum of the numbers of protons and neutrons (called nucleons collectively) in the atomic nucleus is the atomic mass number (*A*).

Definition of half-life

The exact masses of nuclei and neutral atoms are measured in atomic mass units (amu), a scale in which the carbon-12 atom is defined as exactly 12 mass units. On this scale most atoms have masses that are nearly but not exactly whole numbers. All the atoms of any specific element have the same number of protons (the atomic number) but may have different numbers of neutrons; thus, for one atomic number (Z) there may be more than one stable atomic mass number. Nuclei with the same number of protons but different number of neutrons are called isotopes. The term nuclide refers to a particular nuclear species with its electron cloud. The nucleons are strongly bonded together by a force peculiar only to neutrons and protons, called a nucleon–nucleon force. Certain combinations of nucleons produce exceedingly stable nuclei, whereas other combinations are unstable to varying degrees. An unstable nucleus, or radioisotope, will decompose spontaneously, or decay, into a more stable configuration but will do so only in a few specific ways by emitting certain particles or certain entities (quanta) of electromagnetic energy, called photons. Radioactive decay is a property of several naturally occurring elements as well as of artificially produced isotopes of the elements. The rate at which a radioactive element decays is expressed in terms of its half-life; *i.e.*, the time required for one-half of any given quantity of the isotope to decay. Half-lives range from those too long to measure (greater than 1,000,000,000 [10^9] years for some nuclei) to those too short to measure (less than 10^{-9} second; see below Rates of radioactive transitions). The product of a radioactive decay process (called the daughter of the parent isotope) may itself be unstable, in which case it, too, will decay. The process continues until a stable nuclide has been formed.

The symbolization of a particular nuclide consists of the symbol for the element—say, C for carbon—with the atomic number (Z) as a prefix subscript and the atomic mass (*A*) as a prefix superscript $_Z^AC$; thus, the notation $_6{}^{12}C$ means one atom of the carbon isotope that has six protons and six neutrons; this isotope is referred to as carbon-12. The isotope carbon-14 has eight neutrons and is symbolized $_6{}^{14}C$. Sometimes the number of neutrons is affixed as a subscript; *e.g.*, $_6{}^{14}C_8$ (ordinarily this means the number of these atoms in a molecule).

On the submicroscopic scale of the nucleus the manifestations of quantum mechanics are dominant; *i.e.*, a nuclear system can exist only in certain discrete energy states. Each energy state has particular properties besides its unique energy, the important properties being the spin (more correctly, total angular momentum), magnetic dipole moment, and electric quadrupole moment. The spin is, as the word implies, the mechanical angular momentum of rotation, expressed as an integral (0, 1, 2, . . .) or half-integral ($\frac{1}{2}$, $\frac{3}{2}$, $\frac{5}{2}$, . . .) number, the units being $\hbar$, Planck's constant, h, divided by 2π. The magnetic moment, symbolized by the Greek letter mu, μ, is the measure of the strength of magnetism along the spin direction and is measured in units of the nuclear magneton ($e\hbar/2m_pc$, e representing the elementary unit of charge, m_p the mass of the proton, and c the speed of light). The spectroscopic quadrupole moment, Q_{spec}, is the measure of the nonsphericity of the charge distribution of the nucleus with respect to the spin axis. A positive value of Q_{spec} denotes a cigar-shaped nucleus; a zero value denotes a spherical nucleus; and a negative value denotes a doorknob-shaped nucleus. The usual unit of Q_{spec} is the barn, an area unit equal to 10^{-24} square centimetre, comparable to the cross-sectional area of nuclei. A further property, the parity, concerned with the symmetry of a quantum-mechanical wave equation that describes it, is denoted by a + (even) or − (odd) sign following the spin number.

II. The general phenomena of radioactivity

THE NATURE OF RADIOACTIVE EMISSIONS

Positive and negative beta particles

The emissions of the most common forms of spontaneous radioactive decay are the alpha (a) particle, the beta (β) particle, the gamma (γ) ray, and the neutrino. The alpha (a) particle, or ray, is actually the nucleus of a helium-4 atom, with two positive charges $_2{}^4He$. Such charged atoms are called ions. The neutral helium atom has two electrons outside its nucleus balancing these two charges. Beta particles may be negatively charged (beta minus, symbol e−), also called a negatron, or positively charged (beta plus, symbol e^+), also called a positron. (The beta minus [β^-] particle is actually an electron created in the nucleus during beta decay without any relationship to the orbital electron cloud of the atom.) The positron is regarded as the antiparticle of the negatron, because the two particles when brought together will mutually annihilate each other. Gamma rays are electromagnetic radiations like radio waves, light, and X-rays. Beta radioactivity also produces particles called the neutrino and antineutrino, with no charge and no rest mass, symbolized by the Greek letter nu, ν and $\bar{\nu}$, respectively. The energies associated with these particles are usually expressed in terms of the work required to move a unit electrical charge through a potential of one volt; the unit is the electron volt, or eV (an MeV is 1,000,000 electron volts and is equal to 1.6 × 10^{-13} joule; one kilo-electron volt [keV] is 1,000 electron volts). In the less common forms of radioactivity there may be emitted fission fragments, neutrons, or protons. Fission fragments are themselves complex nuclei with usually between one-third and two-thirds the charge Z and mass A of the parent nucleus. The neutrons and protons, as mentioned above, are the basic building blocks of the complex nuclei, having approximately unit mass on the atomic scale and having zero charge or unit positive charge, respectively. The neutron cannot long exist in the free state. It is rapidly captured by nuclei in matter; otherwise, in free space it will undergo beta-minus decay to a proton, a negatron, and an antineutrino with a half-life of 12.8 minutes. The proton is the nucleus of ordinary hydrogen and is stable.

TYPES OF RADIOACTIVITY

The early work on natural radioactivity associated with uranium and thorium ores identified two distinct types of radioactivity: alpha and beta decay.

Alpha decay. In alpha decay, an energetic helium ion (alpha particle) is ejected, leaving a daughter nucleus of atomic number two less than the parent and of atomic mass number four less than the parent. An example is the decay (symbolized by an arrow) of the abundant isotope of uranium, uranium-238, to a thorium daughter plus an alpha particle. The energy released (*Q*) in million electron volts (MeV) and the half-life ($t_{1/2}$) will be stated for this reaction and those that follow. It should be noted that in every reaction the charges, or number of protons, shown in subscript are in balance on both sides of the arrow, as are the atomic masses, shown in superscript.

$$^{238}_{92}U \longrightarrow {}^{234}_{90}Th + {}^4_2He \qquad \begin{array}{l} Q_\alpha = 4.268 \text{ MeV} \\ t_{1/2} = 4.51 \times 10^9 \text{ years} \end{array}$$

Beta-minus decay. In beta-minus decay, or negatron emission, an energetic negative electron is emitted, producing a daughter nucleus of one higher atomic number and the same mass number. An example is the decay of the uranium daughter product thorium-234 into protactinium-234.

$$^{234}_{90}Th \longrightarrow {}^{234}_{91}Pa + e^- + \bar{\nu} \qquad \begin{array}{l} Q_{\beta^+} = 0.263 \text{ MeV} \\ t_{1/2} = 24.1 \text{ days} \end{array}$$

In the above reaction for beta decay, $\bar{\nu}$ represents the antineutrino, an uncharged particle with zero rest mass. (In the accepted relativistic theory a particle has a total mass m, given by Einstein's famous equation E = mc^2, in which c is the speed of light, and E is the total energy. Most particles have a characteristic rest mass, m_0, their mass when they are not moving. Neutrinos and antineutrinos have zero rest mass; hence, they must always have the speed of light.)

Gamma rays—photon emission

Gamma decay. A third type of radiation, gamma radiation, usually accompanies alpha or beta decay. Gamma rays are photons, the elementary packets of electromagnetic radiation, and are without rest mass or charge. Alpha or beta decay may simply proceed directly to the

Table 1: Notable Events in Understanding of Nuclei and Radioactivity

researcher	year	event
Wilhelm Konrad Rontgen	1895	discovery of X-rays
Henri Becquerel	1896	discovery of radioactivity of uranium
Joseph John Thomson	1897	discovery of electron as constituent of atoms
Marie Sktodowska-Curie and Pierre Curie	1898	discovery of highly radioactive elements polonium and radium in uranium ores
Ernest Rutherford	1899	discovery of radioactive gaseous emanation from thorium
Pierre Curie	1900	classification of two different kinds of radiation, alpha and beta rays, from radium
Paul-Ulrich Villard	1900	discovery of third type of radiation, gamma rays
André-Louis Debierne	1900	discovery of actinium
William Crookes	1900	discovery of radioactive UX_1 (thorium-234)
Max Planck	1901	proposal that electromagnetic radiation comes in packets (quanta)
Ernest Rutherford and Frederick Soddy	1902	recognition that nuclear changes (transmutations) accompany radioactive emissions
William Ramsay and Frederick Soddy	1903	discovery that helium is formed by decay of radium
Albert Einstein	1905	theory of equivalence of mass and energy ($E = mc^2$)
Egon Ritter von Schweidler	1905	formulation of statistical laws of radioactive decay
Otto Hahn	1905	discovery of radioactive thorium (thorium-228)
Norman Robert Campbell and A. Wood	1906	discovery of radioactivity in natural potassium and rubidium
Bertram Borden Boltwood	1907	discovery of ionium (thorium-230)
Johannes Wilhelm Geiger	1908	experiments on alpha-particle scattering by thin foils
Frederick Soddy	1910	originated concept of radioactive isotopes—that a given element can consist of different nuclear species
Ernest Rutherford	1911	proposal that positive charge and most of mass of atoms is concentrated in tiny central nucleus
Niels Bohr	1913	quantum-mechanical theory of planetary atom
Otto Hahn and Lise Meitner	1918	discovery of protactinium
Ernest Rutherford	1919	first artificial transmutatron of nuclei (alpha particles on nitrogen)
Francis William Aston	1919	mass-spectrograph construction and mass measurement of isotopes
Louis de Broglie	1924	proposal that moving particles have wavelike nature
Werner Heisenberg, Erwin Schrodinger. P.A.M. Dirac, Wolfgang Pauli, Max Born, and others	1924–27	formulation of the new quantum mechanics
George Gamow, Ronald Gurney, and Edward Uhler Condon	1928	quantum-mechanical tunnelling theory of rates of alpha decay
Wolfgang Pauli	1931	neutrino hypothesis in beta decay
Harold Urey	1932	discovery of heavy hydrogen (deuterium)
James Chadwick	1932	discovery of neutron
Carl David Anderson	1933	discovery of positron
Frédérick Joliot-Curie and Irbne Joliot-Curie	1934	first artificially produced radioactivity, discovery of positron beta decay
Enrico Fermi	1934	first production of radioactivity by neutron capture, formulation of beta decay rate theory
Hideki Yukawa	1935	meson theory of nuclear forces
Hideki Yukawa and Shoichi Sakata	1936	theoretical proposal of orbital electron capture beta decay
Niels Bohr	1936	compound nucleus model of nuclear reactions
Gregory Breit and Eugene Wigner	1936	single-level resonance formula for nuclear states
C. Perrier and Emilio Segrè	1937	discovery of technetium, element 43
D. Bayley and Horace Richard Crane	1937	discovery of beta-delayed alpha emission
Louis W. Alvarez	1938	experimental discovery of electron capture decay
Hans Bethe	1939	proposal of nuclear mechanisms for energy production in Sun and stars
Otto Hahn and Fritz Strassmann	1939	discovery of nuclear fission
Lise Meitner and Otto R. Frisch	1939	calculation of the huge energy release expected in fission
Niels Bohr and John Archibald Wheeler	1939	comprehensive theory of fission
Richard B. Roberts, R. C. Mayer, and P. Wang	1939	discovery of beta-delayed neutron emission after fission
Georgy Nikolayevich Flerov and Konstantin Antonovich Petrzhak	1940	discovery of spontaneous fission
Martin David Kamen and Samuel Ruben	1940	discovery of carbon-14
Edwin M. McMillan and Philip H. Abelson	1940	discovery of first transuranium element, neptunium, element 93
Enrico Fermi and others	1942	first continuous chain-reacting fission reactor operated
Glenn T. Seaborg, Edwin M. McMillan, Arthur Charles Wahl, and Joseph W. Kennedy	1946*	synthesis of plutonium, element 94
Jacob A. Marinsky. Lawrence E. Glendelii, and Charles D. Coryell	1946*	synthesis of promethium, element 61

*Work done during World War II and kept secret until publication year listed. Discovery of plutonium was in 1941. curium in 1944, americium in 1945.

ground (lowest energy) state of the daughter nucleus without gamma emission, but the decay may also proceed wholly or partly to higher energy states (excited states) of the daughter. In the latter case, gamma emission may occur as the excited states transform to lower energy states of the same nucleus. (Alternatively to gamma emission, an excited nucleus may transform to a lower energy state by ejecting an electron from the cloud surrounding the nucleus. This orbital electron ejection is known as internal conversion and gives rise to an energetic electron and often an X-ray as the atomic cloud fills in the empty orbital of the ejected electron. The ratio of internal conversion to the alternative gamma emission is called the internal-conversion coefficient.)

Isomeric transitions. There is a wide range of rates of half-lives for the gamma-emission process. Usually dipole transitions (see below ***Gamma transitiort rates),*** in which the gamma ray carries off one $\hbar$ unit of angular momentum, are fast, less than nanoseconds (one nanosecond equals 10^{-9} second). The law of conservation of angular momentum requires that the sum of angular momenta of the radiation and daughter nucleus is equal to the angular momentum (spin) of the parent. If the spins of initial and final states differ by more than 1, dipole radiation is forbidden, and gamma emission must proceed more slowly by a higher multipole (quadrupole, octupole, etc.) gamma transition. If the gamma-emission half-life exceeds about a nanosecond, the excited nucleus is said to be in a metastable, or isomeric, state (the names for a long-lived excited state), and it is customary to classify the decay as another type of radioactivity, an isomeric transition. An example of isomerism is found in the protactinium-234 nucleus of the uranium-238 decay chain:

$$^{234m}_{91}\text{Pa} \longrightarrow {}^{234}_{91}\text{Pa} + \gamma \qquad Q_\gamma = 0.0698 \text{ MeV} \quad t_{1/2} = 1.17 \text{ min}$$

The letter *m* following the mass number stands for metastable and indicates a nuclear isomer.

Beta-plus decay. In the 1930s, new types of radioactivity were found among the artificial products of nuclear reactions: beta-plus decay, or positron emission, and electron capture. In beta-plus decay an energetic positive electron is created and emitted, along with a neutrino, and the nucleus transforms to a daughter, lower by 1 in atomic number and the same in mass number. For instance, carbon-11 ($Z = 6$) decays to boron-11 ($Z = 5$) plus one positron and one neutrino:

$$^{11}_{6}\text{C} \longrightarrow {}^{11}_{5}\text{B} + e^{+} + \nu \qquad Q_{\beta^+} = 0.97 \text{ MeV} \quad t_{1/2} = 20.4 \text{ min}$$

Electron capture. Electron capture is a process in which decay follows the capture by the nucleus of an orbital electron. Electron capture (EC) is similar to positron decay in that the nucleus transforms to a daughter of one lower atomic number. Electron capture differs in that an orbital electron from the cloud is captured and annihilated by the nucleus with subsequent emission of an atomic X-ray as the orbital vacancy is filled by an electron

Table 1: Notable Events in Understanding of Nuclei and Radioactivity (continued)

researcher	year	event
Glenn T. Seaborg, Ralph A. James, and Albert Ghiorso	1949*	synthesis of curium, element 96
Glenn T. Seaborg, Ralph A. James, and Leon O. Morgan	1949"	synthesis of americium, element 95
Maria Goeppert Mayer, Otto Haxel, Johannes Hans Daniel Jensen, and Hans E. Suess	1949	spherical nuclear shell model theory
Stanley Gerald Thompson, Albert Ghiorso, and Glenn T. Seaborg	1950	synthesis of berkelium, element 97
Stanley Gerald Thompson Kenneth Street, Jr., Albert Ghiorso, and Glenn T. Seaborg	1950	synthesis of californium, element 98
Victor F. Weisskopf and John M. Blatt	1952	formulation of rate theory for gamma transitions
Aage Niels Bohr and Ben R. Mottelson	1952	nuclear rotational motion theory; spheroidal nuclear shell model theory
Albert Ghiorso, Stanley G. Thompson, Gary H. Higgins, Glenn T. Seaborg, Martin H. Studier, Paul R. Fields. Sherman M. Pried, Herbert Diamond, Joseph F. Mech, Gray Lucas Pyle, John R. Huizenga, Albert Hirsch, Winston M. Manning, Charles I. Brown, Jr., Helen Louise Smith, and Roderick W. Spence	1955†	synthesis of einsteinium, element 99, and synthesis of fermium, element 100
Albert Ghiorso, Bernard G. Harvey, Gregory Robert Choppin, Stanley Gerald Thompson, and Glenn T. Seaborg	1955	synthesis of mendelevium, element 101
Willard F. Libby and others	1955	radiocarbon dating method developed
Sven Gosta Nilsson	1955	orbital energy calculations for nucleons in spheroidal nuclei
Tsung-Dao Lee and Chen Ning Yang	1956	proposal that mirror symmetry (parity conservation) may be violated in beta decay
Chien-Shiung Wu, Ernest Ambler, Raymond W. Hayward, Dale D. Hoppes, and Ralph P. Hudson	1957	experimental proof of parity nonconservation in beta decay
Nikolay Bogolyubov, Aage Niels Bohr, Ben R. Mottelson, and David Pines	1958	pairing force theory of nuclear superfluidity
Maurice Goldhaber, Lee Grodzins, and Andrew W. Sunyar	1958	measurement of right-handed spin sense of neutrino
Rudolf Ludwig Mossbauer	1958	discoverv of recoilless nuclear gamma ray scattering
Albert Ghiorso, Torbjorn Sikkeland, John R. Walton, Jr., and Glenn T. Seaborg	1958	synthesis of nobelium, element 102
Frederick Reines and Clyde L. Cowan	1960	experimental detection of neutrino
Albert Ghiorso, Torbjorn Sikkeland, Almon E. Larsh, and Robert M. Latimer	1961	synthesis of lawrencium, element 103
S.M. Polikanov, V.A. Druin, V.L. Mikheyev, V.A. Karnaukhov, N.K. Skobelev, A.A. Pleve, V.A. Fomichev, V.G. Subbotin, and G.M. Ter-Akopian	1962	discovery of fissioning nuclear shape isomers
R. Barton, R. McPherson, R.E. Bell, William R. Frisken, W.T. Link, and R.B. Moore	1963	discovery of beta-delayed proton emission
Georgy Nikolayevich Flerov, Yury Ts. Oganesyan, Yury V. Lobanov, V.I. Kuznetsov, V.A. Druin, V.P. Perelygin, K.A. Gavrilov, S.P. Tretyakova. and V.M. Ptolko	1964	rival claims for first synthesis of element 104
Villen M. Strutinsky	1967	shell structure theory for deformed nuclei
Albert Ghiorso, Matti Nurmia, James Harris, Kari Eskola, and Pirko Eskola	1969	rival claims for first synthesis of element 104
Georgy Nikolayevich Flerov, Yury T. Oganesyan, Yury V. Lobanov, Yury A. Lazarev, and Svetlana P. Tretyakova	1970	rival claims for first synthesis of element 105
Albert Ghiorso, M.J. Nurmia. K. Eskola. J. Harris, and P. Eskola	1970	rival claims for first synthesis of element 105
Joseph Cerny, K.P. Jackson, C.U. Cardinal, H.C. Evans. and N.A. Jelley	1970	discovery of proton emission radioactivity (excited cobalt-53*m*)

*See footnote for 1946. †Discovered in debris of first thermonuclear bomb test and kept secret until 1955. Einsteinium was discovered in 1952 and fermium in 1953.

from the cloud about the nucleus. An example is the nucleus of beryllium-7 capturing one of its inner electrons to give lithium-7:

$$^{7}_{4}\mathrm{Be} + e^- \longrightarrow {}^{7}_{3}\mathrm{Li} + \nu \qquad Q_{\mathrm{EC}} = 0.8616\ \mathrm{MeV} \qquad t_{1/2} = 53\ \mathrm{days}$$

The main features of radioactive decay of a nuclear species are often displayed in a decay scheme. Figure 1 shows the decay scheme of beryllium-7, taken from the *Table of Isotopes* (first corrected printing, 1965) of C. Michael Lederer, Jack M. Hollander, and Isadore Perlman. The scheme shows the half-life of the parent and that of the excited daughter state as well as its energy, 0.4774 MeV. The spins and parities of all three states are indicated on the upper left-hand side of the level. The multipolarity of the gamma ray (magnetic dipole, M1, plus 0.005 percent electric quadrupole, E2) is indicated above the vertical arrow symbolizing the gamma transition. The slanted arrows symbolize the electron-capture (EC) decay with labels giving the percentage of decay directly to ground state (89.7 percent) and the percentage' of electron-capture decay going via the excited state (10.3 percent). The boldface numbers following the percentages are so-called log ft values, to be encountered below in connection with beta-decay rates. The overall energy release, Q_{EC}, is indicated below. The Q_{EO} is necessarily a calculated value, because there is no general practical means of measuring the neutrino energies accompanying EC decay. With a few electron-capturing nuclides it has been possible to measure directly the decay energy by measurement of a rare process called inner bremsstrahlung (braking radiation). In this process the energy release is shared between the neutrino and a gamma ray. The measured distribution of gamma-ray energies indicates the total energy release. Usually there is so much ordinary gamma radiation with radioactive decay that the inner bremsstrahlung is unobservable.

From C.M. Lederer, J.M. Hollander, and I. Perlman, *Table of Isotopes*

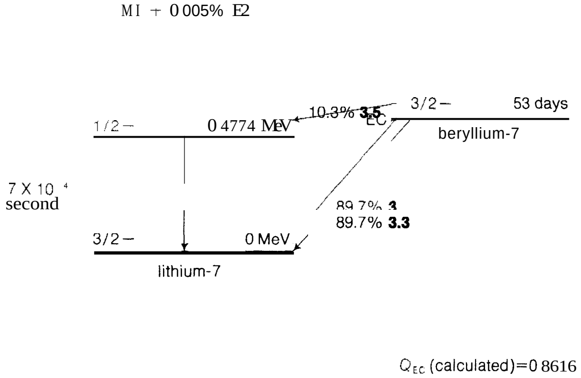


Figure 1: Radioactive decay of beryllium-7 to lithium-7 by electron capture (EC; see text).

Inner bremsstrahlung

Spontaneous fission. Yet another type of radioactivity is spontaneous fission. In this process the nucleus splits into two fragment nuclei of roughly half the mass of the parent. This process is only barely detectable in competition with the more prevalent alpha decay for uranium, but for some of the heaviest artificial nuclei, such as fermium-256, $_{100}{}^{256}\mathrm{Fm}$, spontaneous fission becomes the predominant mode of radioactive decay. Kinetic energy releases from 150 to 200 MeV may occur as the fragments are accelerated apart by the large electrical repulsion between their nuclear charges. The reaction is as follows:

$$\begin{array}{ll} {}^{256}_{100}\mathrm{Fm} \longrightarrow {}^{140}_{54}\mathrm{Xe} + {}^{112}_{46}\mathrm{Pd} + 4n & Q = 150\text{-}200\ \mathrm{MeV} \\ \quad +\text{other fission products} & t_{1/2} = 2.7\ \text{hours} \end{array}$$

Only one of several product sets is shown. A few neutrons are always emitted in fission of this isotope, a feature essential to chain reactions. Spontaneous fission is not to be confused with induced fission, the process involved in nuclear reactors. It is a property of uranium-235, plutonium-239, and other isotopes to undergo fission after absorption of a slow neutron. Other than the requirement of a neutron capture to initiate it, induced fission is quite similar to spontaneous fission regarding total energy release, numbers of secondary neutrons, and so on.

Proton radioactivity. Proton radioactivity, discovered in 1970, is exhibited by an excited isomeric state of cobalt-53, $^{53m}\mathrm{Co}$, 1.5 percent of which emits protons.

$$\begin{tikzcd}[row sep=small] & {}^{52}_{26}\mathrm{Fe} + p & Q_p = 1.57\ \mathrm{MeV} \\ {}^{53m}_{27}\mathrm{Co} \arrow[ur, "1.5\%"] \arrow[dr, "98.5\%"'] & & \\ & {}^{53}_{26}\mathrm{Fe} + e^+ + \nu & t_{1/2} = 0.243\ \mathrm{sec} \end{tikzcd}$$

In addition to the above types of radioactivity, there is a special class of rare beta-decay processes that gives rise to heavy-particle emission. In these processes the beta decay partly goes to a high excited state of the daughter nucleus, and this state rapidly emits a heavy particle.

Special beta-decay processes. ***Beta-delayed neutron emission.*** An example of this type of beta decay is the following reaction:

$$\begin{tikzcd}[row sep=small] {}^{17}_{7}\mathrm{N} \arrow[r] & \overset{*}{{}^{17}_{8}\mathrm{O}} + e- + \bar{\nu} \arrow[dr] & & Q_{\beta^-} = 8.68\ \mathrm{MeV} \\ & & {}^{16}_{8}\mathrm{O} + n & E_{max\ n} = 1.81\ \mathrm{MeV} \\ & & & t_{1/2} = 4.16\ \mathrm{sec} \end{tikzcd}$$

(Note: the asterisk denotes the short-lived intermediate excited states of oxygen-17, and $\boldsymbol{E_{max}}$ ▪ denotes the maximum energy observed for emitted neutrons.) There is a small production of delayed neutron emitters following nuclear fission, and these radioactivities are especially important in providing a reasonable response time to allow control of nuclear fission reactors by mechanically moved control rods.

Beta-delayed alpha emission. Among the positron emitters in the light-element region, a number beta decay partly to excited states that are unstable with respect to emission of an alpha particle. Thus, these species exhibit alpha radiation with the half-life of the beta emission. Both the positron decay from boron-8 and negatron decay from lithium-8 are beta-delayed alpha emission, because ground (unexcited) as well as excited states of beryllium-8 are unstable with respect to breakup into two alpha particles. Another example, sodium-20 (^{20}**Na**) to give successively neon-20 (^{20}**Ne**; the asterisk again indicating the short-lived intermediate state) and finally oxygen-16 is listed below:

$$\begin{tikzcd}[row sep=small] {}^{20}_{11}\mathrm{Na} \arrow[r] & {}^{20}_{10}\mathrm{Ne} + e^+ + \nu \arrow[dr] & & Q_{\beta^+} = \mathbf{13.0\ MeV} \\ & & {}^{16}_{8}\mathrm{O} + \alpha & E_{max\ \alpha} = 4.44\ \mathrm{MeV} \\ & & & t_{1/2} = 0.39\ \mathrm{sec} \end{tikzcd}$$

Beta-delayed proton emissions. In a few cases, positron decay leads to an excited nuclear state not able to bind a proton. In these cases proton radiation appears with the half-life of the beta transition. The combination of high positron-decay energy and low proton-binding energy in the daughter ground state is required. An example is given below, in which tellurium-111 (^{111}**Te**) gives successively antimony-111 (^{111}**Sb**) and then tin-110 (^{110}**Sn**):

$$\begin{tikzcd}[row sep=small] {}^{111}_{52}\mathrm{Te} \arrow[r] & \overset{*}{{}^{111}_{51}\mathrm{Sb}} + e^+ + \nu \arrow[dr] & & Q_{\beta^+}\ \text{uncertain} \\ & & {}^{110}_{50}\mathrm{Sn} + p & E_{max\ p} = 3.7\ \mathrm{MeV} \\ & & & = 19.5\ \mathrm{sec} \end{tikzcd}$$

OCCURRENCE OF RADIOACTIVITY

Some species of radioactivity occur naturally on Earth. A few species have half-lives comparable to the age of the elements (about **6** × **10**^{9} years), so that they have not decayed away after their formation in stars. Notable among these are uranium-238, uranium-235, and thorium-232, the basis of the atomic energy industry. Also there is potassium-40, the chief source of irradiation of the body through its presence in potassium of tissue. Of lesser significance are beta emitters vanadium-50 (^{50}**V**), rubidium-87 (^{87}**Rb**), indium-115 (^{115}**In**), tellurium-123 (^{123}**Te**), lanthanum-138 (^{138}**La**), lutetium-176 (^{176}**Lu**), and rhenium-187 (^{187}**Re**), and the alpha emitters cerium-142 (^{142}**Ce**), neodymium-144 (^{144}**Nd**), samarium-147, 148 ($^{147,\ 148}$**Sm**), gadolinium-152 (^{152}**Gd**), dysprosium-156 (^{156}**Dy**), hafnium-174 (^{174}**Hf**), platinum-190 (^{190}**Pt**), and lead-204 (^{204}**Pb**). Besides these approximately 109-year species, there are the shorter-lived daughter activities fed by one or another of the above species; *e.g.*, by various nuclei of the elements between lead (Z = 82) and thorium (Z = 90).

Cosmic ray bombardment

Another category of natural radioactivity includes species produced in the upper atmosphere by cosmic ray bombardment. Notable are 5,720-year carbon-14 and 12.3-year tritium (hydrogen-3), 53-day beryllium-7, and 2,700,000-year beryllium-10. Meteorites are found to contain additional small amounts of radioactivity, the result of cosmic ray bombardments during their history outside the Earth's atmospheric shield. Activities as short-lived as 35-day argon-37 have been measured in fresh falls of meteorites. Nuclear explosions since 1945 have injected additional radioactivities into the environment, consisting of both nuclear fission products and secondary products formed by action of bomb neutrons on surrounding matter.

The fission products encompass most of the known beta emitters in the mass region 75–160. They are formed in varying yields, rising to maxima of about 7 percent per fission in the mass region 92–102 (light peak of the fission yield versus atomic mass curve) and 134–144 (heavy peak). Two kinds of delayed hazards caused by radioactivity are recognized. First, the general radiation level is raised by fallout settling to Earth. Protection can be provided by concrete or earth shielding until the activity has decayed to a sufficiently low level. Second, ingestion or inhalation of even low levels of certain radioactive species can provide a special hazard, depending on the half-life, nature of radiations, and chemical behaviour within the body.

Some of these internally hazardous products of nuclear explosions are listed in Table 2 below.

Nuclear reactors also produce fission products but under conditions in which the activities may be contained. Containment and waste-disposal practices should keep the activities confined and eliminate the possibility of leaching into groundwaters for times that are long compared to the half-lives. A great advantage of thermonuclear fusion power over fission power, if it can be practically realized, is not only that its fuel reserves, heavy hydrogen and lithium, are vastly greater than uranium, but also that the generation of radioactive fission product wastes can be avoided. In this connection it may be noted that a major source of heat in the Earth's interior and the Moon's interior is provided by radioactive decay. Theories of the formation and evolution of Earth, Moon, and planets must take into account these large heat production sources.

Desired radioactivities other than natural activities and fission products may be produced either by irradiation of certain selected target materials by reactor neutrons or by charged particle beams or gamma ray beams of accelerators.

Table 2: Fall-out Activities Posing Ingestion Hazards

nucleus	formation process	half-life	remarks
^{90}Sr (strontium-90)	fission (5 percent yield)	29 years	long-term deposition in bones, irradiating marrow, the site of red blood cell formation; moves through food chain in milk
^{137}Cs (cesium-137)	fission (6.5 percent)	30 years	body eliminates over several months; localized in the soft tissues
^{89}Sr (strontium-89)	fission (5 percent)	51 days	same remarks as above for ^{90}Sr, but less hazardous since shorter lived
^{131}I (iodine-131)	fission (3 percent)	8 days	short-lived, but can concentrate in food chain, via milk to thyroid gland
^{14}C (carbon-14)	formed by neutrons on atmospheric nitrogen	5,720 years	atmospheric $^{14}CO_2$ incorporated into green plant, then animal material
3H (tritium)	formed by neutrons on lithium-6; released in thermonuclear explosions	12 years	incorporated into living material via tritiated water (^3H_2O); radiations are soft (18.6 keV maximum) and short-ranged

III. Interaction of radiation with matter

The operation of radiation-detection instruments, absorption and shielding materials, and biological effects are all dependent on the interactions of radiation with matter. The types of interaction are described briefly in this article.

For additional details, refer to the articles RADIATION DETECTION AND CHARACTERIZATION; RADIATION EFFECTS ON MATTER; NUCLEUS, ATOMIC; ATOMIC STRUCTURE; X-RAYS; RADIATION, BIOLOGICAL EFFECTS OF; RADIATION INJURY.

Several types of radiation are associated with radioactivity, and a presentation of their interactions calls for recognition of the following types of radiation: charged particles (*a*, *β*, p, fission fragments); photons (y, X-rays); neutral heavy particles (n); and neutrinos.

CHARGED PARTICLES

Role of ionization

Electrically charged particles at the energies occurring in radioactive decay slow down in passage through matter mainly by ionization (ejection of bound electrons) of the electron clouds of the atoms near which they pass. The rate of loss of kinetic energy is thus greater the higher the charge of the particle and indeed is proportional to the square of the charge. The rate of loss is greater the slower the speed (for energies up to about twice the rest mass energy, $2m_0c^2$, at which relativistic effects reverse the trend), because the slower moving particle spends more time close to an atom of the medium and therefore has a greater chance to ionize it. The rate of loss is greater in a medium with lower ionization potential than in one with higher.

Closely related to stopping rate is specific ionization. In a gas and in some insulators and semiconductors, the ionized electrons (those removed from atoms) move far enough from the positively charged parent ions that the electric charge may be collected on electrodes. The number of ion pairs produced per unit path length is thus measurable. A plot of this specific ionization versus path length is called a Bragg curve.

The tracks of alpha particles and fission fragments may be registered in photographic emulsions or in cloud chambers (vessels containing supersaturated vapour, which, rapidly expanded, provides a medium in which the trails of the particles can be traced by trails of visible droplets), and they are seen to be essentially straight paths. Because they move in straight lines and lose energy essentially continuously in a large number of ionizations, heavy charged fragments of a given energy possess a definite range. Ranges may be expressed in distance units but are usually given in terms of mass (mg/cm^2, milligrams per square centimetre of absorber). The straight-line behaviour is a consequence of the fact that alpha and fission-fragment masses exceed by several thousandfold the masses of the electrons slowing them down. The ranges of alpha particles of energies encountered in radioactivity are exceedingly small. For example, the 5.157-MeV alpha particles of plutonium-239 have a range of 4.45 mg/cm^2 of air or 3.75 centimetres (1.5 inches) of dry air at 25° C (77" F) and one atmosphere pressure. The ranges in mass per unit area for other stopping materials are similar.

Beta particles (electrons of both signs), compared to alpha particles and fission fragments, are observed to follow more erratic wandering paths as they slow down in matter, particularly near the ends of their paths. The ionization process is still dominant in slowing down, but the electrons being ionized are of equal rest mass; hence, like a billiard ball colliding with balls of equal mass, the fast electron will usually undergo large deflections in slowing down. These abrupt deflections of beta particles in matter also lead to a conversion of part of their energy into photons by the bremsstrahlung (braking-radiation) process; acceleration or deceleration of charged particles can lead to energy loss by electromagnetic radiation.

Ranges of beta particles

The maximum ranges of beta particles of energies encountered commonly in radioactivity are much longer than for alpha particles, mentioned above. For example, in strontium-90 and yttrium-90 fission-product decay, beta particles are emitted up to a maximum of 2.27 MeV energy. For beta energies between 0.6 MeV and 15 MeV the following simple range relationship was proposed by Norman Feather, a British physicist, for aluminum, and it holds for most other materials as well: the range (R) is proportional to the beta-particle energy (E), or R $= 0.543E - 0.160$. The range (R) is in grams per square centimetre of absorber, and E is in MeV. Thus, for the strontium–yttrium-90 example, the range is 1.07 g/cm^2. The range in distance units is obtained by dividing the density. For water or living tissue with density near unity, the range is 1.07 centimetres.

ELECTROMAGNETIC RADIATION

In contrast to charged-particle stopping, the electromagnetic photons (gamma rays and X-rays) are not gradually slowed down, because they must travel with the speed of light if they survive an encounter. They interact with matter in three kinds of discrete events, described below, in which they disappear or are converted into photons of lower energy.

The photoelectric *effect*. In this type of encounter with orbital electrons of absorber atoms, the photon disappears and gives up its entire energy to the ionization of an orbital electron in the matter. Elements of high atomic number, such as lead or uranium, are far more effective than light elements for the absorption of gamma-ray energies in the 0.1–1 MeV range encountered in radioactivity. The photoeffect probability goes as approximately the fifth power of the atomic number of the absorbing material. The photoeffect probability increases with decreasing photon energy but drops as the energy falls below binding energies of particular electron orbitals.

The Compton effect. In this type of encounter with an electron of the medium, the photon is converted to a lower energy photon that may go off in any direction, the energy and momentum differences being taken up by the electron. The Compton effect is more weakly dependent on atomic number and photon energy than the photoeffect.

Pair production. Gamma rays of energy greater than 1.02 MeV, which is twice the rest mass, m_0c^2, of an electron, may on encountering matter disappear, creating an electron–positron pair. This process rises with photon energy past the 1.02 MeV energetic threshold.

Concept of half-thickness

Absorption calculations. Because gamma-ray absorption is by discrete all-or-nothing events (rather than a slowing down, as with charged particles), gamma rays do not possess definite ranges in matter. The absorption is instead characterized by a half-thickness, the amount of absorber reducing the gamma-ray flux to half its incident value (*i.e.*, half the number of photons). Twice this thickness will cut the flux down to a fourth, three half-thick-

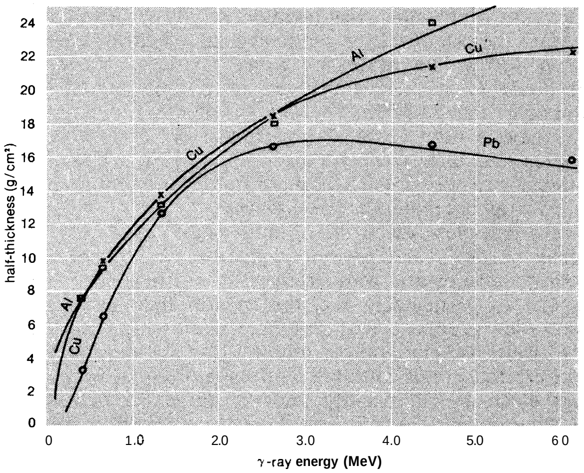


Figure 2: Gamma-ray absorption in aluminum (Al), copper (Cu), and lead (Pb) as a function of energy.

From G. Friedlander, J.W. Kennedy, J.M. Miller Nuclear and Radiochemistry, 2nd ed. (copyright 1964); *used by* permission *of john* Wiley & *Sons,* Inc.

nesses to an eighth, and so on. The combined effect of the three absorption processes gives the half-thickness curves shown in Figure 2 for three materials, aluminum (Al), copper (Cu), and lead (Pb), for gamma-ray energies up to 6 MeV. Lead is the most effective of the three absorbers at all energies, though the differences among materials are small, in the neighbourhood of 1.7 MeV.

As an example of the use of Figure 2, consider the beta emitter cobalt-60, widely used for cancer radiation therapy and for radiation research generally; the important radiations in this regard are the two gamma rays of 1.173 MeV and 1.332 MeV. It is required to ascertain the thickness of lead shielding for a cobalt-60 shipping container to reduce the gamma-ray flux of the 1.332-MeV photon by a factor of 1,024 ($= 2^{10}$, or ten half-thicknesses). One half-thickness value (about 12 g/cm^2) may be read from the curve for lead for an energy of 1.3 MeV. Then the required ten half-thicknesses are 120 g/cm^2 of lead, or, dividing by the density of 11.3 grams per cubic centimetre, a thickness of 10.6 centimetres is obtained.

NEUTRONS AND NEUTRINOS

Shielding problems for neutrons do not arise in the handling of radioactivity, except for the heaviest materials, such as californium-252, for which there is spontaneous fission with its concomitant emission of fast (about one-MeV) neutrons. Neutron slowing down and absorption, however, is a subject of great importance for the operation of fission reactors and the shielding around nuclear accelerators.

Neutrons, having no charge, do not slow down by ionization. They slow down (are moderated) by a series of collisions with nuclei of the medium (called the moderator). The most effective moderator is ordinary hydrogen, because this nucleus, of all nuclides, is of nearly equal mass to the neutron and can thus take up, on the average, the most energy per collision. The slowing down goes on until the neutrons are thermalized; that is, possess the same average thermal kinetic energy as atoms of the moderator. Before and after thermalization, neutrons are subject to capture by various nuclei, some with large probabilities (usually expressed as an effective cross section in barn units, one barn being 10^{-24} cm^2). Water, paraffin, and concrete are employed commonly for neutron shielding because of their large hydrogen contents. This shielding may be combined with materials containing boron or cadmium for more effective capture of the thermalized neutrons.

To a first approximation, neutrinos do not interact with matter, and therefore shielding cannot be used, nor need there be concern about biological radiation effects. Unusually large and specialized detectors have been built for neutrino study, which have detected a few neutrinos through a process of reverse beta decay.

IV. Energetics and kinetics of radioactivity

ENERGY RELEASE IN RADIOACTIVE TRANSITIONS

Consideration of the energy release of various radioactive transitions leads to the fundamental question of nuclear binding energies and stabilities. A much-used method of displaying nuclear-stability relationships is an isotope chart, those positions on the same horizontal row corresponding to a given proton number (Z) and those on the same vertical column to a given neutron number (N). Such **a** map is shown in Figure **3.** The irregular bold line surrounds the region of presently known nuclei. The area encompassed by this is often referred to as the valley of stability, because the chart may be considered a map of a binding energy surface, the lowest areas of which are the most stable. The most tightly bound nuclei of all are the abundant iron and nickel isotopes. Near the region of the valley containing the heaviest nuclei (largest mass number A; that is, largest number of nucleons, N + Z), the processes of alpha decay and spontaneous fission are most prevalent; both these processes relieve the energetically unfavourable concentration of positive charge in the heavy nuclei.

Along the region which borders on the valley of stability on the upper left-hand side are the positron-emitting and electron-capturing radioactive nuclei, with the energy release and decay rates increasing the farther away the nucleus is from the stability line. Along the lower right-hand border region, beta-minus decay is the predominant process, with energy release and decay rates increasing the farther the nucleus is from the stability line.

The grid lines of the graph are at the nucleon numbers corresponding to extra stability, the "magic numbers," to be treated in the section below on ***Nuclear*** *models.* The circles labelled "deformed regions" enclose regions in which nuclei should exhibit cigar shapes; elsewhere the nuclei are spherical. Outside the dashed lines nuclei would be unbound with respect to neutron or proton loss and would be exceeding short-lived (less than 10^{-19} second).

From Proceedings of the International Conference on Properties of Nuclei Far from the Region of Beta-Stability, Leysin, 1970, (CERN 70-30)

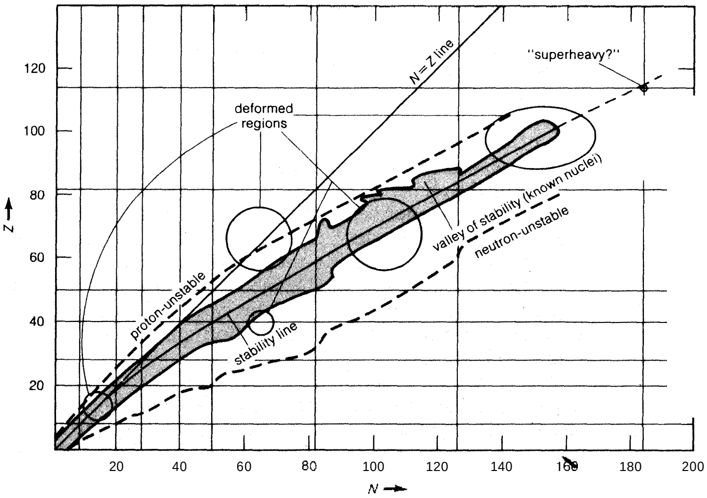
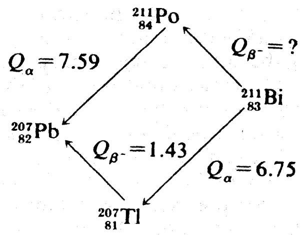


Figure 3: Map of the nuclei.
The coordinates are number of neutrons N and number of protons Z. The bold line encloses the region of known nuclei. The dashed lines are estimated limits for barely binding nucleons. The stability line gives the locus of nuclei most stable with respect to beta decay. Nuclei within the circular regions are cigar-shaped; otherwise the nuclei are spherical. The grid line intersections of the graph indicate the closed shell "magic numbers" of extra stability. A possible region of undiscovered longer lived nuclei is indicated as "superheavy?"

Calculation and measurement of energy. By the method of closed energy cycles, it is possible to **use** measured radioactive-energy-release (Q) values for alpha and beta decay to calculate the energy release for unmeasured transitions. An illustration is provided by the cycle of four nuclei below:

Closed energy cycles



In this cycle, energies from two of the alpha decays and one beta decay are measurable. The unmeasured beta-decay energy for bismuth-211, Q_{β^-}(Bi), is readily calculated, because conservation of energy requires the sum of Q values around the cycle to be zero. Thus, Q_{β^-}(Bi) + 7.59 − 1.43 − 6.75 = 0. Solving this equation gives Q_{β^-}(Bi) = 0.59 MeV. This calculation by closed energy cycles can be extended from stable lead-207 back up the chain of alpha and beta decays to its natural precursor uranium-235 and beyond. In this manner the nuclear binding energies of a series of nuclei can be linked together. Because alpha decay decreases the mass number *A* by 4, and beta decay does not change *A*, closed *α–β*-cycle calculations based on lead-207 can link up only those nuclei with mass numbers of the general type *A* = 4n + 3, in which n is an integer. Another, the 4n series, has as its natural precursor thorium-232 and its stable end product lead-208. Another, the 4n + 2 series, has uranium-238 as its natural precursor and lead-206 as its end product.

In the earlier research on natural radioactivity the classification of isotopes into the series cited above was of great significance, for they were identified and studied as families. Newly discovered radioactivities were given symbols relating them to the family and order of occurrence therein. Thus, thorium-234 was known as UX_1, the isomers of protactinium-234 as UX_2 and UZ, uranium-234 as U_{II}, etc. These original symbols and names are sometimes encountered in more recent literature but are mainly of historical interest. The remaining 4n + 1 series is not naturally occurring but comprises well-known artificial activities decaying down to stable thallium-205.

To extend the knowledge of nuclear binding energies it is clearly necessary to make measurements to supplement the radioactivedecay energy cycles. In part, this extension can be made by measurement of Q values of artificial nuclear reactions (see NUCLEUS, ATOMIC). For example, the neutron-binding energies of the lead isotopes necessary to link the energies of the four radioactive families together can be measured by determining the threshold gamma-ray energy to remove a neutron (photonuclear reaction); or the energies of incoming deuteron (nucleus hydrogen-2, denoted by d, of the heavy hydrogen isotope deuterium) and outgoing proton in the reaction, symbolized *d*, p, can be measured to provide this information.

Further extensions of nuclear-binding-energy measurements rely on precision mass spectroscopy (see MASS SPECTROMETRY, PRINCIPLES OF). By ionizing, accelerating, and magnetically deflecting various nuclides, their masses can be measured with great precision. A precise measurement of the masses of atoms involved in radioactive decay is equivalent to direct measurement of the energy release in the decay process. The atomic mass of naturally occurring but radioactive potassium-40 is measured to be 39.964008 amu (atomic mass units). Masses are usually given as the mass of the electrically neutral atom (nucleus plus orbital electrons) on a scale fixed by the carbon-12 atomic mass being'defined as exactly 12 units. Potassium-40 decays predominantly by P-emission to calcium-40, having a measured mass 39.962589. Through Einstein's equation, energy is equal to mass (*m*) times velocity of light (c) squared, or E = *mc*², the energy release (Q) and the mass difference, *Am,* are related, the conversion factor being one amu, equal to 931.478 MeV. Thus, the excess mass of potassium-40 over calcium-40 appears as the total energy release Q_β in the radioactive decay Q_{β^-} = (39.964008 − 39.962589) × 931.478 MeV = 1.31 MeV. The other neighbouring isobar (same mass number, different atomic number) to argon40 is also of lower mass, 39.962384, than potassium-40. This mass difference converted to energy units gives an energy release of 1.5 MeV, this being the energy release for electron-capture decay to argon-40. The maximum energy release for positron emission is always less than that for electron capture by twice the rest mass energy of an electron ($2m_0c^2$ = 1.022 MeV); thus, the maximum positron energy for this reaction is 1.5 − 1.02, or 0.48 MeV.

Energy–mass conversions

To connect alpha-decay energies and nuclear mass differences requires a precise knowledge of the alpha-particle (helium-4) atomic mass. The mass of the parent minus the sum of the masses of the decay products gives the energy release. Thus, for alpha decay of plutonium-239 to uranium-235 and helium-4 the calculation goes as follows:

M (^{239}Pu)	239.05216
−*M* (^{235}U)	−235.04393
−*M* (^{4}He)	− 4.00260
	0.00563 × 931.478

Q_α = 5.24 MeV

By combining radioactive-decay-energy information with nuclear-reaction Q values and precision mass spectroscopy, extensive tables of nuclear masses have been prepared. From them the Q values of unmeasured reactions or decay processes may be calculated.

Alternative to the full mass, the atomic masses may be expressed as mass defect, symbolized by the Greek letter delta, A (the difference between the exact mass M and the integer *A*, the mass number), either in energy units or atomic mass units.

Absolute nuclear binding energy. The absolute nuclear binding energy is the hypothetical energy release if a given nuclide were synthesized from *Z* separate hydrogen atoms and *N* (equal to *A* − *Z*) separate neutrons. An example is the calculation giving the absolute binding energy of the stablest of all nuclei, iron-56.

26 × M (^{1}H)	26 × 1.007825 =	26.20345
30 × *M* (*n*)	30 × 1.008665 =	30.25995
M (^{56}Fe)		− 55.93493

binding energy = 0.52847 × 931.478 = 492.58 MeV

average binding energy per nucleon of ^{56}Fe = 492.58/56 = 8.796 MeV

A general survey of the average binding energy per nucleon (for nuclei of all elements grouped according to ascending mass) shows a maximum at iron-56 falling off gradually on both sides to about 7 MeV at helium-4 and to about 7.4 MeV for the most massive nuclei known. Most of the naturally occurring nuclei are, thus, not stable in an absolute nuclear sense. Nuclei heavier than iron would gain energy by degrading into nuclear products nearer to iron, but it is only for the elements of greatest mass that the rates of degradation processes such as alpha decay and spontaneous fission attain observable rates. In a similar manner, nuclear energy is to be gained by fusion of most elements lighter than iron. The coulombic (electrostatic) repulsion between nuclei, however, keeps the rates of fusion reactions unobservably low unless the nuclei are subjected to temperatures of greater than 10⁷° C. Only in the hot cores of the Sun and stars or in thermonuclear bombs or controlled fusion plasmas are these temperatures reached and nuclear-fusion energy released.

NUCLEAR MODELS

The liquid-drop model. The average behaviour of the nuclear binding energy can be understood with the model of a charged liquid drop. In this model, the aggregate of nucleons has the same properties of a liquid drop, such as surface tension, cohesion, and deformation. There is a dominant attractive-binding-energy term proportional to the number of nucleons *A*. From this must be subtracted a surface-energy term proportional to surface area and a coulombic repulsion energy proportional to the square of the number of protons and inversely proportional to the nuclear radius. Furthermore, there is a symmetry-energy term of quantum-mechanical origin favouring equal numbers of protons and neutrons. Finally, there is a pairing term that gives slight extra binding to nuclei with even numbers of neutrons or protons.

The pairing-energy term accounts for the great rarity of odd-odd nuclei (the terms odd-odd, even-even, even–odd, and odd–even refer to the evenness or oddness of proton number, *Z*, and neutron number, N, respectively) that are stable against beta decay. The sole examples are deuterium, lithium-6, boron-10, and nitrogen-14. A few other odd–odd nuclei, such as potassium-40, occur in nature, but they are unstable with respect to beta decay. Furthermore, the pairing-energy term makes for the larger number of stable isotopes of even-Z elements, compared to odd-Z, and for the lack of stable isotopes altogether in element 43, technetium, and element 61, promethium.

Mirror nuclei

The beta-decay energies of so-called mirror nuclei afford one means of estimating nuclear sizes. For example, the neon and fluorine nuclei, ${}_{10}{}^{19}\mathrm{Ne}_9$ and ${}_{9}{}^{19}\mathrm{F}_{10}$, are mirror nuclei, because the proton and neutron numbers of one of them equal the respective neutron and proton numbers of the other. Thus, all binding-energy terms are the same in each except for the coulombic term, which is inversely proportional to the nuclear radius. Such calculations along with more direct determinations by high-energy electron scattering and energy measurements of X-rays from mu-mesic atoms (hydrogen atoms in which the electrons are replaced by negative mu mesons, a type of fundamental particle) establish the nuclear charge as roughly uniformly distributed in a sphere of radius 1.2 $A^{1/3} \times 10^{-13}$ centimetre. (The length unit 10^{-13} centimetre occurs so commonly in nuclear physics that it has a special symbol, fm, and is known as the fermi, honouring the great pioneer of nuclear science Enrico Fermi.) That the radius is proportional to the cube root of the mass number has the great significance that the average density of all nuclei is nearly constant.

Careful examination of nuclear-binding energies reveals periodic deviations from the smooth average behaviour of the charged-liquid-drop model. An extra binding energy arises in the neighbourhood of certain numbers of neutrons or protons, called "magic numbers." These numbers are 2, 8, 20, 28, 50, 82, and 126. Nuclei such as ${}_2{}^4\mathrm{He}_2$, ${}_8{}^{16}\mathrm{O}_8$, ${}_{20}{}^{40}\mathrm{Ca}_{20}$, ${}_{20}{}^{48}\mathrm{Ca}_{28}$, and ${}_{82}{}^{208}\mathrm{Pb}_{126}$ are especially stable species, doubly magic, in view of their having both proton and neutron numbers magic. These doubly magic nuclei are situated at the intersections of grid lines on Figure 3.

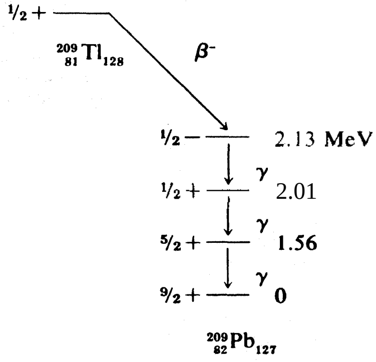
The shell model. In the preceding section the overall trends of nuclear binding energies were described in terms of a charged-liquid-drop model. Yet there were noted periodic binding-energy irregularities at the magic numbers. The periodic occurrence of magic numbers of extra stability is strongly analogous to the extra electronic stabilities occurring at the atomic numbers of the noble-gas (helium, neon, argon, krypton, xenon, and radon) atoms (*i.e.*, 2, 10, 18, 36, 54, 86). The explanations of these stabilities are quite analogous in atomic and nuclear cases as arising from filling of particles into quantized orbitals of motion. The completion of filling of a shell of orbitals is accompanied by an extra stability. The nuclear model accounting for the magic numbers is known as the shell model. The shell model in its simplest form can account for the occurrence of spin zero for all even-even nuclear ground states; the nucleons fill pairwise into orbitals with angular momenta cancelling. The shell model also readily accounts for the observed nuclear spins of the odd-mass nuclei adjacent to doubly magic nuclei, such as ${}_{82}{}^{208}\mathrm{Pb}$. Here, the spins of ½ for neighbouring ${}_{81}{}^{207}\mathrm{Tl}$ and ${}_{82}{}^{207}\mathrm{Pb}$ are accounted for by having all nucleons fill pairwise into the lowest energy orbits and putting the odd nucleon into the last available orbital before reaching the doubly magic configuration (the Pauli exclusion principle dictates that no more than two nucleons may occupy a given orbital, and their spins must be oppositely directed); calculations show the last available orbitals below lead-208 to have angular momentum ½. Likewise, the spins of 9/2 for ${}_{82}{}^{209}\mathrm{Pb}$ and ${}_{83}{}^{209}\mathrm{Bi}$ are understandable, because spin-9/2 orbitals are the next available orbitals beyond doubly magic lead-208. Even the associated magnetization, as expressed by the magnetic dipole moment, is rather well explained by the simple spherical-shell model.

The spherical-shell-mode1 orbitals are labelled in a notation close to that for electronic orbitals in atoms. The orbital configuration of calcium-40 has protons and neutrons filling the following orbitals: $1s_{1/2}$, $1p_{3/2}$, $1p_{1/2}$, $1d_{5/2}$, and $1d_{3/2}$. The letter denotes the orbital angular momentum in usual spectroscopic notation, in which the letters s, p, d, *f*, g, *h*, *i*, etc. represent integer values of *l* running from zero for *s* (not to be confused with spins) through six for *i*. The fractional subscript gives the total angular momentum j with values of $l + ½$ and $l - ½$ allowed, as the intrinsic spin of a nucleon is ½. The first integer is a radial quantum number taking successive values 1, 2, 3, etc., for successively higher energy values of an orbital of given *l* and *j*. Each orbital can accommodate a maximum of 2j + 1 nucleons. Table 3 lists the orbitals comprising each shell, the exact order of various orbitals within a shell differing somewhat for neutrons and protons. The parity associated with an orbital is even (+) if *l* is even (*s*, d, g, i) and odd (—) if *l* is odd (p, *f*, *h*).

Table 3: Spherical-Shell-Model Orbitals

shell closure number	
2	$1s_{1/2}$
8	$1p_{3/2}$, $1p_{1/2}$
20	$1d_{5/2}$, $2s_{1/2}$, $1d_{3/2}$
28	$1f_{7/2}$
50	$2p_{3/2}$, $1f_{5/2}$, $2p_{1/2}$, $1g_{9/2}$
82	$1g_{7/2}$, $2d_{5/2}$, $1h_{11/2}$, $2d_{3/2}$, $3s_{1/2}$
126	$2f_{7/2}$, $1h_{9/2}$, $1i_{13/2}$, $3p_{3/2}$, $2f_{5/2}$, $3p_{1/2}$
184 (?)	$2g_{9/2}$, $1i_{11/2}$, $1j_{15/2}$, $3d_{5/2}$, $2g_{7/2}$, $4s_{1/2}$, $3d_{3/2}$

An example of a spherical-shell-mode1 interpretation is provided by the betadecay scheme of 2.2-minute thallium-209 shown below, in which spin and parity are shown for each state.


½ +
${}^{209}_{81}\mathrm{Tl}_{128}$
β⁻
½ —
2.13 MeV
γ
½ +
2.01
γ
5/2 +
1.56
γ
9/2 +
0
${}^{209}_{82}\mathrm{Pb}_{127}$

The ground and lowest excited states of lead-209 are to be associated with occupation by the 127th neutron of the lowest available orbitals above the closed shell of 126. From the last line of Table 3 it is to be noted that there are available $g_{9/2}$, $d_{5/2}$, and $s_{1/2}$ orbitals with which to explain the ground and first two excited states. Low-lying states associated with the $i_{11/2}$ and $j_{15/2}$ orbitals are known from nuclear-reaction studies, but they are not populated in the beta decay.

The 2.13-MeV state that receives the primary beta decay is not so simply interpreted as the others. It is to be associated with promotion of a neutron from the

$3p_{1/2}$ orbital below the 126 shell closure. The density (number per MeV) of states increases rapidly above this excitation, and the interpretations become more complex and less certain.

Doubly magic nuclei

By suitable refinements, the spherical-shell model can be extended further from the doubly magic region. Primarily, it is necessary to drop the approximation that nucleons move independently in orbitals and to invoke a residual force, mainly short-range and attractive, between the nucleons. The spherical-shell model augmented by residual interactions can explain and correlate around the magic regions a large amount of data on binding energies, spins, magnetic moments, and the spectra of excited states.

The unified model. For nuclei more removed from the doubly magic regions, the spherical-shell model encounters difficulty in explaining the large observed electric quadmpole moments indicating cigar-shaped nuclei. For these nuclei a hybrid of liquiddrop and shell models, the unified model, has been proposed. (See the circular regions of Figure 3 for occurrence of cigar-shaped nuclei.)

Nucleons can interact with one another in a collective fashion to deform the nuclear shape to a cigar shape. Such large spheroidal distortions are usual for nuclei far from magic, notably with 150 $\lesssim A \lesssim$ 190, and 224 $\lesssim A$ (the symbol $<$ denotes less than, and $\sim$ means that the number is approximate). In these deformed regions the unified model prescribes that orbitals be computed in a cigar-shaped potential and that the relatively low-energy rotational excitations of the tumbling motion of the cigar shape be taken into account. The unified model has been highly successful in correlating and predicting nuclear properties in the deformed region. An example of a nuclear rotational band (a series of adjacent states) is provided by the decay of the isomer hafnium-180*m* (^{180m}Hf), Figure 4, through a cascade of gamma rays down the ground rotational band (see below ***Gamma transition rates*** for explanation of M2, E1, E2, and *E3).*

From C.M. Lederer, J.M. Hollander. and I. Perlman, Table of *Isotopes*

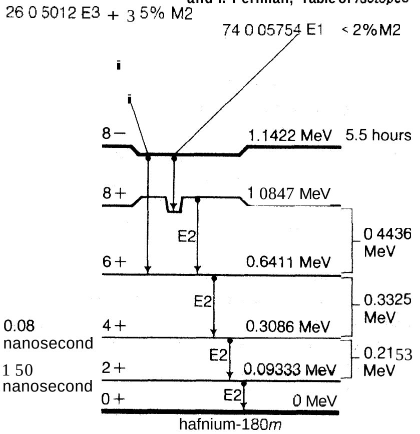


Figure 4: The decay scheme of hafnium-180*m* (^{180m}Hf). See text.

RATES OF RADIOACTIVE TRANSITIONS

There is a vast range of the rates of radioactive decay, from undetectably slow to unmeasurably short. Before presenting detailed consideration of the factors governing particular decay rates, it is well to review the mathematical equations governing radioactive decay and to consider general methods of rate measurement in different ranges of half-life.

Exponential-decay law. Radioactive decay occurs as a statistical exponential rate process. That is, the number of atoms likely to decay in a given infinitesimal time interval (dN/dt) is proportional to the number (N) of atoms present. The proportionality constant, symbolized by the Greek letter lambda, λ, is called the decay constant. Mathematically, this statement is expressed by the first-order differential equation

$$-\frac{dN}{dt} = AN. \tag{1}$$

This equation is readily integrated to give

$$N(t) = N_0 e^{-\lambda t}, \tag{2}$$

in which N_0 is the number of atoms present when time equals zero. From the above two equations it may be seen that a disintegration rate, as well as the number of parent nuclei, falls exponentially with time. An equivalent expression in terms of half-life $t_{1/2}$ is

$$N(t) = N_0(½)^r; r = t/t_{1/2}.$$

It can readily be shown that the decay constant, A, and half-life $(t_{1/2})$ are related as follows: $A = \log_e 2/t_{1/2} = 0.693/t_{1/2}$. The reciprocal of the decay constant λ is the mean life, symbolized by the Greek letter tau, τ.

For a radioactive nucleus, such as potassium-40, that decays by more than one process (89 percent β^-, 11 percent electron capture), the total decay constant is the sum of partial decay constants for each decay mode. (The partial half-life for a particular mode is the reciprocal of the partial decay constant times 0.693.) It is helpful to consider a radioactive chain in which the parent (generation 1) of decay constant λ_1 decays into a radioactive daughter (generation 2) of decay constant λ_2. The case in which none of the daughter isotope (2) is originally present yields an initial growth of daughter nuclei followed by its decay. The equation giving the number (N_2) of daughter nuclei existing at time t in terms of the number $N_1(0)$ of parent nuclei present when time equals zero is

$$N_2(t) = \lambda_1 N_1(0)\,\frac{e^{-\lambda_1 t} - e^{-\lambda_2 t}}{\lambda_2 - \lambda_1}, \tag{3}$$

in which e represents the logarithmic constant 2.71828.

Radio-active decay chain

The general equation for a chain of n generations with only the parent initially present (when time equals zero) is as follows:

$$N_n(t) = N_1(0)\,(C_1 e^{-\lambda_1 t} + C_2 e^{-\lambda_2 t}$$
$$+ \ldots .\, C_n e^{-\lambda_n t})\lambda_1\lambda_2 \ldots \lambda_{n-1}, \tag{4}$$

in which ***e*** represents the logarithmic constant 2.71828.

$$C_1 = 1/(\lambda_2 - \lambda_1)\,(\lambda_3 - \lambda_1) \ldots (\lambda_n - \lambda_1),$$

$$C_2 = 1/(\lambda_1 - \lambda_2)\,(\lambda_3 - \lambda_2) \ldots (A_{,,} - \lambda_2),$$

$$C_n = 1/(\lambda_1 - \lambda_n)\,(\lambda_2 - \lambda_n) \ldots (\lambda_{n-1} - \lambda_n)$$

These equations can readily be modified to the case of production of isotopes in the steady neutron flux of a reactor or in a star. In such cases the chain of transformations might be mixed with some steps occurring by neutron capture and some by radioactive decay. The neutron-capture probability for a nucleus is expressed in terms of an effective cross-sectional area. If one imagines the nuclei replaced by spheres of the same cross-sectional area, the rate of reaction in a neutron flux would be given by the rate at which neutrons strike the spheres. The cross section is usually symbolized by the Greek letter sigma, σ, with the units of barns (10^{-24} cm^2) or millibarns (10^{-3} b) or microbarns (10^{-6} b). Neutron flux is often symbolized by the letters nv (neutron density, n, or number per cubic centimetre, times average speed, v) and given in neutrons per square centimetre per second.

The modification of the transformation equations merely involves substituting the product $nv\sigma_i$ in place of λ_i for any step involving neutron capture rather than radioactive decay. Reactor fluxes nv even higher than 10^{15} neutrons per square centimetre per second are available in several research reactors, but usual fluxes are somewhat lower by a factor of 1,000 or so. Tables of neutron-capture cross sections of the naturally occurring nuclei and some radioactive nuclei can be used for calculation of isotope production rates in reactors.

Measurement of half-life. The measurement of half-lives of radioactivity in the range of seconds to a few years commonly involves measuring the intensity of radiation at successive times over a time range comparable to the half-life. The logarithm of the decay rate is plotted against time, and a straight line is fitted to the points. The time interval for this straight-line decay curve to fall by a factor of 2 is read from the graph as the half-life, by virtue of equations (1) and (2). If there is more than one activity present in the sample, the decay curve will not be a straight line over its entire length, but it should be resolvable graphically (or by more sophisticated statistical analysis) into sums and differences of straight-line exponential terms. The general equations (4) for chain decays show a time dependence given by sums and differences of exponential terms, though special modified equations are required in the unlikely case that two or more decay constants are identically equal.

Decay curve for two or more activities

For half-lives longer than several years it is usually not feasible to measure sufficiently accurately the decrease in counting rate over a reasonable length of time. In such cases a measurement of specific activity may be resorted to. That is, a carefully weighed amount of the radioactive isotope is taken for counting measurements to determine the disintegration rate, D. Then by equation (1) the decay constant λ_i may be calculated. Alternately, it may be possible to produce the activity of interest in such a way that the number of nuclei, N, is known, and again with a measurement of D equation (I) may be used. The number of nuclei, N, might be known from counting the decay of a parent activity or from knowledge of the production rate by a nuclear reaction in a reactor or accelerator beam.

Half-lives in the range of 100 microseconds to one nanosecond are usually measured electronically in coincidence experiments. The radiation giving birth to the species of interest is detected to provide a start pulse for an electronic clock, and the radiation by which the species decays is detected in another detector to provide a stop pulse. The distribution of these time intervals is plotted semi-logarithmically, as discussed for the usual decay-rate treatment, and the half-life is determined from the slope of the straight line.

Half-lives in the range of 100 microseconds to one second must often be determined by special techniques. For example, the activities produced may be deposited on rapidly rotating drums or moving tapes, with detectors positioned along the travel path. The activity may be produced so as to travel through a vacuum at a known velocity and the disintegration rate measured as a function of distance, but this method usually applies to shorter half-lives in or beyond the range of the electronic circuit.

Species with half-lives shorter than the electronic measurement limit are not considered as separate radioactivities, and the various techniques of determining their half-lives will hence not be cited here.

Decay-rate considerations for various types of radioactivity will now be given in the same order as listed above in the section *Types of radioactivity.*

Alpha ***decay.*** Alpha decay, the emission of helium ions, exhibits sharp line spectra when spectroscopic measurements of the alpha-particle energies are made. For even–even alpha emitters the most intense alpha group or line is always that leading to the ground state of the daughter. Weaker lines of lower energy go to excited states, and there are frequently numerous lines observable. The alpha spectrum of curium-242 is shown in Figure 5.

The main decay group of even–even alpha emitters exhibits a highly regular dependence on the atomic number, ***Z,*** and the energy release, Q. (Total alpha energy release, ***Q,,*** is equal to alpha-particle energy, E_{α}, plus daughter recoil energy needed for conservation of momentum; $E_{recoil} = (m_{\alpha}/[m_{\alpha} + M_d])E_{\alpha}$, with m, equal to the mass of the alpha particle and M_d the mass of the daughter product.) As early as 1911, the German physicist Johannes Wilhelm Geiger, together with the British physicist John Mitchell Nuttall, noted the regularities of rates for even-even nuclei and proposed a remarkably successful equation for the decay constant, log $\lambda = a +$ b log r, in which r is the range in air, b is a constant, and ***a*** is given different values for the different radioactive series. The decay constants of odd alpha emitters (odd ***A*** or odd Z or both) are not quite so regular and may be much smaller. The values of the constant b that were used by Geiger and Nuttall implied a roughly 90th-power dependence of λ on Q_{α}. There is a tremendous range of known half-lives from the 2 × 10^{15} years of $_{60}{}^{144}$**Nd** (neodymium) with its 1.83-MeV alpha-particle energy (E,) to the 0.3 microsecond of $_{84}{}^{212}$**Po** (polonium) with E, = 8.78 MeV.

Geiger–Nuttall law

From I. Perlman, F. *Asaro,* and *S.G.* Thompson, *Physical Review,* vol. 92, p. **694** (1953)

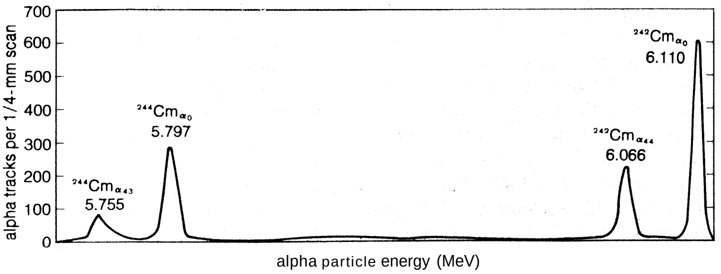


Figure 5: Spectrum of curium-242 and curium-244, showing relative peak intensities for principal alpha particle groups. Ground state alpha groups are labelled α_0.

The theoretical basis for the Geiger–Nuttall empirical rate law remained unknown for 17 years, until the formulation of wave mechanics. A dramatic early success of wave mechanics was the quantitative theory of alpha-decay rates. One of the curious features of wave mechanics is that particles may have a nonvanishing probability of being in regions of negative kinetic energy. In classical mechanics a ball that is tossed to roll up a hill will slow down until its gravitational potential energy equals its total energy, and then it will roll back toward its starting point. In quantum, or wave, mechanics the ball has a certain probability of tunnelling through the hill and popping out on the other side. For objects large enough to be visible to the human eye, the probability of tunnelling through energetically forbidden regions is unobservably small. For submicroscopic objects like alpha particles, nucleons, or electrons, however, quantum mechanical tunnelling can be an important process—as in alpha decay.

The logarithm of tunnelling probability on a single collision with an energy barrier of height B and thickness D is a negative number proportional to thickness D, to the square root of the product of B and particle mass *m.* The size of the proportionality constant will depend on the shape of the barrier and will depend inversely on Planck's constant h (see the article MECHANICS, QUANTUM).

Tunnelling of alpha particles

In the case of alpha decay, the electrostatic repulsive potential between alpha particle and nucleus generates an energetically forbidden region, or potential barrier, from the nuclear radius out to several times this distance. The maximum height (B) of this alpha barrier is given approximately by the expression $B = 2Ze^2/R$, in which Z is the charge of the daughter nucleus, *e* is the elementary charge in electrostatic units, and R is the nuclear radius. Numerically, B is roughly equal to $2Z/A^{1/3}$, with ***A*** the mass number and B in energy units of MeV. Thus, although the height of the potential barrier for $_{84}{}^{212}$**Po** decay is nearly 28 MeV, the total energy released is Q, = 8.95 MeV. The thickness of the barrier (***i.e.,*** distance of the alpha particle from the centre of the nucleus at the moment of recoil) is about twice the nuclear radius of 8.8 × 10^{-13} centimetre. The tunnelling calculation for the transition probability (P) through the barrier gives approximately

$$P = \exp\left[\left(-\frac{\sqrt{2MB}\,R}{\hbar}\right)\left(\frac{\pi B^{1/2}}{Q^{1/2}} - 4\right)\right], \qquad (5)$$

in which M is the mass of the alpha particle and ***A*** is Planck's constant ***h*** divided by 2π. By making simple assumptions about the frequency of the alpha particle striking the barrier, the penetration formula (5) can be used to calculate an effective nuclear radius for alpha decay. This method was one of the early ways of estimating nuclear sizes. In more sophisticated modern techniques the radius value is taken from other experiments, and alpha-decay data and penetrabilities are used to calculate the frequency factor.

The form of equation (5) suggests the correlation of decay rates by an empirical expression relating the half-life ($t_{1/2}$) of decay in seconds to the release energy (Q_α) in MeV:

$$\log t_{1/2} = \frac{a}{\sqrt{Q_\alpha}} + b. \qquad (6)$$

Values of the constants a and b that give best fits to experimental rates of even–even nuclei with neutron number greater than 126 are given in Table 4. The nuclei with 126 or fewer neutrons decay more slowly than the heavier nuclei, and constants a and b must be readjusted to fit their decay rates.

Table 4: Semi-empirical Constants*

	a	*b*
98 californium (Cf)	152.86	–52.9506
96 curium (Cm)	152.44	–53.6825
94 plutonium (Pu)	146.23	–52.0899
92 uranium (U)	147.49	–53.6565
90 thorium (Th)	144.19	–53.2644
88 radium (Ra)	139.17	–52.1476
86 radon (Rn)	137.46	–52.4597
84 polonium (Po)	129.35	–49.9229

*From correlation of ground-state decay rates of even–even nuclei with N > 126. See equation (6) in text.

The alpha-decay rates to excited states of even-even nuclei and to ground and excited states of nuclei with odd numbers of neutrons, protons, or both may exhibit retardations from equation (6) rates ranging to factors of thousands or more. The factor by which the rate is slower than the rate formula (6) is called the hindrance factor. The existence of uranium-235 in nature rests on the fact that alpha decay to the ground and low excited states exhibits hindrance factors of more than 1,000. Thus, the uranium-235 half-life is lengthened to 7 × 10^8 years, a time barely long enough compared to the age of the elements in the solar system for uranium-235 to exist in nature today.

The alpha hindrance factors are fairly well understood in terms of the orbital motion of the individual protons and neutrons that make up the emitted alpha particle. The alpha-emitting nuclei heavier than radium are considered to be cigar-shaped, and alpha hindrance factor data have been used to infer the most probable zones of emission on the nuclear surface—whether polar, equatorial, or intermediate latitudes.

Alpha-decay rates are also a sensitive measure of the tendency of nucleons to pair or cluster in nuclei, and a considerable amount of information about pairing has been deduced therefrom.

The weak interaction

Beta decay. The processes separately introduced at the beginning of this article as beta-minus decay, beta-plus decay, and orbital electron capture can be appropriately treated together. They all are processes whereby neutrons and protons may transform to one another by what is called weak interaction. In striking contrast to alpha decay, the electrons (minus or plus charged) emitted in beta-minus and beta-plus decay do not exhibit sharp, discrete energy spectra but have distributions of electron energies ranging from zero up to the maximum energy release, Q_β. Furthermore, careful measurements of heat released by beta emitters (most radiation stopped in surrounding material is converted into heat energy) show a substantial fraction of the energy, Q_β, is missing. These observations, along with other considerations involving the spins or angular momenta of nuclei and electrons, led an Austrian physicist, Wolfgang Pauli, in 1931 to postulate the simultaneous emission of the neutrino. The neutrino, as a light and uncharged particle with nearly no interaction with matter, was supposed to carry off the missing heat energy. Today, neutrino theory is well accepted with the elaboration that there are four kinds of neutrinos, the electron neutrino and mu neutrino and corresponding antineutrinos of each. The electron neutrinos are involved in nuclear beta-decay transformations, and the mu neutrinos are encountered in decay of muons (mu mesons) to electrons (see further PARTICLES, SUBATOMIC).

Although in general the more energetic the beta decay the shorter is its half-life, the rate relationships do not show the clear regularities of the alpha-decay dependence on energy and atomic number.

The first quantitative rate theory of beta decay was given by Fermi in 1934, and the essentials of this theory form the basis of modern theory. As an example, in the simplest beta-decay process, a free neutron decays into a proton (p), a negative electron (*e*-), and an antineutrino (symbolized by the Greek letter nu with a superior bar; $\bar{\nu}$): $n \rightarrow p + e^- + \bar{\nu}$. The weak interaction responsible for this process, in which there is a change of species (*n* to *p*). by a nucleon with creation of electron and antineutrino, is characterized in Fermi theory by a universal constant, g. The sharing of energy between electron and antineutrino is governed by statistical probability laws giving a probability factor for each particle proportional to the square of its linear momentum (defined by mass times velocity for speeds much less than the speed of light and by a more complicated. relativistic relation for faster speeds). The overall probability law from Fermi theory gives the probability per unit time per unit electron energy interval, *P*(W), as follows:

$$P(W) = \frac{64\pi^4 m_0^5 c^4 g^2}{h^7}\, W(W^2-1)^{1/2}(W_0-W)^2, \qquad (7)$$

in which W is the electron energy in relativistic units ($W = 1 + E/m_0c^2$) and W_0 is the maximum ($W_0 = 1 + Q_\beta/m_0c^2$), m_0 the rest mass of the electron, c the speed of light, and h Planck's constant. This rate law expresses the neutron beta-decay spectrum in good agreement with experiment, the spectrum falling to zero at lowest energies by the factor W and falling to zero at the maximum energy by virtue of the factor $(W_0 - W)^2$.

In Fermi's original formulation the spins of an emitted beta and neutrino are opposing and so cancel to zero. Later work showed that neutron beta decay partly proceeds with the ½$\hbar$ spins of beta and neutrino adding to one unit of *A* (Planck's constant divided by 2π). The former process is known as Fermi decay (F) and the latter Gamow–Teller (GT) decay, after the Russian–born George Gamow and the Hungarian–born Edward Teller, the physicists first proposing it. The interaction constants are determined to be in the ratio $g_{GT}^2/g_F^2 =$ 1.4. Thus, g^2 in equation (7) should be replaced by $(g_F^2 + g_{GT}^2)$.

Gamow–Teller decay

The scientific world was shaken in 1957 by the measurement in beta decay of maximum violation of the law of conservation of parity. The meaning of this nonconservation in the case of neutron beta decay considered above is that the preferred direction of electron emission is opposite to the direction of the neutron spin of ½$\hbar$. (The direction of nuclear spin is an arrow upward in the direction of the thumb of one's right hand, in which the extended fingers point in the direction of rotation). By means of a magnetic field and low temperature it is possible to cause neutrons in cobalt-60 and other nuclei, or free neutrons, to have their spins set preferentially in the up direction perpendicular to the plane of the coil generating the magnetic field. The fact that beta decay prefers the down direction for spin means that the reflection of the experiment as seen in a mirror parallel to the coil represents an unphysical situation: conservation of parity, obeyed by most physical processes, demands that experiments with positions reversed by mirror reflection should also occur. Further consequences of parity violation in beta decay are that spins of emitted neutrinos and electrons are directed along the direction of flight, totally so for neutrinos and partially so by the ratio of electron speed to the speed of light for electrons.

The overall half-life for beta decay of the free neutron, measured as 12 minutes, may be related to the interaction constants g^2 (equal to $g_F^2 + g_{GT}^2$) by integrating (summing) probability expression (7) over all possible electron energies from zero to the maximum. The result for the decay constant is

$$\lambda = \frac{64\pi^4 m_0^5 c^4 g^2}{h^7}\left\{(W_0^2-1)^{1/2}\left(\frac{W_0^4}{30}-\frac{3W_0^2}{20}-\frac{2}{15}\right)+ \right.$$
$$\left. +\frac{W_0}{4}\ln[W_0+(W_0^2-1)^{1/2}]\right\}, \qquad (8)$$

in which W_0 is the maximum beta-particle energy in relativistic units ($W_0 = 1 + Q_\beta/m_0c^2$), with m_0 the rest mass of the electron, c the speed of light, and h Planck's constant. The best g value from decay rates is approximately 10^{-49} erg centimetre^3. As may be noted from equation (8), there is a limiting fifth-power energy dependence for highest decay energies.

Beta spectra decay

In the case of a decaying neutron not free but bound within a nucleus, the above formulas must be modified. First of all, as the nuclear charge Z increases, the relative probability of low-energy electron emission increases by virtue of the coulombic attraction. For positron emission, which is energetically impossible for free protons but can occur for bound protons in proton-rich nuclei, the nuclear coulomb charge suppresses lower energy positrons from the shape given by equation (7). This equation can be corrected by a factor $F(Z,W)$ depending on the daughter atomic number Z and electron energy W. The factor can be calculated quantum mechanically. The coulomb charge also affects the overall rate expression (8) such that it can no longer be expressed as an algebraic function, but tables and graphs are available for analysis of beta decay rates. The rates are analyzed in terms of a function $f(Z,Q_\beta)$ calculated by integration of equation (7) with correction factor $F(Z,W)$. The β^- and β^+ spectra of copper-64 are shown in Figure 6.

From G. Friedlander, J.W. Kennedy. J.M. Miller. *Nuclear and Radiochemistry*, 2nd ed. (copyright 1964); used by permission of John Wiley & Sons. Inc.

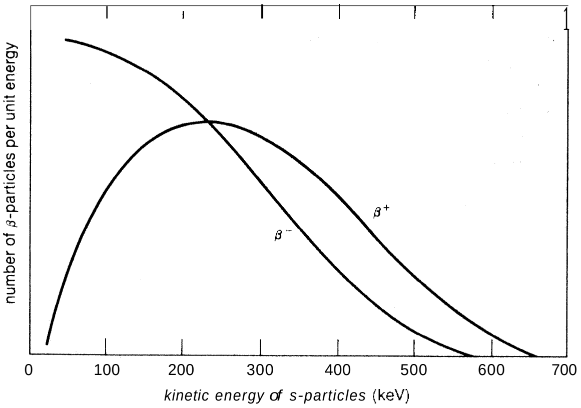


Figure 6: β^- and β^+ spectra cf copper-64.

Approximate expressions for the f functions usable for decay energies Q between 0.1 MeV and 10 MeV, in which Q is measured in MeV, and Z is the atomic number of the daughter nucleus, are as follows (the symbol ≈ means approximately equal to):

$$f_{\beta^-} \approx 6.0Q^{4-0.005(Z-1)}\cdot 10^{Z/50},$$

$$f_{\beta^+} \approx 6.2Q^4/10^{0.007Z}\cdot 10^{0.009Z(\log 1/3Q)^2}.$$

For electron capture, a much weaker dependence on energy is found:

$$f_{EC} \approx (Z+1)^{3.5}Q/4\times 10^5 .$$

The basic beta decay rate expression obeyed by the class of so-called superallowed transitions, including decay of the neutron and several light nuclei. is

$$\lambda_\beta = \frac{64\pi^4 m_0^5 c^4 g^2}{h^7} f_\beta \qquad (9)$$

Like the ground-to-ground alpha transitions of even–even nuclei, the superallowed beta transitions obey the basic rate law, but most beta transitions go much more slowly. The extra retardation, as with alpha hindrance, is explained in terms of mismatched orbitals of neutrons and protons involved in the transition. For the superallowed transitions the orbitals in initial and final states are almost the same. Most of them occur between mirror nuclei, with one more or less neutrons than protons; *i.e.,* beta minus decay of hydrogen-3, electron capture of beryllium-7 and positron emission of carbon-11, oxygen-15, neon-19, . . . titanium-43.

The nuclear retardation of beta decay rates below those of the superallowed class may be expressed in a fundamental way by multiplying the right side of equation (9) by the square of a nuclear matrix element (a quantity of quantum mechanics), which may range from unity down to zero depending on the degree of mismatch of initial and final nuclear states of internal motion. A more usual way of expressing the nuclear factor of the beta rate is the log ft value, in which f refers to the function $f(Z,Q_\beta)$. Because the half-life is inversely proportional to the decay constant λ, the product $f_\beta t_{1/2}$ will be a measure of (inversely proportional to) the square of the nuclear matrix element. For the log ft value, the beta half-life is taken in seconds, and the ordinary logarithm to the base 10 is used. The superallowed transitions have log ft values in the range of 3 to 3.5. Beta log ft values are known up to as large as —23 in the case of indium-115. There is some correlation of log ft values with spin changes between parent and daughter nucleons, the indium-115 decay involving a spin change of four, whereas the superallowed transitions all have spin changes of zero or one. Various classifications according to spin change may be made for beta decay (allowed, first forbidden, second forbidden, etc.).

Beta *ft* values

Gamma transition. The nuclear gamma transitions belong to the large class of electromagnetic transitions, encompassing radio frequency emission by antennas or rotating molecules, infrared emission by vibrating molecules or hot filaments, visible light, ultraviolet light, and X-ray emission by electronic jumps in atoms or molecules. The usual relations apply for connecting frequency ν, wavelength λ, and photon quantum energy E with speed of light c and Planck's constant h, namely, $\lambda = c/\nu$ and $E = h\nu$. It is sometimes necessary to consider the momentum (p) of the photon given by $p = E/c$.

Classically, radiation accompanies any acceleration of electrical charge. Quantum mechanically there is a probability of photon emission from higher to lower energy nuclear states, in which the internal state of motion involves acceleration of charge in the transition. Thus, purely neutron orbital acceleration would carry no radiative contribution.

A great simplification in nuclear gamma transition rate theory is brought about by the circumstance that the nuclear diameters are always much smaller than the shortest wavelengths of gamma radiation in radioactivity. That is, the nucleus is too small to be a good antenna for the radiation. The simplification is that nuclear gamma transitions can be classified according to multipolarity, or amount of spin angular momentum carried off by the radiation. One unit of angular momentum in the radiation is associated with dipole transitions (a dipole consists of two separated equal charges, plus and minus). If there is a change of nuclear parity, the transition is designated electric dipole (E1) and is analogous to the radiation of a linear half-wave dipole radio antenna. If there is no parity change, the transition is magnetic dipole (M1) and is analogous to the radiation of a full-wave loop antenna. With two units of angular momentum change the transition is electric quadrupole (E2), analogous to a full-wave linear antenna of two dipoles out-of-phase, and magnetic quadrupole (M2), analogous to coaxial loop antennas driven out-of-phase. Higher multipolarity radiation also frequently occurs with radioactivity.

Multipolarities are experimentally measured by angular dependence of radiation or by ratios of orbital electron ejection to photon emission (internal conversion).

Transition rates are usually compared to the single-proton theoretical rate, or Weisskopf formula, named after a United States physicist, Victor Frederick Weisskopf, who developed it. Table 5 gives the theoretical

Table 5: Gamma Transition Rates*

transition type	partial half-life *t* (seconds)	illustrative t_γ valuer for $A = 125, E = 0.1$ MeV (seconds)
E1	$5.7 \times 10^{-15} E^{-3} A^{-2/3}$	2×10^{-13}
E2	$6.7 \times 10^{-9} E^{-5} A^{-4/3}$	1×10^{-6}
E3	$1.2 \times 10^{-2} E^{-7} A^{-2}$	8
E4	$3.4 \times 10^{4} E^{-9} A^{-8/3}$	9×10^{7}
E5	$1.3 \times 10^{11} E^{-11} A^{-10/3}$	1×10^{15}
M1	$2.2 \times 10^{-14} E^{-3}$	2×10^{-11}
M2	$2.6 \times 10^{-8} E^{-5} A^{-2/3}$	1×10^{-4}
M3	$4.9 \times 10^{-2} E^{-7} A^{-4/3}$	8×10^{2}
M4	$1.3 \times 10^{5} E^{-9} A^{-2}$	8×10^{9}
M5	$5.0 \times 10^{11} E^{-11} A^{-8/3}$	1×10^{17}

*The energies *E* are expressed in MeV. The nuclear radius has been taken as 1.3 fermis. It is tu be noted that t_γ is the partial half-life for γ emission only; the occurrence of internal conversion will always shorten the measured half-life.

reference rate formulas in their dependence on nuclear mass number ***A*** and gamma ray energy Eγ (in MeV).

It is seen for the illustrative case of gamma energy 0.1 MeV and mass number 125 that there occurs an additional factor of 10^7 retardation with each higher multipole order. For a given multipole, magnetic radiation should be a factor of 100 or so slower than electric. These rate factors generally ensure that nuclear gamma transitions are nearly purely one multipole, the lowest permitted by the nuclear spin change. There are numerous exceptions, however; mixed M1–E2 transitions are common, because E2 transitions are often much faster than the Weisskopf formula gives and M1 transitions are generally slower. All E1 transitions encountered in radioactivity are much slower than the Weisskopf formula. The other higher multipolarities show some scatter in rates, ranging from agreement to considerable retardation. In most cases the retardations are well understood in terms of nuclear model calculations.

Electric monopole transitions

Though not literally a gamma transition, electric monopole (EO) transitions may appropriately be mentioned here. These may occur when there is no angular momentum change between initial and final nuclear states and no parity change. For spin-zero to spin-zero transitions, single gamma emission is strictly forbidden. The electric monopole transition occurs largely by the ejection of electrons from the orbital cloud in heavier elements and by positron–electron pair creation in the lighter elements.

V. Applications of radioactivity

MEDICAL APPLICATIONS

Radioisotopes have found extensive use in medical therapy. Radium-226 and its radon daughter have been used since early days of radiology as radiation sources to arrest development of cancer. These isotopes are now largely replaced by cobalt-60. Many medical applications use isotopes not as radiation sources but as tracers localizing in particular organs and giving diagnostic information. In a 1969 speech, U.S. nuclear chemist Glenn T. Seaborg, then chairman of the U.S. Atomic Energy Commission, summarized the status of medical applications noting that cobalt-60 (and to some extent cesium-137) was used to treat cancer and as a suppressant of the immune reaction occurring after the transplanting of human organs; iodine-131 (and in some cases other radioactive iodine isotopes, such as iodine-125) was used to diagnose various thyroid disorders, treat hyperthyroidism and functional metastatic thyroid cancer, diagnose kidney and liver disorders and make tests of the functioning of these organs, screen for pulmonary emboli by lung scans, locate tumours, and make blood-volume and cardiac-output studies; carbon-14 was used to study abnormalities of metabolism that underlie diabetes, gout, anemia, and acromegaly; gallium-68, strontium-85, strontium-87*m*, fluorine-18, and calcium-47 were used for bone scanning in suspected metabolic disorders and malignancies; chromium-51, to measure cell mass; arsenic-74, to locate brain tumours; yttrium-90, for treating pituitary gland disorders; gold-198, as a palliative in lung cancer; mercury-197 and mercury-203, for renal function tests; selenium-75 (selenomethionine), to scan the pancreas for cancer detection; sodium-22 and sodium-24, to measure exchangeable sodium and diagnose circulatory disorders; phosphorus-32, to treat the blood disorders polycythemia vera and chronic leukemias, and in some cases even bone cancer; iron-59 and iron-55, to measure the rate of formation of red-cell mass, red-cell survival times, and intestinal absorption rates; and cobalt-57 (and cobalt-60), for vitamin B_{12} absorption tests. Combinations of radioisotopes are used in many applications, such as the measurement of red blood cells and plasma volume for the control of blood transfusions. Seaborg also noted the uses of six-hour technetium-99*m*, widely used for its nearly ideal properties in diagnosing thyroid, liver, brain, and kidney diseases.

As more hospitals acquire low power isotope production reactors or cyclotrons, the medical use of a wide range of short-lived isotopes is increasing. Table 6 lists some medically useful short-lived isotopes.

Table 6: Some Medically Useful Short-lived Gamma-emitting Isotopes

	isotope	half-life (hours)
Sodium	^{24}Na	15.0
Magnesium	^{28}Mg	21.0
Silicon	^{31}Si	2.6
Potassium	^{42}K	12.5
Calcium	^{47}Ca	113.0
Manganese	^{56}Mn	2.6
Nickel	^{65}Ni	2.6
Zinc	^{69m}Zn	13.8
Gallium	^{72}Ga	14.0
Arsenic	^{76}As	26.0
Bromine	^{82}Br	36.0
Strontium	^{87m}Sr	2.8
Molybdenum	^{99}Mo	67.0
Technetium	^{99m}Tc	6.0
Iodine	^{123}I	13.0
Iodine	^{130}I	12.5
Barium	^{137m}Ba	0.043
Lanthanum	^{140}La	40.0
Praseodymium	^{142}Pr	19.2
Dysprosium	^{165}Dy	2.3
Tungsten	^{187}W	24.0
Mercury	^{197}Hg	65.0

Positron diagnostics

Positron emitters offer special possibilities in diagnostic scanning of radioactive tracer movement in the body, for positrons rapidly annihilate with electrons in matter. The annihilation is usually accompanied by emission of two penetrating (0.51 MeV) gamma rays in opposite directions. Two counters on opposite sides of the body recording coincidences (simultaneous arrival of two gammas) can more sharply and efficiently locate the position of the tracer activity than can scanners based on measurement of a single gamma ray. Positron emitters generally cannot be produced in reactors and require charged particle accelerators such as cyclotrons. A table of medically useful positron emitters is given in Table 7.

Table 7: Some Medically Useful Positron Emitting Isotopes Produced by Cyclotrons

	isotope	half-life
Carbon*	^{11}C	20 minutes
Nitrogen	^{13}N	10 minutes
Oxygen	^{15}O	2 minutes
Fluorine*	^{18}F	1.9 hours
Sodium	^{22}Na	2.6 years
Chromium	^{49}Cr	42 minutes
Iron	^{52}Fe	8.3 hours
Cobalt	^{55}Co	18.2 hours
Copper*	^{64}Cu	12.8 hours
Gallium	^{68}Ga	1.1 hours
Arsenic	^{72}As	1.1 days
Arsenic	^{74}As	17.5 days
Technetium	^{94}Tc	54 minutes
Iodine	^{124}I	4.5 days

*Can also be produced in a reactor.

INDUSTRIAL APPLICATIONS

Foremost in applications is the nuclear energy industry based on release of fission energy of uranium through neutron chain reacting reactors. This application is extensively treated in other articles (see NUCLEAR FISSION and NUCLEAR REACTOR).

There are many industrial applications of radioisotopes in measurements of thickness of metal or plastic sheets by absorption of radiation, or in density measurements by backscattering of radiation. Radioisotopes may be used in place of large X-ray machines for X-ray examinations of manufactured metal parts.

Rapid nondestructive analysis of solid materials for the elements composing them can be quickly made by X-ray fluorescence analysis for all but the five lightest elements. The material is exposed to low-energy gamma radiation and the resulting X-rays from the material are detected by an energy-sensitive detector like lithium-drifted silicon or germanium. The elements are measured by the intensity of their characteristic X-rays.

A highly sensitive method of analysis for a number of elements is neutron activation analysis. The material is irradiated with neutrons and the resulting radioactivities are identified as a measure of the elements present in the sample.

OTHER APPLICATIONS

Radioactive isotopes are mixed with fluorescent paints in order to provide the energy for luminous signs or watch dials. Tritium is increasingly being used because of its lack of gamma radiation; it is rapidly replacing radium and thorium.

Because of the constancy in half-life, radioactivities can serve as clocks for a wide variety of scientific deductions. In geochronology and cosmochronology various radioactivities have served to determine the ages of various minerals, the age of the Earth itself, and the time that has passed since the stellar formation of the elements (see DATING, RELATIVE AND ABSOLUTE).

Radioactive isotopes alone can provide small and compact sources of electrical power. The 86-year alpha emitter, plutonium-238, has served as a compact heat source, with negligible radiation outside thin shielding, generating electricity through thermoelectric junction circuits. The 18-year curium-244 and strontium-90 are also promising in these applications.

BIBLIOGRAPHY

General works: GERHART FRIEDLANDER ***et al., Nuclear and Radiochemistry,*** 3rd ed. (1981); and BERNARD G. HARVEY, ***Introduction to Nuclear Physics and Chemistry,*** 2nd ed. (1969), are two excellent introductory texts on nuclear phenomena; AAGE BOHR and BEN ROY MOTTELSON, ***Nuclear Structure,*** vol. 1 (1969), is a scholarly, more advanced reference work; C. MICHAEL LEDERER and VIRGINIA S. SHIRLEY, ***Table of Isotopes,*** 7th ed. (1978), a comprehensive table that lists all of the known radioactive and stable isotopes, together with their properties.

Biological and medical aspects: EDITH HINKLEY QUIMBY, ***Radioactive Nuclides in Medicine and Biology,*** 3rd ed. (1968, reissued 1981), an excellent text and reference work, especially for those whose primary interest is in the biological and medical applications of radioactivity; JOHN H. LAWRENCE, BERNARD MANOWITZ, and BENJAMIN S. LOEB, ***Radioisotopes and Radiation: Recent Advances in Medicine, Agriculture and Industry*** (1969), an illustrated survey of various practical applications of radioisotopes and radiation.

Historical: FREDERICK SODDY, ***Radioactivity and Atomic Theory*** (1975), edited with commentary by THADDEUS J. TRENN; ALFRED ROMER (ed.),***The Discovery of Radioactivity and Transmutation*** (1964), collections of original articles and reports; LAWRENCE BADASH, ***Radioactivity in America*** (1979), a study of developments in the United States from 1896 to 1920; ***The Restless Atom*** (1960, reprinted 1982), a popular account of the discovery of and research on radioactivity.

(J.O.R.)

Radio-Frequency Heating

Radio-frequency heating is the process of heating materials by exposing them to the output of a radio-frequency generator operating at frequencies above 70,000 cycles per second (hertz). This type of heating can be generated exactly where desired to the exclusion of other areas, making the technique essentially efficient.

Radio-frequency heating can be broken down into two groups: inductive heating, which is used for heating metals and other materials that are good conductors of electricity, and dielectric (capacitance) heating, which is used for materials that are poor conductors of electricity, such as insulators. Both techniques have broad applications in industry.

INDUCTION HEATING

Operating principles. Induction-heating technology is based on the fact that electromagnetic energy can be transmitted very efficiently over relatively short distances (up to a few centimetres) without benefit of conductors. Investigative work on this phenomenon was begun in 1935, and studies undertaken at that time provided the foundation for several efficient heating methods now used extensively in various technological processes.

In induction heating, the material to be heated, which must be a fairly good conductor of electricity, is placed in a high-frequency electromagnetic field generated by a conductor or coil (usually called the inductor) connected to a high-frequency generator. This field causes electrical currents (eddy currents) to be excited or induced in the regions of the material located within the field of the inductor. These eddy currents heat up the material; the amount of heat produced depends on the magnitude of the induced currents, the resistance of the material to the flow of electrical current, and the length of time the material is in the field.

During heating, the inductor, usually made of copper tubing, is kept cool by passing water through it. The heated object is located a fraction of a centimetre to several centimetres from the inductor, depending on the technological requirements and the shape of the product, and nowhere does it make contact with the inductor.

"Skin effect"

A phenomenon known as the "surface effect," or "skin effect," causes uneven distribution of alternating current over the cross section of the object being heated and must be allowed for in induction-heating technology. As the frequency of the generator is increased, the induced currents tend to crowd to the surface of the heated material and drop off very rapidly towards the interior. The heat developed in the material follows the same pattern, as the heating is a direct result of the eddy currents.

For practical calculations a procedure based on several assumptions, although slightly reducing the accuracy of the results, is used extensively in high-frequency heating technology. The calculation is based on the equation:

$$p = 5.03 \,.\, 10^4 \sqrt{\rho/\mu f} \text{ millimetres} \qquad (1)$$

in which p is a factor indicating the depth of current penetration, p the electrical resistivity, μ the magnetic permeability of the material with respect to air or vacuum, and f the operating frequency in cycles per second.

As temperature increases p and μ change markedly, and this in turn results in an increase in the depth of penetration of the current. The specific resistance p of steel increases during heating; for example, in the 20"–1,000" C (70°–1,800° F) temperature range, the specific resistance of low-alloy steel increases nearly tenfold. Magnetic permeability μ of steel depends on the intensity of the electromagnetic field created by the inductor and the chemical composition and structure of the steel. Magnetic permeability drops sharply at a critical temperature known as the Curie point and remains close to unity with any further increase in temperature. For example, the Curie point of carbon eutectoid steel containing 0.83 percent carbon is 729° C (1,344" F), while the magnetic properties of pure iron vanish at 768° C (1,414" F). Nonmagnetic steels of the austenite class, which have a constant $\mu = 1$ that does not depend on temperature, are exceptions.

Curie point

As can be seen from Equation 1, the increase in resistivity and the decrease in permeability as the temperature increases both contribute to an increase in the depth of penetration. The depth of penetration for low-alloy structural steel, for example, increases nearly 30-fold in the 20"–1,000" C temperature range.

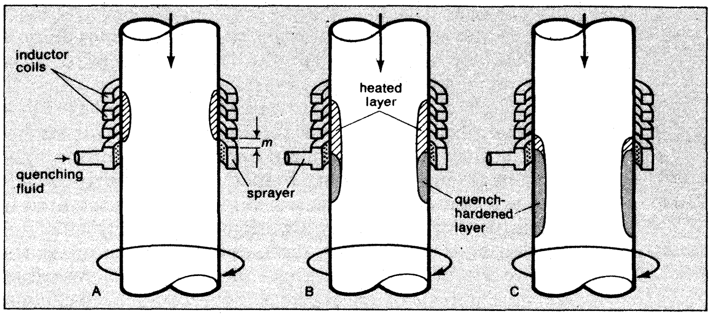


Figure 1: Induction heating and quench hardening of the surface of a rotating cylindrical shaft (see text).

Adapted from M.G. Lorinskii, *Industrial Applications* of *Induction Heating* (1970); Pergamon Press Ltd.

Applications. Induction heating permits the efficient completion of such technological tasks as: (1) case hardening of the working surfaces of steel products; (2) through heating (heating clear through the material as opposed to surface heating) of billets or individual zones of bars for stampings and forgings; (3) heating for soldering with hard and soft solder, permitting manufacture of complete assemblies from individual parts; (4) tempering and annealing of selected portions of a part or component; (5) fusion of metals and production of high-quality alloys.

Surface ***and zonal*** *hardening.* Induction heating makes it possible to reduce manyfold the production time of parts requiring heat treatment for hardening because in most cases heating requires only a few seconds. Moreover, surface hardening can be carried out directly on the production line, leading to total automation of all heating and hardening processes, uniform quality of all processed products, and substantial improvement in working conditions.

There are three basic heating methods for surface hardening: simultaneous, when the entire surface to be processed is subjected at once to heating and subsequent hardening; sequential, the successive heating for hardening of individual areas of the same part; and continuous-sequential heating and hardening, during which the processed object is moved continuously through the zone of influence of the inductor, heating the desired surface layer to the hardening temperature and then to the coolant, where hardening takes place. Greatest productivity can be achieved using the simultaneous heating method, but in this case the generator power must be sufficient to ensure heating of the entire area to be processed at once.

Optimal frequency

The optimal frequency for the induction-heating process, which can be found from empirical formulas, depends on the configuration of the products to be treated and the thickness of the layer to be hardened.

For heating parts with simple shapes (bodies of revolution, planes), the minimum frequency will be

$$f \cong 50{,}000/S^2 \text{ hertz} \qquad (2)$$

in which δ is the thickness of the layer to be heated in millimetres.

For heating parts with complex configurations (with protuberances and depressions in the processed zone), the optimal frequency is about ten times higher, and for heating gear teeth with a magnetic permeability of between four and ten the frequency is still higher.

The time required for heating a cylindrical surface of thickness δ from one to four millimetres can be calculated using the following formula:

$$t \cong \delta^2/2 \text{ seconds.} \qquad (3)$$

The generator power can be approximated by using the relation

$$p_g = \Delta QS/t \text{ kilowatts} \qquad (4)$$

in which ΔQ is the average power required for heating the material, including radiation loss, and S is the area to be heated. For a layer approximately two millimetres thick, Q is 2.5 kilowatt-seconds per square centimetre. Thus for a heating time of two seconds and an area of 100 square centimetres, the generator power must be 2.5 × 100/2 or 125 kilowatts.

The sequential, or successive, method of hardening individual portions of a part, a variation of the simultaneous heating method, is used in production of internal-combustion-engine crankshafts, with the individual journals of the shafts alternately heated and hardened; gear cogs, with sequential hardening "cog by cog"; and other objects.

In many cases, particularly when high productivity is not required, as in processing large parts, a continuous-sequential "feed" method is employed (Figure 1). Figure 1A illustrates the moment at which heating begins. The piece to be processed (in this case a cylindrical shaft) is fed, for instance, from top to bottom through an inductor. When this method is used, feed can also be horizontal or at various angles to the vertical.

In addition to the sequential motion, the shaft is also rotated around its axis to ensure uniform heating of the layer being hardened regardless of the accuracy of positioning relative to the inductor walls. The rate of rotation is usually between 100 and 300 revolutions per minute.

The layer that is heated to hardening temperature is shown by the heavy diagonal shading. At a distance *m* from the inductor, usually about five times the thickness of the heated layer, is situated a sprayer (water or emulsion shower). The coolant emerges through holes 1.O–1.2 millimetres in diameter bored on the inside of the sprayer surface. The height of the sprayer is about ten times the depth of penetration.

The process of hardening the heated layer is illustrated in Figure 1B. The hardened layer is shown in black on the shaft section. If the high-frequency current generator is turned off at the moment shown in Figure 1C, then the shaft must be fed downward for a certain additional distance to harden the heated zone then in the inductor.

For hardening layers of thickness $\delta = 1$–5 millimetres on parts made of structural steel, the minimum feed rate can be calculated by using the empirical formula:

$$v_{\min} = 5/\delta \text{ millimetres per second.} \qquad (5)$$

The maximum feed rate is limited only by the generator power. As generator power increases, the inductor height can be increased. Feed rate on lines for continuous-sequential hardening usually runs between 0.2 and 2.0 centimetres per second.

In cases in which the thickness of the layer to be hardened is greater than the depth of penetration of the eddy currents into the steel, the remaining thickness is heated essentially by thermal conduction from the heated to the unheated region.

The inductor height is selected as a function of the available power and workpiece diameter.

The use of the continuous-sequential heating and hardening method for processing large parts, because of the lower productivity, permits a substantial reduction in generator power capacity.

Inductor configuration

The inductor configuration is one of the most important elements in designing equipment for heating specific areas of parts of varying configurations.

The size of the zone to be heated depends on the shape of the inductor and its location in relation to the part

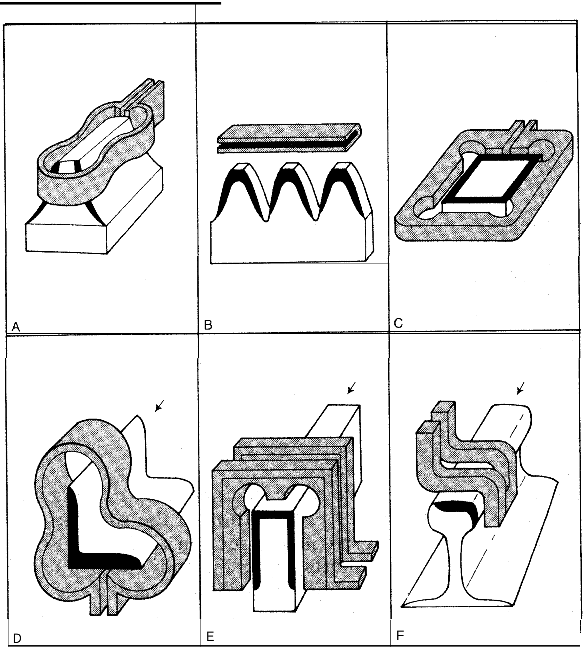


Figure 2: Representative inductor configurations for heating steel bodies of various shapes (see text).

being treated. The air gap between the inductor surface and the treated area is usually about two to five millimetres. When a vacuum-tube generator is used, a smaller gap can result in electrical breakdown between the inductor and the part being treated, due to thermionic emission (emission of electrons), which takes place at hardening temperatures.

An increase in the air gap results in increased losses as a result of diffusion of the electromagnetic field around the inductor and, other conditions being equal, reduces the intensity of induction heating.

Figure 2 indicates several characteristic inductor shapes for heating and tempering steel components of different configurations. The top three (A,B,C) are used for simultaneous heating and the bottom three (D,E,F) for heating parts while they are moving through the inductor field.

The cost of surface hardening depends largely on the required hardened layer thickness, type of apparatus, and degree of automation in the feeding of the processed parts. In the U.S.S.R., for example, the cost of hardening one square centimetre of surface to a depth of three millimetres can vary between 0.1 and 1 kopeck. The cost of heating a uniform mass-produced item is lower because of mechanization and the automation of all processes. Surface hardening of individually produced parts is more expensive than that of mass-produced parts, since the specific cost of preparatory operations (manufacture of inductor and equipment for holding the parts, time spent in readjustment, etc.) must be higher.

Through heating. High-frequency electrothermics—induction heating of billets (blocks of metal) for stamping and forging—is a standard operation in industry. The advantages include a sharp reduction of the heating cycle to a few minutes as compared to long heating required in fossil-fuel furnaces, thus lowering losses due to the formation of scale. Further, the method nearly doubles the service life of dies and reduces the margin on the subsequent treatment of the forgings. Moreover, the technique makes it possible completely to automate the preparation of billets for forging and, consequently, to produce uniform high-quality products. Another valuable feature is the possibility of reducing by up to 50 percent the floor area required in the production areas of the forging mills because of acceleration of the heating cycle and because the high-frequency equipment requires much less area than large furnaces. The power consumption needed in the prolonged process of heating up the furnaces is also eliminated, and a substantial improvement in the working conditions of the forging and heating workers is achieved.

A graph for the approximate determination of the frequency range for through induction heating for stamping and forging structural steel billets measuring between 10 and *200* millimetres in diameter is given in Figure **3.**

The induction method can also be used successfully for heating billets of nonferrous metals such as aluminum, copper alloys, etc. In these uses, heating efficiency will be less than for steel, and the optimum frequency will be lower. For a given frequency, the diameter of an aluminum billet would, thus, be much smaller than that of a steel billet.

The approximate time required for the through induction heating to forging temperature of carbon steel varies from *20* seconds for a ten-millimetre-diameter billet to five minutes for one of 100-millimetre diameter. For alloyed structural steel, the time is about 20 percent greater.

Losses due to radiation from the heated surface increase very rapidly as the temperature is increased; therefore, in selecting the heating regime, the shortest duration possible of the heating cycle must be achieved to reduce radiation losses.

The average power consumption for heating one kilogram of steel is approximately one-half kilowatt-hour. The required generator power can be determined with the formula:

$$P_g = 30\,G/t \text{ kilowatts} \qquad (6)$$

in which G is the weight of the billet, and t, the heating time in minutes.

Adapted from M.G. Lozinskii, *Industrial Applications of Induction Heating* (1970); Pergamon Press Ltd.

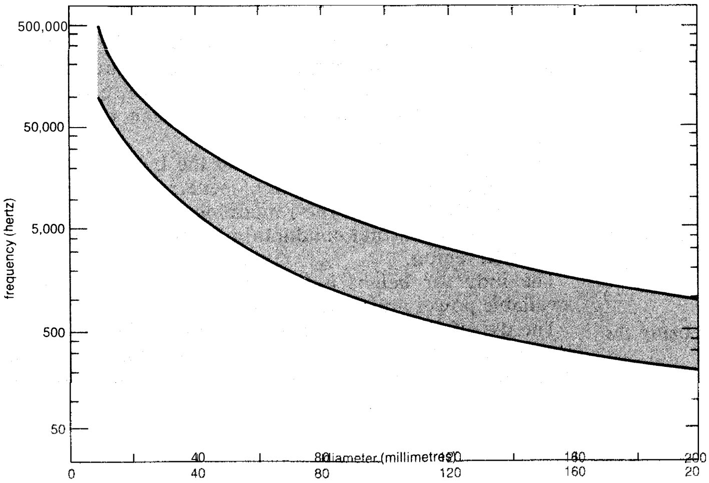


Figure 3: Recommended frequency ranges for heating solid cylindrical steel billets of various diameters.

Heating for soldering and brazing. Joining of steel parts after induction heating by means of high-speed soldering is a widely employed technique. Such soldering, when carried out with hard solders, retains and even increases the strength of the parts being joined while substantially reducing the consumption of materials for their production.

For through heating for soldering solid cylindrical parts, the choice of frequency is the same as for stamping and forging heating. Correct frequencies for cases in which it is necessary to heat a part of a hollow cylindrical steel product for soldering are shown in the graph in Figure 4.

The strength of the joint made by high-frequency soldering depends largely on the cleanness of the soldered surfaces, the gap between them, correct choice of heating routine (duration of the process and heating temperature), quality of the solder and flux (the material that cleans and prepares the surface for soldering), and subsequent heat treatment.

Flux

The quality of the flux is of great importance in high-frequency soldering. Calcined (anhydrous) borax, finely pulverized and mixed with solvent to a pasty consistency, is widely used. The use of pastry flux instead of powder permits precise dosing and preparation of the parts to be joined long before soldering.

The soldering of both hard alloy plates and high-speed (heat-resistant) steel plates to tool mounts and milling cutter bodies is employed extensively in industry.

Copper is usually used for soldering hard alloys and a flux consisting of 50 percent pulverized ferromanganese and 50 percent powdered glass is used in soldering high-speed steel plates. In the latter case the heating temperature is about 1,300"–1,350" C (2,360"–2,450" F), which makes it possible to follow the soldering operation with the hardening of the soldered plates (the haidening temperature of high-speed steels falls within the above-specified temperature range).

Tempering** of **steel. Local tempering to reduce residual stresses and also to decrease the hardness of the steel can be carried out successfully after local surface hardening by the use of high-speed induction heating. Short-term (for a few seconds) heating of the zone to be treated up to a temperature of, for example, 500" C will produce the same results as two hours of tempering in an oven at 150°–180° C.

In many cases, "self-tempering" can be achieved after surface hardening. For this purpose the heat intensity and duration of cooling are both determined in such a way that the heat remaining in the part after termination of cooling will be sufficient to lower the residual stresses and hardness of the hardened zone to within the required limits.

Adapted from M.G. Lozinskii, *Industrial Applications of Induction Heating* (1970); Pergamon Press Ltd.

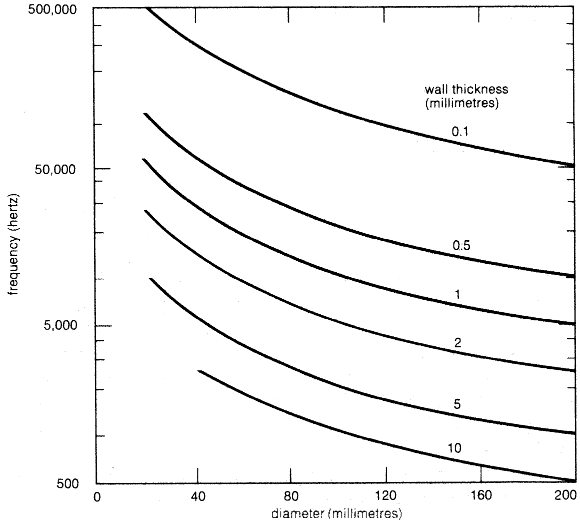


Figure 4: Optimum frequencies for heating hollow steel cylinders of various diameters and wall thicknesses preparatory to soldering.

The same inductors that are employed for hardening are also used for tempering. Heating regimes for tempering and the duration of the cooling process in hardening with subsequent self-tempering are determined by experimentation.

Induction furnaces. Metallurgy is the oldest branch of the comparatively new induction heating technology. The production of magnetic and heat-resistant alloys, and of other alloys with various special properties in an induction furnace, is an important part of metallurgy.

The shape of the induction heating coils and the frequency requirements for an induction furnace are similar to those used for through heating of billets. The average specific power consumption for furnaces with a capacity of 150–200 kilograms and more of steel is about 0.6 kilowatt-hours per kilogram and increases somewhat for smaller furnaces.

DIELECTRIC HEATING

Basic principles. The dielectric ("capacitance") heating method is based on the utilization of heat created in materials that are poor electrical conductors, when placed in high-frequency electromagnetic fields. The heat is formed as a result of losses that occur in a material located between metal walls that form a sort of capacitor connected to the high-frequency generator. The amount of the loss is a characteristic of the material called its loss factor, or "loss angle."

"Loss angle"

In contrast to induction heating, in which nonuniform heating is possible, capacitance heating provides comparatively uniform through heating of objects made of various dielectrics (wood, plastic, rubber, food products, etc.), which are located in the high-frequency field of the capacitor.

The dielectric heating method has been successful in solving problems in which it was necessary to heat poor conductors. Depending on the structure and physical properties of these poor conductors, it is quite possible to select the optimal frequency and applied voltage for heating them.

Heating insulating materials and poor conductors usually requires substantially less specific power than heating metals. Thus, to heat a layer of metal of a thickness, for example, of 3 millimetres for surface hardening, a thousand or more watts of high-frequency energy are required for each square centimetre of surface. For the high-frequency drying of large volumes of wood, on the other hand, the average power is usually less than one watt per square centimetre. But because of the high electrical resistance of the insulators, generating the necessary heat requires much higher frequencies than in the induction heating of metals.

At low frequencies even high voltage does not ensure the required power. Therefore, the frequencies used for dielectric heating vary from hundreds of thousands of hertz to many gigahertz (1,000,000 hertz).

A type of electron tube known as the magnetron is widely used in industry to generate high-frequency power in the gigahertz range. Developed about 40 years ago, magnetrons were used during World War II in radar generators, combining high output powers with extremely high frequencies (see also ELECTRON TUBE). At the present time there are magnetrons capable of continuous operation with effective power outputs of the order of several kilowatts at frequencies in the 3,000 to 30,000 gigahertz (and even higher) range.

Applications. Of many industrial applications of dielectric heating, only a few will be discussed.

Drying of wood. The average power consumption in evaporating one kilogram of water from wood is about 2.5 kilowatt-hours. To determine the parameters of the apparatus, it is necessary to determine the weight of the water to be removed during drying on the basis of the job specifications.

High-frequency drying is normally used only for the final processing of lumber that has been seasoned in kilns, containing, for example, 20 percent moisture. The final moisture content in such cases is usually about 10 percent. Thus, the power and time requirements for a drying operation are directly related to the amount of moisture represented by this drop from 20 percent to 10 percent.

Cracking

The duration of the drying process, however, is controlled by the danger of cracking the lumber due to the formation of steam and the rupture of the fibres. In general, the time will be around one hour per ten millimetres of thickness of the boards. For example, a board 40 millimetres thick should dry in not less than four hours, and a board 20 millimetres in thickness may be dried in two hours.

Heating of ceramic products. Many ceramic firms employ dielectric heating of ceramic products during their production cycles. Capacitor-type heaters are used with the ceramic products placed between the plates of the capacitor. In many cases a conveyer belt is used to transport the products between the plates of the capacitor, thus making it possible to automate the process.

Confectionery and culinary industries. One of the relatively new applications of the dielectric heating technique is in the confectionery and culinary industries. In these applications the product is placed between capacitor plates connected to a radio-frequency generator. The heated products never come into contact with the conductors; the heat generated within them is proportional to the frequency and depends on their physical properties.

Baking and preparing roasted meat dishes by no means exhaust the culinary possibilities of this technique; it is widely used in the food industry, and equipment is available for both restaurants and homes.

Melting glass and other dielectrics. Many materials that are good dielectrics at room temperature become conductors when heated. Glass, for example, heated close to its melting point, about 800" C, becomes a comparatively good conductor.

Radio-frequency heating can be used successfully for melting glass and glasslike materials (including various types of rocks), but the process is not effective until the temperature is increased to a point at which the material becomes conducting. Thus to start melting, a small portion of the material must be heated close to the melting point by some other heat source, such as gas burners. The high-frequency power is then switched on and the preheated conducting zone begins to increase its temperature and goes into the liquid state.

A valuable feature of the above-described method is the possibility of enlarging or reducing the dimensions of the molten zone, depending on the amount of generator power. A dynamic equilibrium is established; there is a certain minimum power below which the process does not take place, and the molten material hardens. In this regard the heating of glass differs substantially from the high-frequency heating of metals, which may be heated as much or as little as desired.

BIBLIOGRAPHY. G BABAT and M.G. LOZINSKII, "La Trempe superficielle de l'acier par chauffage au moyen de courants à haute fréquence," ***Revue Gén. Élect.***, 44:495–510 (1938), one of the first works to contain a description of induction heating of metals and its industrial applications; V.P. VOLOGDIN, *Поверхностная закалка индукционным способом* (1939), a discussion of various features of induction heating of steel, including some information on apparatus employed; G.H. BROWN, C.N. HOYLER, and R.A. BIERWIRTH, ***Theory and Application of Radio-Frequency Heating*** (1947), a discussion of the theory and application of radio-frequency heating in the metal industry; L.L. LANGTON, ***Radio Frequency Heating Equipment*** (1949), a detailed account of the principles of induction heating; F.W. CURTIS, ***High-Frequency Induction Heating,*** 2nd ed. (1950), a detailed account of the early history of induction heating and a description of plants utilizing it; A.M. THOMAS, ***Differential Effects in High-Frequency Dielectric Heating,*** British Electrical and Allied Industries Research Association Report WT/20 (1951), a report on plants utilizing dielectric heating in materials that are poor conductors of electricity; J.W. CABLE, ***Induction and Dielectric Heating*** (1954), a detailed account of the methods of induction and dielectric heating of various objects and the types of equipment used; D.A. COPSON, ***Microwave Heating: In Freeze-Drying, Electronic Ovens, and Other Applications*** (1962), a ,study of the application of microwave heating; H. PUSCHNER, ***Wärme durch Mikrowellen*** (1964; Eng. trans. ***Heating with Microwaves,*** 1966), a report on microwave heating equipment and a description of its applications for cooking; M.G. LOZINSKII, ***Industrial Applications of Induction Heating*** (1969; pub. orig. in Russian, 1958), a basic discussion of the theory and practice of induction heating, with a description of the types of equipment and inductors used.

(M.G.L.)

Radioisotopes, Applications of

Isotopes are atoms of an element that are slightly different structurally from each other but have the same chemical properties. For example, the element tin has ten different stable isotopes of varying atomic weight. Some isotopes are stable; that is, they retain their character in all normal circumstances. Certain isotopes, however, are unstable, or radioactive, because the ratio of neutrons to protons is either too low or too high for stability; that is, they tend to break down spontaneously into entirely different elements, giving off radiation in the process. Isotopes of all of the elements heavier than bismuth are radioactive; for a detailed discussion see RADIOACTIVITY.

Radioactive isotopes, or radioisotopes, have proved to be valuable tools in the laboratory, in industrial plants, and in agriculture. The radiation they give off, called alpha, beta, and gamma rays, can be readily detected by such detectors as Geiger counters, scintillation counters, and the like. Since the rates of emission of the rays and their penetrating abilities are known and since their intensities are reduced when they travel through materials, they can be used in making highly accurate scientific measurements, in testing the operation of mechanical systems without disassembling, in tracing flow in pipes, in measuring nutrient intake of plants, and in innumerable other applications.

Radiation doses must be controlled within amounts judged safe to the personnel involved in manufacturing processes and to the consumers of the products having residual radiation.

Residual radiation in products

In the United States, federal government approval has customarily been granted by the Atomic Energy Commission (AEC) to specific products whose radiation does not exceed that of cane sugar sold in stores. The AEC has ruled that products are considered safe for release to the public when radiation does not exceed the concentration established many years ago for water.

Disposal of radioactive waste into sewers or the atmosphere may come under government regulation in the future.

Potential health hazards are discussed fully in the article RADIATION INJURY.

EXPERIMENTATION

First use

The first experimental use of radioisotopes occurred in Austria in 1913 just ten years after the award of the Nobel Prize to Henri Becquerel and to Pierre and Marie Curie for the discovery of radioactivity. Hungarian physicist George Charles de Hevesy, working in the Vienna Institute for Radium Research, used what was then known as "radium D," actually an isotope of lead (lead-210), to study the solubility in water of lead sulfide and lead chromate. A little later he used "thorium B (lead-212) to study the uptake of lead by growing plants.

For the next several years a variety of experiments were carried out with naturally occurring radioisotopes in biology and medicine. The invention of the cyclotron in the early 1930s brought a new source of isotopes, but they became plentiful only with the development of nuclear reactors in the 1950s. Measurement and control systems based on radioisotopes were then devised, and new fields of application for the radiations opened up: food processing, medical-supply sterilization, and sources of yower and heat. In the early 1970s tracer applications were increasing steadily; food processing and other process-radiation applications also were expanding.

Naturally occurring radioisotopes. About 55 radioactive isotopes can be found in nature and about 20 more radioactive-decay products are always associated with some of them. They range from traces of carbon-14 and potassium-40 in the body to radium, thorium, and uranium, which are mined commercially. Table 1 lists these nuclides or atoms of a specific isotope with their half-lives (the time required for half the atoms in a given quantity of material to decay) and natural abundances. Most naturally occurring radioisotopes are insignificant as sources of radioactivity either because they are not abundant or because they have low activities as a result of their long half-lives. The exceptions are radium, used in medical therapy, and uranium and thorium, used in the nuclear industry as fuel and in "breeder" reactors, *i.e.*, reactors that produce more fissionable material than they burn.

Table 1: Naturally Occurring Radioactive Isotopes

isotope	abundance (percent)	half-life (years)	isotope	abundance (percent)	half-life (years)
^{14}C	trace	5.8 X 10^{3}	^{149}Sm	13.83	4 × 10^{14}
^{40}K	0.012	1.3 X 10^{9}	^{152}Gd	0.20	1.1 X 10^{14}
^{48}Ca	0.18	>2 X 10^{16}	^{159}Tb	100	5 × 10^{16}
^{50}V	0.24	6 × 10^{14}	^{165}Ho	100	6 × 10^{16}
^{64}Zn	48.89	1 × 10^{15}	^{169}Tm	100	5 × 10^{16}
^{70}Zn	0.62	1 × 10^{15}	^{175}Lu	97.41	1 × 10^{17}
^{76}Ge	7.76	2 × 10^{16}	^{176}Lu	2.60	2.1 X 10^{10}
^{82}Se	9.15	1 X 10^{17}	^{174}Hf	0.18	4.3 X 10^{15}
^{87}Rb	27.85	4.7 X 10^{10}	^{180}Ta	0.012	I × 10^{13}
^{96}Zr	2.80	3 × 10^{17}	^{180}W	0.14	3 × 10^{14}
^{92}Mo	15.84	4 X 10^{18}	^{182}W	26.41	2 × 10^{17}
^{100}Mo	9.13	3 × 10^{17}	^{183}W	14.40	1 × 10^{17}
^{113}Cd	12.26	1 X 10^{15}	^{186}W	28.41	6 X 10^{15}
^{116}Cd	7.58	1 × 10^{17}	^{187}Re	62.93	7 × 10^{10}
^{115}In	95.72	6 X 10'4	^{192}Os	41	1 × 10^{14}
^{124}Sn	5.94	2 × 10^{17}	^{190}Pt	0.013	7 × 10^{11}
^{123}Sb	42.75	I × 10^{16}	^{192}Pt	0.78	1 X 10^{15}
^{123}Te	0.87	1.2 × 10^{13}	^{198}Pt	7.21	1 × 10^{15}
^{130}Te	34.48	8 × 10^{20}	^{196}Hg	0.146	1 × 10^{14}
^{138}La	0.09	1.1 X 10^{11}	^{209}Bi	L00	1 X 10^{14}
^{136}Ce	0.193	3 X 10^{11}	^{222}Rn	from ^{226}Ra	3.8 days
^{142}Ce	11.07	5 × 10^{15}	^{226}Ra	from ^{230}Th	1622
^{141}Pr	100	2 X 10^{16}	^{230}Th	from ^{234}U	8 × 10^{4}
^{144}Nd	23.85	5 X 10^{15}	^{232}Th	100	1.4 × 10^{10}
^{150}Nd	5.62	1 × 10^{16}	^{234}U	0.006	2.5 X 10^{5}
^{147}Sm	14.97	I X 10^{11}	^{235}U	0.72	7.1 X 10^{8}
^{148}Sm	11.24	1.2 × 10^{13}	^{238}U	99.27	4.5 X 10^{9}

Manufactured radioisotopes. When atoms of uranium-235 or plutonium-239 are bombarded with neutrons in a controlled chain reaction, a series of fission fragments is produced (see also NUCLEAR REACTOR). Many of these particles are themselves radioactive and continue to decay. Some, either stable or with very long half-lives, remain in the fission products; and the individual elements can be separated from the rest of the fission products.

The fission products are by analogy the ashes from the burning of nuclear fuels. Finding uses for them is extremely important because as nuclear wastes they must be disposed of. Table 2 lists some important fission products along with their abundances and half-lives. The first artificial radioisotopes were identified in the early 1930s as a result of transmutations (nuclear reactions changing one chemical element into another). By 1935 the University of California cyclotron, the earliest type of charged-particle accelerator, was being used to make a number of radioisotopes. After World War II, neutron-produced radioisotopes became available. Since then, production of radioisotopes has increased enormously.

Table 2: A Few Important Fission Products

nuclide	half-life	fission yield (percent)	nuclide	half-life	fission yield (percent)
^{85}Kr	10.4 years	0.3	^{106}Ru	1 year	0.38
^{89}Sr	50.4 days	4.6	^{131}I	8.05 days	2.8
^{90}Sr	28 years	5.3	^{133}Xe	5.27 days	4.5
^{91}Y	57.5 days	5.9	^{137}Cs	30 years	5.9
^{95}Zr	65 days	6.4	^{140}Ba	12.8 days	6.1
^{95}Nb	35 days	6.4	^{144}Ce	285 days	6.1
^{99}Tc	2.1 X 10^{5} years	6.1	^{147}Pm	2.5 years	2.6
^{129m}Te	33 days	0.19			

First identification of artificial radioisotopes

Target atoms to be bombarded with neutrons may be sealed in quartz vials that are welded into aluminum containers and lowered into the core of a nuclear reactor. After bombardment the products are processed to a form suitable for ultimate use. This processing may involve simply dissolving the irradiated material or may require a more complex treatment. In all cases handling is carried out remotely in hot cells to protect workers from radiation (see Figure 1).

BY courtesv of Oak Ridge National Laboratory, Oak Ridge, Tennessee

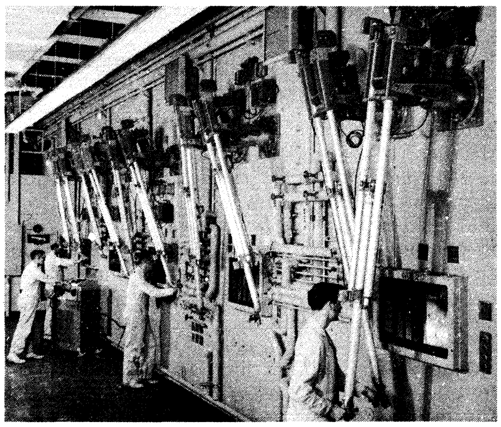


Figure 1: Hot cells at the Oak Ridge National Laboratory for processing fission products.

Reactors produce only a limited number of isotopes with a low ratio of neutrons to protons; charged-particle accelerators are a more efficient means of producing such isotopes. Energies of the accelerators can be varied over wide ranges, and the nuclear reaction can be controlled.

Because most elements contain several isotopes, their bombardment often results in mixtures containing undesirable radioisotopes. Targets are therefore usually enriched with a specific isotope, resulting in the production of a very high proportion of the desired radioisotope.

ISOTOPIC TRACERS

All uses of radioisotopes depend upon the radiations that result from the decay of the unstable nuclei. When a radioisotope, or radioactive element, is added in small smounts to relatively large amounts of the stable element, it will behave chemically exactly the same as the stable species, but it can be tracked or traced by a Geiger counter or other device. This characteristic leads to the largest variety of radioisotope applications, including studies of wear, tracing of fluid flow and locating leaks in pipes and hydraulic systems, analyzing metabolism of foods and drugs, and studying fertilizer uptake by plants.

In agriculture and animal husbandry. In the mid-1930s radioisotopes began to be used in studies to trace the uptake of fertilizer. The addition of phosphorus-32 to phosphate fertilizers allowed the study of rate of fertilizer uptake and its distribution in the plant. As a result, it was discovered that application of nutrients on leaves is about 95 percent efficient but application on roots is only about 10 percent efficient.

Radioisotopic tracers have been used to help evaluate the effects in plants of growth stimulators, such as auxins and gibberelins, the function of certain proteins and nucleic acids, and the effects of plant-growth regulators. Thousands of plant metabolic studies have been carried out on amino acids and proteins, sulfur, nitrogen, and phosphates. One of the most significant early applications, involving carbon-14, assisted by the stable isotope oxygen-18, was in the study of photosynthesis, at the University of California, beginning about 1948. These studies revealed that none of the oxygen released during photosynthesis comes from the carbon dioxide, but instead from the water.

Photosynthesis studies

In the evaluation of insecticides, pesticides, and weed killers, it is desirable to follow the life cycles of organisms to determine specific effects of the chemical on the biological system. Such isotopes as carbon-14, sulfur-35, phosphorus-32, and iodine-131 have been useful in this work.

Many radioisotopes have been involved in investigating the body processes and habits of animals. Weight gain as related to types of feed and to hormones, the significance of trace elements and vitamins in diets, and the effect of tranquillizers in reducing weight loss during shipment have all been studied with isotopes.

Radioisotopes also are used to tag insects, animals, and fish for study of their life cycles; migratory patterns and ranges; mating and feeding habits; and identification of parasites, hosts, and prey.

In medicine. Some of the earliest uses of radioisotopes were as tracers in medical diagnosis and in metabolic research. More than 100 radioisotopes have been used in medicine since about 1945, though five have been used for diagnosis much more than the others: chromium-51, iodine-131, phosphorus-32, iron-59, and technetium-99*m*. Chromium-51 is used for measurement of total volume and residence time of red blood cells; measurement of total volume of plasma and blood; measurement of cardiac output; location of the exact position of the placenta in expectant mothers; determination of amount and location of bleeding from the gastrointestinal tract. Iodine-131 is used for determination of blood volume, cardiac output, plasma volume, liver activity, fat metabolism, thyroid-cancer metastases, brain tumours, and particularly the size, shape, and activity of the thyroid gland. Phosphorus-32 can help identify tumours because cancerous cells tend to concentrate phosphates more than do healthy cells. Iron-59 is used for measurement of the rate of formation of red blood cells, their lifetime, and their volume; and for measurement of absorption of iron by the digestive tract to evaluate the effectiveness of various iron compounds in revitalizing blood. Technetium-99*m* is extremely valuable for diagnosis of brain tumours using scanning devices.

Other radioisotopes used frequently in medicine are sodium-24 in circulatory studies; cobalt-58 and cobalt-60 added to vitamin B_{12} in diagnosis of pernicious anemia; arsenic-74 and copper-64 in locating tumours; hydrogen-3 (as tritiated water) in measuring total body water.

In the pharmaceutical field, isotopes are used in studying the reactions involved in making new compounds.

In industry. During chemical or physical processing, the radioisotopes retain their chemical identities, and the process streams can be followed regardless of changes in the system or its components. Radioisotopes are widely used in industry to trace the flow of materials; *e.g.*, the flow of a lubricant in a sealed engine; often there is no other way to do this. In the 1970s, the continuous use of safe but detectable amounts of radioisotopes promises to become routine for many industrial processes and products.

In blast-furnace operations, iron-59 has been used to study distribution of constituents and their residence times in a number of metallurgical processes. The use of radioisotopes results in improvements in mixing during formation of alloys.

Wear studies

Tracers yield information about the rate of tool wear in punching and machining operations and on the wear of gear trains and similar moving parts, permitting evaluation of the characteristics of alloys and the stability of cutting oils. Automobile manufacturers can test such components as engine cylinders and pistons, piston rings, bearings, and lubricating oil by irradiating piston rings and cylinder liners (iron-59) or bearings (copper-64 and zinc-65) and noting the change in radioactivity in the lubricant. In contrast to other wear-measuring techniques, radioisotopes can be used to test the system while operating without disassembling, so that effects of idling, acceleration, speed, starting, stopping, coolant temperature, and the viscosity and composition of the lubricant can be related directly to friction and wear.

One of the original applications of radioisotope tracing was the location of obstructions in underground pipelines. An isotope source, mounted in a small rubber ball or other suitable carrier, is introduced into the blocked pipe. It moves along until it reaches the obstruction, where it stops. It can then be located exactly from above ground by use of a radiation detector, and the blocked section can be replaced without costly searching. Leak testing is extremely important in industry; in underground piping, iodine-131, bromine-82, and sodium-24 are used; in heat exchangers, hydrogen-3 (tritium) is used; in telephone cables, krypton-85.

Other typical industrial tracer applications include measurement and control of the time a chemical remains resident in a solution, uniformity of mixing components, catalyst flow, and efficiency of separation, as in distillation. Important information about hydrodynamic relationships in oil-bearing geological formations and about flow patterns in subterranean rock formations may be obtained by using hydrogen-3 and iodine-131 tracers. This information is useful because. oil yields can be increased by water-flooding of wells. Tracers are used in transcontinental oil lines to identify the boundaries between different types of oil being pumped through the line.

In environmental and ocean studies. Tracers are widely used in studies of air pollution from smokestack-gas effluents and of water pollution in lakes and rivers and are also used to measure deepwater currents in lakes and oceans.

Sand and silt labelled with the isotopes gold-198 or xenon-133 are being used to study drift and transport of sand and silt in waterways, thus making the maintenance and restoration of beaches and harbours more efficient.

Carbon dioxide with carbon-14 is used to study photosynthesis in marine algae, providing information important to long-range sea-farming plans.

Snow-management research

Radioisotopes have been used in gauges to monitor snow density and as tracers to study water movements in snowpacks; snow-management research is designed to provide ways for delaying snow melt, measuring snow-water content, and increasing total-water content from snow- and watersheds. Sediment density in rivers and streams can be measured continuously with radioisotopes. A nuclear sediment-density gauge has been developed to give a continuous measurement of suspended matter in rivers and streams, thereby providing data as to what happens during and immediately after heavy rainfall.

Tracers are also used in many other areas of research, such as the kinetics of reactions, reaction mechanisms, diffusion, permeability of membranes, sorption on surfaces, vapour pressures of metals, and the mechanism of passivity, which makes metals resistant to certain environments.

USE OF DIRECT RADIATION

Direct application of penetrating radiation from radioisotopes has found increasing applications in medical therapy, in agriculture for pest-control purposes, and in industry.

In medicine and biology. The therapeutic uses of X-rays—and now of radioisotopes—are based upon the facts that these radiations cause ionization of the matter through which they pass and that malignant or otherwise diseased cells are usually more affected by this radiation than are healthy cells. Diseased cells may be killed by the radiations, while nearby healthy cells, although damaged, usually recover. Cobalt-60, iridium-192, and cesium-137 have been used primarily to replace X-ray therapeutic units, with the advantage that rotational isotope units that minimize damage to the skin and other healthy cells are less cumbersome than similar X-ray machines. A more subtle advantage of radioisotopes is that they can be used on localized areas or organs. For example, iodine concentrates in the thyroid and can be tagged with radioactive iodine for radiation in the thyroid to reduce the activity of that gland. Phosphate containing phosphorus-32 concentrates in tissues that utilize phosphates in their metabolism, including the blood-forming tissues. This concentration provides a treatment for *polycythemia vera*; the hematopoietic, or actively forming, cells are sensitive to radiation and are selectively destroyed. Gold-198 in the form of a colloid (finely distributed in a liquid) has been used in palliative treatment to control the accumulation of excess fluid in the chest and abdominal cavities from their linings as a result of the growth of some types of malignant tumours. The treatment is not curative, but it is not as dangerous or uncomfortable as frequent surgical drainage. Beads and needles containing, or made of, cobalt-60, yttrium-90, or gold-198 can be implanted directly in tumours to increase the effectiveness of radiation and to cut down on peripheral cell damage.

Medical sterilization

Biocidal uses. A major use of gamma radiation is in the medical- and hospital-supply business for sterilizing bandages, surgical sutures, plastic blood-donor kits, rubber gloves, disposable hypodermic syringes, catheters, surgical blades, needles, and other instruments. This method of sterilization eliminates the need for boiling, autoclaving, or chemical treatment, and materials can be packaged before sterilization instead of aseptically afterward. The process for the sterilization of catgut sutures is particularly complicated when heat treatment is used but simple with gamma radiation.

One of the world's largest cobalt-60 irradiation facilities, located at Dandenong, Australia, is used for the disinfestation of goat hair. The radiation destroys the anthrax spores but does not harm the fibres.

For insect eradication. Ionizing radiations may be used to eradicate insects by sexual sterilization. The classic application concerns the elimination of the screwworm fly, indigenous to large parts of the southern U.S., the Caribbean, and Mexico. The fly lays its eggs in the navels of newborn animals or in open wounds of livestock; the resulting maggots burrow into the animals, killing them. In the southeastern U.S., the annual loss varied from $25,000,000 to $50,000,000. In the pupal stage, these insects may be sexually sterilized by low doses (2,500 roentgens) of radiation. Sterile males compete on a par with normal males for mates, and females mate only once. Saturation of an area with up to 50,000,000 sterilized males at a time rapidly reduced to zero the number of eggs that hatched from the normal native flies. Over a period of a year and a half, during which more than 2,000,000,000 sterilized male flies were released, the screwworm fly was completely eliminated.

Industrial uses. Radioisotopes can be used as a source of gamma rays for radiography. The principle is the same as in conventional X-radiography: the radiations of short wavelength can penetrate materials opaque to ordinary light, shadowing the more dense parts on photographic film to produce the image, or radiograph.

Radioisotopes have certain advantages over conventional X-ray sources: portability, freedom from external power supplies, applicability to odd and complex shapes, and the capability for inside-out exposures. The large number of available isotopes provides a wide range of energies. A disadvantage is that the radioisotope source cannot be turned off. Point radioisotope sources yield radiographs with good resolution.

Portable radioisotope units can be used for taking radiographs of an accident victim on the scene before deciding whether or not he should be moved; one such iodine-125 unit weighs only two pounds.

Neutron radiography was discovered about 1930, but the practical applications became significant only in the late 1960s. The X-rays or gamma rays used to make conventional radiographs interact with orbital electrons, but neutrons react with the nucleus; and the radiographs made by the two methods show quite different details.

In luminescent devices. Just as electrons are used to excite the phosphors in a television tube, isotopic beta emitters are used in luminescent devices such as exit signs in aircraft and runway markers. The original application was in radium-activated luminescent watch dials. Some compasses have isotope-activated luminous dials. Illuminants such as tritium, krypton-85, and promethium-147 are independent of external power supplies and have a lower maintenance cost than conventional systems.

The radiations from radioisotopes are reduced in intensity when they travel through materials, the amount of the reduction depending on the thickness and density of the materials. This principle, involving a radioisotopic source and a radiation detector, is used to determine and control the thicknesses and densities of various substances and to locate and control levels of liquids and solids. Radiation detectors provide continuous, nondestructive measurements and can be used along with an automatic-control system for regulating thickness and density of manufactured products.

In gauging. The material being processed is assumed to have uniform density, so that any change in radiation intensity measures a change in thickness. An increase in radiation intensity indicates a decrease in thickness and vice versa. The radiation-detector signal may be used to adjust and control a thickness-regulating device, such as a roller setting. The method can be applied to metal, paper, plastic, or coatings. A technique known as backscattering may also be used to measure thickness, weight, or density. When radiation enters a material, some is reflected, or backscattered, and the amount can be related directly to thickness or density. This technique is particularly effective in estimating the extent of corrosion of boiler shells, pipe walls, and ships' hulls.

Radioisotope gauges are used in agriculture to measure eggshell thickness. The breakage of eggs during shipment is decreased by selection of thick-shelled eggs.

In gauging density, the dimensions of the material are kept constant so that changes in radiation intensity are caused by density changes. Densities of liquids, powders, granular solids, and slurries may be estimated by placing a source and detector on opposite sides of a pipe or container.

Isotope gauges are used in agriculture to measure silage density and to control the density of condensed milk and the air content of ice cream.

Level gauging

In gauging a level, the radiation detector senses the change that occurs when material is or is not in the container at the prescribed level. More radiation is transmitted through the container when the level of material is below the radiation source than when it is above and the radiation must pass through it. A practical application is the automatic coatrol of liquid filling devices such as are used in agriculture for filling milk cans and cartons. The detector is placed at the level to which the container is to be filled, and when the product reaches this level, the change in radiation intensity can be used to shut off the filling device.

As an inspection-control mechanism, a detector may be set at a level just below that to which the containers must be filled; any container that does not meet the minimum-fill specifications will allow a surge in radiation that can be used to activate a reject mechanism.

Polymerization reactions

In chemical processing. In one of the first commercial applications of ionizing radiations from a radioisotope source to initiate a chemical reaction, cobalt-60 has been used since 1962 to produce ethyl bromide by the reaction between ethylene and bromine. Later, a radiation-induced sulfoxidation of hydrocarbons produced biodegradable detergents that break down without leaving a pollution residue. Cobalt-60 can be used in a continuous industrial chemical process for polymerizing ethylene to yield a basic raw material for packaging, laboratory tubing, and many other uses. Radiation can induce chemical cross-linking (bonding) in polymers (chemical compounds consisting of repeated structural units); polyethylene cross-linked by radiation shows increased mechanical strength at elevated temperature, and irradiated stretched tubing can be shrunk by heating to give adherent insulation.

Organic monomers (compounds that may become a structural unit in a polymeric compound) may be polymerized within the structure of wood to give wood plastics. Wood is placed under vacuum, the pores are filled with a monomer such as monomethylmethacrylate, and the combination is exposed to gamma radiation from a radioisotope source. Not only is the monomer polymerized, but some of it is apparently grafted to the cellulose of the wood, forming a new product with improved abrasion and water resistance, dimensional stability, and shear- and static-bending strength. These wood-polymer combinations are particularly suited for use as veneers, in certain types of furniture, parquet flooring, and such specialty products as golf club heads, billiard cues, and rifle stocks. A similar treatment gives concrete-polymer combinations that have improved qualities over untreated concrete. A possible use of this technique would be to form lightweight building blocks from compacted refuse–polymer composites, a process that might help solve the solid-waste-disposal problem.

In chemical analysis. Radioisotopes may be used as a source of radiations in devices for chemical analysis of samples. Such a device carried by the space satellite Surveyor 5 used alpha particles from curium-242 and reported on the composition of the surface of the moon.

Use in establishing age of artifacts

Oil paintings can be identified as to the composition of the paint used by a technique called activation analysis, in which they are exposed to a neutron beam and then placed on one photographic film after another for short periods of time. Since the radiations emitted from different elements within the painting have different properties and decay at different rates, they may often be distinguished by the varying images they induce in the chronological series of photographic films.

Pieces of ancient pottery, ancient coins, and other objects can also be identified chemically by activation analysis: the object is exposed to neutrons and the radioisotopes formed from the materials present, including any trace of impurities, usually give a positive identification of the composition of the original object.

The fact that extremely small amounts of certain elements can be detected and identified nondestructively by the activation-analysis technique has been used by law enforcement agents in various ways. Human hair contains small traces of metallic elements, but the quantities vary from person to person. The actual quantities can be determined by activation analysis. A murder conviction in Canada was based partly on activation-analysis evidence that a hair found in the hand of the murder victim matched the suspect's hair with respect to the proportion of metallic elements. Activation analysis of wipings taken from a suspect's hands indicates not only if he has fired a gun recently but also the number of bullets fired and the type of ammunition used. In another case, activation analysis of soil specimens from the scene of a crime and soil residue on a vehicle in the possession of a criminal suspect established a link with the site of a crime 1,000 miles away.

OTHER USES

Carbon dating. Dating, the determination of the age of objects from their radioactivity, is treated fully in the article DATING, RELATIVE AND ABSOLUTE. The use of naturally occurring carbon-14, an excellent method of estimating ages of objects up to about 50,000 years, was developed by U.S. physicist Willard Libby about 1946 and is based on the assumption that natural carbon dioxide in the air contains traces of cosmic-ray-produced carbon-14, but in rather fixed amounts determined by the rate of production of carbon-14 from nitrogen and the rate of decay of carbon-14 already in existence. Because plant life grows by photosynthesis of atmospheric carbon dioxide and animals feed on the plants, living organic matter is also in equilibrium with the carbon-14 in the atmosphere. Once organic matter has ceased to grow, however, there is no replenishment of the carbon dioxide, and the amount of carbon-14 gradually decreases as it decays. Thus the amount of residual carbon-14 when compared to the carbon-14 in living matter permits an estimation of the date at which the object died, as, for example, when a tree was cut.

Plant mutation. Radiation can be used to change the molecular structure of plant genes, inducing mutations in the plants. Successful mutations are rare, but experimentation in this area is simple. Studies since the early 1930s have yielded results throughout the world in producing millions of hectares of superior crops. Nearly 50 new varieties of plants so produced were released to growers in the late 1960s. Crop varieties have been improved and shortages caused by disease have been averted by using radiation to induce mutations. The technique is proving valuable to breeders of flowers and ornamental plants as well as in agriculture.

In Japan the higest yielding rice variety, the Reimei, has been produced by treating dry seeds with gamma rays from cobalt-60. The success of this technique has opened the possibility of improving even high-yielding crop varieties. Radiation-induced mutations may alleviate some population-growth problems.

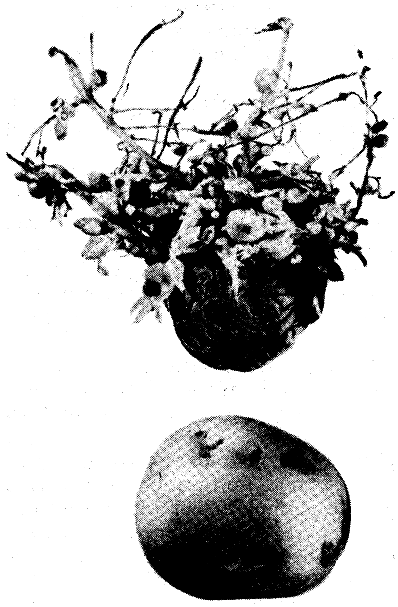


Figure 2: Sprout inhibition of potatoes by gamma irradiation using cobalt-60. (Top) Not treated. (Bottom) Exposed to 20,000 rads of gamma radiation. Photograph taken after both were stored for 16 months.

By courtesy of Oak Ridge National Laboratory, Oak Ridge, Tennessee

Pasteurization

Food processing. Beta and gamma radiation may be used to pasteurize, sterilize, and preserve foods. Food irradiation is discussed fully in the article FOOD PRESERVATION. Radiation treatment can be applied to inhibit the sprouting of potatoes and onions (see Figure 2), to disinfest grain and grain products, and to extend shelf life of fruits, meats, and seafoods. Radiation treatment of foods is relatively cheap, rapid, and effective and requires no heat.

Heat and power. When radiation is slowed down or stopped by matter, the kinetic energy is converted to heat. By using suitable isotopes and shielding materials heat and power can be produced. The heat generated by radioactive decay may be used directly as was done in the lunar missions to keep equipment warm during the cold lunar nights; or the heat may be converted into electrical or mechanical energy. The advantage of isotopes as power sources lies in their long lives that make energy available over long periods without recharging. Many SNAP (Systems for Nuclear Auxiliary Power) devices involving radioisotopes are commercially available and can be used to power navigational lights and buoys or remote and unmanned weather stations. Units have also been carried to the moon to provide electricity for various scientific experiments.

In medical therapy. ***Heart pacemakers.*** The newest application of radioisotopes in the life sciences is as a source of power for cardiac support devices, in particular, for the cardiac pacemaker and the artificial heart involved in programs supported by the U.S. Atomic Energy Commission and the National Heart Institute of the National Institutes of Health. In the pacemakers, plutonium-238 is the isotope source, the heat from it being converted to electricity by means of thermoelectric junctions. In one case, the four-ounce unit supplies 160 microwatts of electric power, sufficient to shock the heart 70 times per minute for a minimum of ten years. The obvious advantage is that replacement of the pacemakers through implantation surgery will have to occur only once in ten years in contrast to every two years for ordinary battery-powered devices.

Circulatory-assist devices, the ultimate of which is a completely implantable artificial heart, involve the heat from radioisotopes to power a Stirling or Rankine-cycle engine that simulates the pumping action of the heart. British surgeons have successfully implanted radioisotope-powered pacemakers, and the French have implanted an isotope-powered artificial heart in animals.

Heat for life-support systems. A potential use of isotopic heat and power is in an integrated life-support system for aerospace application. The design of this unit anticipates the handling of human waste and spacecraft debris for a 180-day mission and includes a waste incinerator, a high-temperature urine-vapourizing unit, and a catalytic oxidizer–incinerator.

GROWTH OF THE RADIOISOTOPE INDUSTRY

National vroduction and consumption

The radioisotope industry is continually growing. In France the vroduction and sale of radioisotopes increased rapidly throughout the 1960s. In 1967 total sales were valued at more than 14,000,000 francs ($3,000,000), excluding taxes, nearly half of which were to other countries.

In East Germany, from 1958 until 1968, the radioisotope centre at Rossendorf made more than 30,000 deliveries of radioactive products, of which almost half were exports. This material represents a total radioactivity of 1.3 kilocuries, valued at 8,000,000 marks. The centre is a leading supplier of radioactive materials in Europe.

In Argentina, from 1954 until 1968, consumption of radioisotopes increased from about 8 curies to 85 curies. In Brazil consumption increased at approximately the same rate. Argentina may soon produce all its own radioisotope requirements, except for high-intensity radiation sources, and may also export radioisotopes to other Latin-American countries.

In Australia, in fiscal 1969, the total income from sales of Australian-produced radioisotopes was $236,000 (U.S. equivalent $260,000), more than double that for 1968. Demand increased particularly for medical products.

In India there has been a steady growth in the use of radioisotopes since 1958. In 1968 about 18,000 consignments of radioisotopes valued at 2,200,000 rupees (U.S. equivalent $300,000) were supplied to 325 institutions, mainly for work in medicine, agriculture, industry, and hydrology.

In Japan the total number of establishments licensed to use radiation increased from 50 in March 1959 to 1,540 in March 1968; the greatest percentage increase was accounted for by hospitals and industry.

In the United States the program for applications of radioisotopes began about 1946, and the overall growth has levelled off to about 15 percent per year. Originally the AEC supplied essentially all of the domestic isotopes, but for a number of years it has been withdrawing from the sale of isotopes that can be produced in commercial facilities. As a result, except for the fission products, the AEC produces only a small proportion of the isotopes used in the United States.

About 100 private U.S. firms produce radioisotopes or convert them into products for medicine, science, and industry. Total sales of these companies are estimated at $53,000,000 annually, consisting of about $8,000,000 in basic radioisotope materials, $16,000,000 in radiochemicals, $25,000,000 in radiopharmaceuticals, and $4.000.000 in radiation sources. In addition, sales of devices in which radioisotopes are employed total about $40,000,000 a year. If the sales of products produced by radiation processing, auxiliary materials, and services related to radioisotopes and radiation uses are included, the total commercial activity in the United States is at a level of several hundred million dollars annually. The cyclotron isotopes account for about $3,000,000 in sales each year. At the end of 1969, 57 Industrial accelerator irradiation units and 18 industrial isotope irradiation units were in operation. Most of the isotope users must be licensed.

Table 3: Global Savings from Radioisotope Use
($000,000)

applications	24 countries, 1961–63	U.S., 1963	U.S.S.R., 1961	total
Gauging	26.7–43.4	35.2–50.4	100*	162–194
Radiography	12.1–28.9	4.0–7.6	22	38–58
Ionization	1–2	†	†	1–2
Tracing	10–40	27-48:	58*	95–146
Total	49–104	66–106	180	296–400

*The exact distribution of savings in certain gauging and tracing is not known. †Included in other groups.
‡Includes also certain gauging and ionization applications.

Table 3 lists savings experienced by some countries through the use of radioisotopes in various applications.

BIBLIOGRAPHY. A. ROMER (comp.), *Radiochemistry and the Discovery of Isotopes* (1970), a collection of original articles of historical interest: F.E. MCKINNEY *et al., Special Sources of Information on Isotopes* (1968), an extensive bibliography on this subject; *Isotope User's Guide* (1969); YEN WANG, *Handbook of Radioactive Nuclides* (1970); C.M. LEDERER and V.S. SHIRLEY, *Table of Isotopes,* 7th ed. (1978), a listing of known isotopes and their properties. See also publications in the "Proceedings Series" and the "Panel Proceedings Series" of the INTERNATIONAL ATOMIC ENERGY AGENCY (IAEA).

Specific Applications: J.H. LAWRENCE *et al., Radioisotopes and Radiation: Recent Advances in Medicine, Agriculture, and Industry* (1969); E. QUIMBY *et al., Radioactive Nuclides in Medicine and Biology,* 3rd ed. (1968, reissued 1981); M.F. L'ANNUNZIATA, *Radiotracers in Agriculrural Chemistry* (1979); G.M. URROWS, *Food Preservarion by Irradiation* (1964); W.R. CORLISS and R.L. MEAD, *Power from Radioisotopes,* rev. ed. (1971), a discussion of the SNAP program, radioisotopes, and generators on Earth; E.W. PHELAN, *Radioisotopes in Medicine* (1966), a discussion of specific radioisotopes as diagnostic and therapeutic tools: R.E. JERVIS, "Present Status of Activation-Analysis Applications in Criminalistics," *Isotopes and Radiation Technology,* 6:57–70 (1968); H.J.M. BOWEN, *Chemical Applications of Radioisotopes* (1969); G. FAURE, *Principles of Isotope Geology* (1977), an introduction to age dating and other geologic applications.

(P.S.B.)

Radiology

Radiology is a branch of medicine that originally dealt with the use of X-rays in the diagnosis of disease and the use of X-rays, gamma rays, and other forms of ionizing radiation (*e.g.*, radiation that causes formation of positive and negative ions in the substance penetrated) in the treatment of disease. In more recent years radiology has also come to include diagnosis by a method of organ scanning with the use of radioactive isotopes and also with nonionizing radiation, such as ultrasound waves. Similarly, the scope of radiotherapy has extended to include, in the treatment of cancer, such agents as hormones and chemotherapeutic drugs.

The topics discussed in this article include the profession of radiologist and the ancillary professions, types of radiation and the radiation sources used, radioactive isotopes, other forms of radiation used in diagnostic and therapeutic radiology, new developments, and needed future developments.

History and associated specialties

DEVELOPMENT OF RADIOLOGY

X-rays were discovered by Wilhelm Conrad Rontgen, a German professor of physics, in his laboratory in the University of Wurzburg on November 8, 1895. It is probable that other workers had produced X-rays previously in their laboratories without realizing their significance, but Rontgen immediately set about a complete analysis of the properties of this phenomenon. Seven weeks later, on January 1, 1896, he published a model exposition of his work, to which little in basic knowledge has been added even to the present day, although multitudinous fringe advancements of knowledge have occurred, and developments in the techniques of the use and applications of X-rays continue to the present time. Within days of his publication there was speculation in newspapers in Germany, Austria, the United States, and England as to the possible medical applications of this new discovery, and so, one of the greatest revolutions in the evolution of modern medicine was started.

In radiodiagnosis use was made early of three of the properties of X-rays—their ability to penetrate the tissues, their photographic effect, and their ability to cause certain substances to fluoresce. In penetrating the tissues, the radiation is absorbed differentially, depending on the densities of the tissues that are being penetrated. The radiation

Introduction of contrast media

emerging from tissues thus produces on a photographic film or a fluorescent screen an image of the structures of differing densities within the body. The limiting factor is the similarity between the densities of adjacent soft tissues within the body, with a resultant failure to produce a contrast between the images of adjacent structures or organs. During the first two decades following their discovery, X-rays were used largely for the diagnosis and control of treatment of fractures and for the localization of foreign bodies, such as bullets, during World War I. Early, the physicians using these methods introduced artificial contrast agents, such as a paste consisting of barium sulfate, which is inert and nontoxic when taken by mouth. When it is taken by mouth or introduced by enema, the various parts of the alimentary tract can be demonstrated and examined. Refinements of this technique continue to the present day, and radiological examination of the alimentary tract is an elegant and precise aid to diagnosis. Later, other contrast media were produced that could be injected into blood vessels. The media could thus be used either to demonstrate those vessels, whether arteries or veins, or after their selective concentration and excretion by the kidneys, to show the urinary tract. Within the first few months after Rontgen's discovery, attempts were made to produce films of moving objects; thus, it was soon realized that radiology might be able to depict function and so demonstrate dynamic physiological functions rather than just static anatomy. Technical difficulties and the hazards of a high dose of radiation to the patient prevented the proper development of this technique. In the 1950s, an electronic method was devised to intensify the image, the so-called image intensifier, which made possible the overcoming of the technical difficulties, and now cineradiography is routinely performed in many departments. During the whole period of development, photographic techniques were also continually improved. Single-coated photographic plates were used at first, and then double-coated photographic films; photographic emulsions have now been developed to such a point that high speed can be provided with good definition and little intrusion of photographic grain into the image. Similarly, processing methods have been improved; automatic processors now can deliver a fully processed dry film in 90 seconds.

Emergence of the radiobiologist

In radiotherapy, use is made of the biological effects of ionizing radiations. The early workers noted that large doses of radiation would cause, after some delay, reddening of the skin, which might lead on to blistering and ulceration. Even small repeated doses, if occurring often enough, might produce serious skin lesions. It was argued, then, that a phenomenon producing such damage to normal tissues might be directed toward abnormal and undesirable tissues, such as cancer. Research into the fundamental nature of the biological action of radiation continues to the present day, and a new race of scientist, the radiobiologist, has emerged. About the same time as the uses of X-rays were first being applied to medicine, radium was discovered, and also the importance of the time factor as a modifier of the reaction of tissue to radiation was established. Thus was born the art of radiotherapy, at first based entirely on an empirical approach. It was soon noted that ionizing radiations also have the effect of alleviating pain, and so in the period of development of this form of treatment it was used rather extensively in the treatment of painful forms of arthritis, swellings of the salivary glands, herpes zoster or shingles, overgrowth of adenoids in children, and several other benign conditions. As knowledge of the possible harmful effects of radiation has grown, many of these applications have been discarded, except in special circumstances and under strict supervision.

The vast bulk of the practice of radiotherapy has to do with cancer, and it is here that the great advances have been made. The first step was to establish a unit of measurement. Since the early days physicians practicing treatment with ionizing radiations have worked in close collaboration with physicists, and much of the fundamental research has been undertaken by radiation physicists working alongside their medical colleagues. The establishment of an internationally accepted unit of measurement at Stockholm in 1925, the roentgen unit, enabled physicists to undertake similar calibration in different centres and provided a means of devising a system of dosimetry.

THE PROFESSIONS ASSOCIATED WITH RADIOLOGY

The radiologist and radiographer

A radiologist is a physician who has specialized in radiodiagnosis or radiotherapy, or both. During the evolution of radiology most physicians in this branch of medicine practiced both forms of the specialty, and in certain parts of the United States, in France, Germany, and some other parts of southern Europe, in South America, and in India, it is still common to practice both. Since the end of World War II, however, in the United Kingdom, the Scandinavian countries, The Netherlands, Australia and New Zealand, and the major cities in the United States, the two have been completely separated, with separate training programs. This is a natural development as equipment and techniques have become more sophisticated and as each branch has tended to absorb more fringe activities, such as the use of ultrasound techniques in diagnostic radiology and the use of chemotherapeutic agents in the treatment of cancer in therapeutic radiology.

In most Western countries now, for a physician who wishes to specialize in radiology, after qualifying as a physician, the training period is four years. The majority of radiologists practice in hospitals. Equipment is now so exotic and the facilities required so demanding that it is usually possible to produce the financial requirements and the supporting staff only within an institution. It is usually also only in large institutions with private or public financial backing that significant research can be undertaken and advances made. Nevertheless, in all countries throughout the world, there is always a demand for private facilities, and these exist at a high level of work for all but the more sophisticated techniques in most countries. The radiologist is the medical practitioner probably most often consulted by other medical practitioners. In radiodiagnosis he has a clinical problem referred to him; he investigates it either by a simple examination, such as for a fracture, or by a more complex technique, such as the introduction of a catheter (tube) into a renal (kidney) artery and showing the renal circulation by injecting a contrast medium; he then interprets the results to the referring physician. The surgeon, the orthopedist, the internist, the gynecologist, the pediatrician, and the general practitioner all refer problems to him. In radiotherapy he again is a man with problems referred to him—*e.g.*, the patient with cancer for advice on treatment. This treatment he takes over and conducts and continues to review in collaboration with his colleagues. In the Western world, with the increasing expectation of life, and therefore the increasing incidence of cancer, he is becoming more and more a key figure in the satisfactory treatment of a major human problem.

Radioactive isotopes are unstable forms of elements that emit radiation in the process of breaking down into other elements. With the growth of the use of these isotopes for diagnosis and therapy, a new branch of medicine is developing—that of the specialist in nuclear medicine. Some of the investigations carried out by such specialists lie outside the normal scope of radiodiagnosis; many of them are more related to physiology and clinical biochemistry. Thus, nuclear medicine exists with part of its frontier overlapping radiology; many radiologists practice nuclear medicine, and also many of those who practice nuclear medicine are drawn from such other disciplines as endocrinology, gynecology, and biochemistry.

The radiologist is supported by a radiographer or radiological technician. These highly trained technicians in most Western countries undergo a two-year period of training and carry out most of the day-to-day conduct of patient investigation or treatment under the guidance or supervision of the radiologist. In many less developed countries, where the need for radiology as an aid to diagnosis is far greater than the ability of the government to provide radiologists, it is essential that technicians be trained and sent to outlying places in order to carry out the technical aspects of such examinations.

In places in which radiotherapy exists on any substantial scale, physicists are needed and in most instances are provided. Medical physics has developed or grown alongside the growth of radiotherapy. Now that medical physicists exist, they provide a service to a greater clientele than radiotherapists; with the increasing need for an awareness of radiation hazards and the constant monitoring of all radiation workers including radiographers, the diagnostic radiologists also receive a service from them.

Radiation: types, sources, and hazards

TYPES AND SOURCES OF RADIATION

X-rays. X-rays form part of the broad spectrum of electromagnetic waves of which radio waves, microwaves, infrared rays, visible light, ultraviolet rays, and cosmic rays are also component parts. These waves are energy waves, the various parts of the electromagnetic spectrum differing in wavelength; whereas the wavelength of radio waves is measured in metres, that of X-rays is measured in angstrom units. (The vast difference in size thus indicated is difficult to grasp: there are 10,000,000,000 angstrom units in a metre.) X-rays have various physical and biological properties that are made use of in both diagnostic and therapeutic radiology. In diagnostic radiology, use is made of their penetrating property, of the fluorescent effect that they have on certain substances, and of their photographic effect. In therapeutic radiology, use is made of their penetrating property and their biological effect.

X-ray production

Production of X-rays is brought about when electrons travelling at high speed collide with matter in any form. The usual method of producing them is by means of an X-ray tube, an electronic device consisting of an evacuated glass tube containing two electrodes to which is applied an alternating current of many thousands of volts. The high voltage is produced by an X-ray generator.

X-ray generators are in fact step-up transformers. The normal domestic or industrial electric supply of alternating current is fed into the transformer and stepped up to the required voltage—in the diagnostic field ranging from 25 to 150 kilovolts or occasionally 200 kilovolts, and in the field of radiotherapy ranging from 10 to 300 kilovolts. Obviously, at the differing kilovoltage levels, the transformers differ considerably in complexity, size, weight, and manoeuvrability. At the lower kilovoltages in some machines, the X-ray tube itself, being a valve, acts as a rectifier (a device for changing alternating to direct current), no current passing from cathode to anode (negative to positive electrode) during half the cycle, the other half being utilized—a process called half-cycle rectification. In the more modem type of machine, full-wave rectification is achieved by inserting four electronic or solid-state rectifiers in the generator between the high-voltage terminals of the transformer and the terminals of the X-ray tube. In the extremely sophisticated machines, a means of producing what amounts to straight-line current across the tube has been achieved by the use of six such rectifiers. At this point the needs and design of sources and other factors for diagnostic radiology and radiotherapy differ.

Description of X-ray tube

X-ray tubes consist of an evacuated diode (electronic tube containing a cathode and an anode) surrounded by a metal casing that is lined with lead except for a small port or window through which the useful beam emerges. The space between the casing and the glass insert is filled with insulating oil, which also has the property of being a good conductor of heat. The cathode houses a filament within a focussing cup, which is supplied with current from a low-voltage transformer within the generator. The filament can thus be heated to incandescence, providing a halo of electrons. The anode consists of copper, a good conductor of heat, with a small block of tungsten (which has a high melting point), called the target, about one-half inch square, set in the anode face. When an extremely high current is applied to the cathode and the anode, the available electrons are attracted to the anode along a path in the vacuum, striking the target with tremendous energy. This impact creates two products—heat and X-rays. The face of the anode is angled in such a way that the maximum beam of X-rays is directed through the port in the tube casing. Only about 1 percent of the energy directed at the anode results in the formation of X-rays, the remainder being wasted by heating of the target. Tube manufacturers have devised various means of preventing the melting of the target—*e.g.*, uses of a metal of high melting point and an anode that is a good heat conductor; oil circulation around the tube insert and anode terminal; and rotation of the anode during exposure so as to expose a changing surface of the target to the stream of electrons.

The higher the current in the cathode filament, the more electrons there are available to join the stream to the anode, and so the greater is the intensity of the X-ray beam. The higher the potential difference (*i.e.*, kilovoltage) between the cathode and the anode, the greater is the speed of the electrons striking the target, which results in producing X-rays of a shorter wavelength and therefore of greater penetrating power. The controls for varying filament current and for varying kilovoltage are within the generator and are manipulated from the control table on the X-ray unit. One further factor influences the quantity of X-radiation produced, the time during which the current, measured in milliamperes, passes across the tube. Intensity of the X-ray beam, therefore, is expressed as the product of tube current and time—*i.e.*, as milliampere seconds.

Radiographic quality is the term used to describe the photographic image produced on photographic film and is dependent on a number of factors. Photographic factors affecting quality are described in a later section, but some physical factors also have an affect and are properly described here. The shorter the wavelength of the radiation used, the greater is the penetrating power. The interpretation of radiographic images depends on producing shadows of adjacent structures. X-radiation is absorbed differentially by tissues of different densities, the resulting photographic image portraying this differential absorption. It is necessary to use X-rays of greater penetrating power to produce satisfactory radiographs (X-ray pictures) of the skull, for example, than of the soft tissues of the leg. Similarly, the photographic blackening of a film depends on the intensity of the beam, which must be strictly controlled.

The X-ray beam emerging from a small target on the X-ray tube acts like a point source of light and produces a sharp image, whereas the beam from a broad target produces a penumbra effect (*i.e.*, a grayish margin). Because a small target cannot be subjected to so intense a bombardment of electrons, however, a compromise must be achieved.

X-radiation is absorbed in air as the square of the distance. Thus, the greater the distance of the object and the film from the source, the less the intensity of the beam will be; the nearer to the source the object and film are placed, the greater will be the geometrical distortion of the image. Thus, optimum conditions for radiographic procedures are essentially a compromise.

Use of grids to reduce X-ray scatter

One further feature is involved. If a photon (an amount of energy) of X-radiation strikes an atom and dislodges an electron and is of greater energy than the energy required to dislodge the electron, the residual energy continues as a photon of scattered radiation in a different direction. The greater the intensity of an X-ray beam and the thicker or denser the part being penetrated, the greater is the scattered radiation. Scattered radiation may produce fogging and blurring of the image. Therefore, instruments known as grids are interposed between the patient and the film. Grids consist of an assembly of lead strips interspaced with strips of wood or plastic or other radiolucent material (material partially penetrable to radiation). The relation of the depth to the width is called the grid ratio; the greater the ratio the greater is the efficiency to absorb scattered radiation. Grids may be stationary or moving, the latter being synchronized with the X-ray exposure, the movement being either unidirectional or oscillating.

X-ray tubes for therapy machines are, in principle, mostly the same as for diagnostic radiology, differing only in detail. Target size is not so critical as for diagnostic tubes so long as excessive penumbra is avoided. Cooling of the anode may be achieved by a variety of means; for example, having the copper anode in metallic connection with a number of metallic fins or sheets outside the tube, which can be cooled by a flow of oil or other suitable material past them. For low-voltage X-ray equipment, the target must be placed close to the skin, and the window must be as thin as possible, so that excessive filtration of the beam is avoided. In supervoltage machines a transmission target is used in which the useful X-ray beam is taken from the target in the same direction as the incident electron stream, so that the emergent beam is filtered by the target material itself.

Dosimetry is the science of the measurement and calculation of dose distribution in the tissues and of depth of dose. The units of dose have been redefined from time to time.

Three types of units for measuring ionizing radiation are in common use: the roentgen, which is the unit for absorption in air; the rad (radiation absorbed dose), which is the unit of absorbed dose; and the rem (roentgen equivalent man), the quantity of any ionizing radiation that has the same effect as one rad of X-rays (of 200–250 kilovolts).

The roentgen is the quantity of X-radiation or gamma radiation that produces in 0.001293 gram of air (one cubic centimetre of dry atmospheric air at 0° C [32° F] and at a pressure of 760 millimetres of mercury—the standard atmospheric pressure at sea level) one electrostatic unit of positive or negative electricity. The rad is the amount of radiation that delivers energy equivalent to 100 ergs to a gram of irradiated tissue.

Several instruments have now been devised for dose measurement. So important is the accurate calculation and planning of dose delivery that the existence of sophisticated physics support—with physicists, physics technicians, and physics workshops—is now regarded as an essential for any centre at which radiotherapy is practiced. The applications of this in clinical practice are considered in more detail in a later section.

Gamma rays and X-rays compared

Gamma rays. Gamma radiation, the wave radiation that arises in the nucleus of a radioactive atom, may occur naturally, as from uranium and radium, or may be produced artificially, as from the radioactive isotopes cobalt-60 and iodine-125. (A radioactive isotope is an unstable variant of a substance that has a stable form; during the process of breaking down, the unstable form emits radiation.) Gamma rays and X-rays have similar characteristics, and differ only in their wave length and their origin. The energies of the gamma-ray quanta given off by a particular radioactive element are characteristic of the nucleus from which they come. Gamma rays from the isotope cobalt-60, for example, are confined to two wavelengths, 10.5 X.U. and 9.3 X.U. (an X.U., or X unit, is approximately 0.00000000001 centimetre). Similarly, the spectrum from radium is complicated but still consists of a number of discrete wavelengths. By comparison, an X-ray beam as generated in a conventional X-ray tube is heterogeneous, consisting of radiation of many wavelengths. The penetration of gamma rays is measured by the half-value layer—the thickness of suitable material that is required to reduce the intensity of the radiation by one-half. Ten millimetres (0.4 inch) of lead or ten centimetres (four inches) of tissue is needed, for example to reduce the radiation of isotope lead-214 by one-half; only two centimetres (0.8 inch) of tissue is required for isotope iodine-125, which has low penetrating power.

Radioactive isotopes. Radioactive isotopes have already been considered in relation to gamma radiation; in later sections their clinical uses in both diagnostic and therapeutic fields are discussed. Here mention need only be made of isotopes as a source of radiation. Radioactivity is the property, possessed by some isotopes, of emitting ionizing radiations spontaneously. These radiations are emitted from the nucleus of atoms when the nucleus is unstable because of its structure. Radioactive isotopes are indistinguishable chemically from the stable form of the element. The radiations emitted consist usually of charged particles, nearly always accompanied by gamma radiation. These radiations, in addition to gamma rays, may consist of: (1) alpha particles, which are positively charged helium atoms travelling with high energy and high velocity; they have little penetrating power but can cause local biological damage. (2) Beta particles, which are negatively charged electrons; they have high energy and high velocity and are relatively easily absorbed. (3) Positrons, which are the same as beta particles except that they carry a positive charge; the range of a positron is small, and on coming to rest it combines with an electron, producing a pair of gamma photons. The radioactive isotopes used in clinical practice are mostly gamma emitters, though some beta emitters are used in both branches of radiology.

Ultrasound waves. Ultrasound waves, produced by a device called a transducer and of about 1,500,000 to 5,000,000 hertz, or cycles per second, are able to pass through body tissues in straight lines until they encounter an interface, or common boundary, between two types of tissue. At this point the waves are reflected as echoes. The echoes, received back at the transducer, can be converted into electrical impulses and shown as waves on the fluorescent screen of an oscilloscope. Two types of scanning with ultrasound waves, known as the **A** scan and the compound B scan, are used to determine the depths of the reflecting interfaces within the body. These diagnostic techniques, analogous to the use of echo sounders to chart the ocean floor, are mentioned further in a later section.

Infrared rays. Thermography, the use of infrared rays diagnostically, is a relatively new diagnostic measure. It represents a spin-off from military advances and rocketry fields. A thermograph is a graphic record of the infrared radiation emitted from the skin. The record is obtained by scanning the skin's surface, and it indirectly relates to cutaneous blood supply. The method is simple, harmless, and relatively inexpensive.

The infrared radiations are concentrated and focussed by a system of mirrors, the area of skin being scanned somewhat as a motion picture camera records the changing events in a scene. The beam of heat radiation is then focussed on a highly sensitive crystal, usually composed of indium and arsenic salts, which is cooled on one surface by liquid nitrogen to below —160" C (—256" F) to increase its sensitivity. The crystal converts the heat energy into units of electrical energy, and the latter can then be displayed on a television monitor. Permanent records of the display can be obtained with Polaroid films. Various modifications can be employed. Thus, isotherms can be drawn, lines connecting points within a given temperature range. A variety of instruments is available. One widely employed unit can record thermographs in the range 0"–1,000" C (32"–1,800" F) when used commercially in blast furnace control, but when adapted to clinical use in accurate to 0.2" C (0.4" F).

RADIATION HAZARDS

Mention has already been made of the dangers of exposure to ionizing radiation. These dangers may be immediate and in the form of a burn, just as a burn may be caused by other forms of electromagnetic radiation, such as visible light or ultraviolet rays, except that in the case of ionizing radiations the bum **may** affect deeper tissues. In radiodiagnosis with modem apparatus, the doses administered, even under prolonged screening procedures, are so small as to be far below the range at which such an effect is likely or possible. In radiotherapy with modem apparatus, the control of beam direction, of field of application, and of depth within the tissues of dosage also practically eliminates this possibility.

Types of remote danger

Remote dangers may be either (1) malignant; *i.e.*, leading to a localized cancer, such as a skin cancer, or to a generalized cancer, such as leukemia; or (2) genetic; the genetic effect results from a genetic mutation, of which the individual is unaware but which may affect future generations, as the affected individuals continue to reproduce.

With regard to malignancy, the doses administered in radiodiagnostic investigations are regarded as being safe, though it is thought that in the early period of the development of the technique of cardiac catheterization (introduction of a tube into a vein and along the vein into the heart) before apparatus had become as refined as it is today and when some of these procedures were not in the hands of radiologists or under their supervision, large bone-marrow doses may have been administered that could subsequently have led to leukemia. It is virtually impossible to prove such a course of events. Similarly, in radiotherapy, dose and beam control are now so highly developed that the possibility of causing cancer is highly unlikely. It is a fact, however, that in the 1930s wide-field radiotherapy was a standard practice in treatment of ankylosing spondylitis, a form of arthritis of the spine occurring often in young men. Follow-up 20 years later of a large group of persons thus treated showed a higher incidence of leukemia than in the ordinary population.

With respect to genetic effect, the doses administered in radiodiagnostic procedures could be significant, especially those involving the lower part of the trunk and the upper part of the thighs. All radiologists and all radiographers in their training receive instruction and guidance about the importance of filtration of the beam, limitation of beam size, sex-gland protection, especially in young persons, and departmental discipline. It is considered that no unnecessary X-ray examination should ever be carried out, and that the physician requesting an investigation should always consider that the medical needs and advantage likely to accrue from the examination are clear and outweigh the admittedly small potential danger. Provided that protective devices are available, that the technicians are protection-conscious, that departmental discipline is good, and that the requesting physician is circumspect, dangers are unlikely to arise. In radiotherapy, dose and beam control render genetic effects unlikely except in certain unavoidable cases. Similarly, the age group and type of condition being predominantly treated often render the likelihood of further procreation small.

Similar effects do not occur with nonionizing radiations.

The practice of radiology: techniques and equipment

DIAGNOSTIC RADIOLOGY

In modem medicine one of the basic essential aids to diagnosis is the science of radiodiagnosis, using X-rays, radioactive isotopes, or ultrasound as a means of demonstrating normal or abnormal anatomy and function.

X-ray techniques. ***Exposure and timing.*** When X-rays are generated by means of a low kilovoltage, the radiation has a long wavelength and a relatively low penetrating power. Radiations produced in the five- to 20-kilovolt range are used mostly for laboratory work. From 25 to 45 kilovolts, one of the few clinical applications is X-ray examination of the breast. Radiations of 50 to 100 kilovolts are used for many of the day-to-day routine examinations--of the skeleton, the chest, the alimentary tract, the urinary tract, and so forth. X-rays generated by means of a kilovoltage above 100 kilovolts have a short wavelength and high penetration. In the diagnostic range it is rare to use a kilovoltage either above or even as high as 200 kilovolts. The greatest field of application is in the range from 100 to 125 kilovolts, with some techniques that make use of up to 150 kilovolts.

Elimination of movement

Wavelength, and therefore penetration, is dependent on the generating kilovoltage, but this is not the only factor concerned in the production of a satisfactory radiograph. Intensity of radiation is a factor, this being dependent on the current passing between the cathode and the anode of the X-ray tube during an exposure and on the length of the exposure. In routine work it is essential to eliminate movement so as to obtain the greatest possible sharpness of image. This is achieved by stabilizing the part under examination, such as a limb, with sandbags, or by inducing the person being examined to hold the breath, and always by using the shortest time exposure possible. Not all movement can be stopped, however, and cardiac (heart) movement can cause blurring. In order to overcome this, timing devices have been designed to give exposures of as little as 1/120 of a second or even shorter, but this in turn has made it necessary to build generators that will provide a greater milliamperage. In conventional diagnostic work, most generators work in the 200 to 500 milliamperage range, but generators have been built that will deliver up to 1,000 milliamperes.

Films and intensifying screens. X-radiation produces a photographic effect on photographic film, but not as great an effect as visible light. X-radiation also causes certain substances, mainly barium platinocyanide, calcium tungstate, and zinc sulfide, to fluoresce, producing visible light. As it is always desirable to keep time exposures to a minimum and also to deliver the least amount of radiation to the subject under examination, use is made of both these features. Double-coated film, with a silver halide photographic emulsion on a cellulose acetate or polyester base, is used for the purpose, so as to achieve maximum film blackening from any exposure. Film cassettes are used that contain a pair of intensifying screens made of fluorescent material between which the film is pressed. Good contact between screens and film is essential to prevent blurring. The major effect on the film, therefore, is due to visible light. Because the grain size of the crystals of fluorescent material can subtract from the sharpness of the image, for examinations in which minute detail is essential (*e.g.*, of hands or feet), the photographic effect of X-rays rather than of visible light is used.

Rapid film processors

Processing of X-ray films. Processing of X-ray films is the same in principle as any photographic processing and is carried out in a darkroom. In many departments this is done manually; it consists of developing, rinsing, fixing, and washing, followed by drying. This must be carried out according to a strict regimen, temperature and time control being of the greatest importance. Development at 68° F (20" C) for five minutes produces a wet film for scrutiny after complete washing in 30 or more minutes. In recent years automatic processors have become available, which, still using standard film, produce a dry film in 11 minutes or as little as seven minutes. These processors have now been introduced into most major departments. More recently, rapid processors have been designed that produce a dry film in 90 seconds. This has been brought about by raising the temperature of development to somewhere in the region of 100" F (about 40° C), by introducing a hardener into the gelatin of the emulsion (so as to prevent melting), and by modifying the sensitivity of the silver halide in the emulsion. Some increased photographic graininess is considered in many busy departments to be more than compensated for by the speed with which dry films can be produced.

Some special procedures. Tomography, also called body-section roentgenography, is a procedure for demonstrating selected planes within the subject under examination. A conventional radiograph is a two-dimensional representation of a three-dimensional subject, all structures lying in the direction of the X-ray beam being superimposed. In tomography a plane in the line of the beam is selected, and the shadows of all structures above and below that plane are blurred. Thus, the shadows of unwanted structures are blurred, leaving in focus those shadows in the selected plane. This is achieved by moving the source of the X-rays (*i.e.*, the tube) in one direction over the subject, and the film in the opposite direction under the subject during the exposure, the tube and film having a connecting link that rotates about the fulcrum at the level of the selected plane. The movement of tube and film are thus at all times opposite and with velocities proportional to one another. This technique, devised in the early 1930s, now has a wide field of application. The apparatus can be simple, but sophisticated apparatus is also now available for this purpose.

Stereoscopy is a method of taking two films of a selected area, the films and subject remaining in identical spatial relationship but the tube being shifted slightly between the two exposures. The shift used is usually six centimetres (2.4 inches), this being the average distance between the pupils of an observer's eyes. The right shift is then viewed with the observer's right eye, and the left with the left eye, special binoculars being available for this purpose; the resultant image is a three-dimensional picture.

Macroradiography is an enlargement technique that can be performed with X-ray tubes having very small targets. If the effective target of the tube is 0.3 millimetre (0.01 inch), this will act as a point source producing little or no penumbra. In conventional radiography the part under examination is placed as close as possible to the film so as to preserve sharpness of the image. With a 0.3-millimetre target the subject can be placed at a distance from the film and enlargements made as great as to twice the original size that show a degree of sharpness comparable to that achieved by conventional procedures. In certain fields of work, scrutiny of such macroradiographs enables interpretation and demonstration to be made more satisfactorily.

Microradiography is a laboratory technique for examining thin sections of specimens. By means of a generator that produces X-rays in the five-kilovolt to 20-kilovolt range—thus of low penetration—images of extremely thin sections can be obtained. Some such machines with a molybdenum target and a beryllium window can produce an almost homogeneous beam of X-rays. Film having a single coat of extremely fine grain or maximum resolution plates are used; the exposed films or plates can subsequently be examined under the microscope by magnification up to X60 without interference from photographic grain.

Contrast media for special procedures. The procedures in which contrast media are used are described in some detail in the section on clinical applications. Contrast is defined as the difference in density between adjacent objects or structures or the difference in light transmission by the shadows of these structures or objects. The whole of radiological interpretation is dependent on the quality of contrast and the quality of sharpness. Into some structures, which do not create a shadow differing in density from the adjacent structures, artificial contrast media are introduced in order to make detailed examination of the structures feasible. Three main groups of contrast media are employed: gas (air, oxygen, or carbon dioxide), barium sulfate, and organic iodine compounds. Because gases absorb less radiation than the surrounding tissues, shadows with increased blackening result. Barium and iodine atoms absorb more radiation than the surrounding tissues, thus causing denser shadows. Gases are used for examination of the ventricular system of the brain (the brain cavities), for intra-abdominal problems, and in some examinations of the alimentary (digestive) tract. There are many and varied organic iodine compounds, with different chemical structures, different degrees of viscosity, and dealt with in different physiological ways by the organs or systems to which they are directed. By use of them the biliary tract, the urinary tract, renal (kidney) function, the vascular system, the chambers of the heart, the spinal canal, the female genital tract, and many other areas can be demonstrated. The essential features of contrast media are that they produce adequate contrast, are simple to handle, and are nontoxic.

Types of contrast media

Fluoroscopy. In fluoroscopy use is made of the fluorescent effect of X-rays. X-rays penetrating the subject under examination impinge on a fluorescent screen; the radiologist observes the image so created and is able to see motion and to move the subject in various directions so as to examine all sides and aspects of a particular structure. There are, however, radiation hazards to subject and operator; because of the increased time required to observe motion, the time factor is long, and tube current is kept low, thus providing only a dim image. To make maximum use of the dim image the operator must be fully dark-adapted. No permanent record of the screen image remains, and the screen image itself is too dim to photograph. Most machines, therefore, are fitted with devices for sliding a film between the subject and the screen and making a conventional X-ray exposure. This cannot be performed simultaneously with observation of the screen image, though many simple and many elaborate devices have been designed to cut down the time interval between the two events. This method was the standard procedure for examination of the alimentary tract for more than three decades.

Image intensifiers, photofluorography, and television. Electronic amplification of the screen image became a reality during the late 1950s and the 1960s. X-rays are used to produce an electron image of the subject. The electron image is then accelerated by a high-voltage field aided by electron lenses and mirrors, and the enhanced electron image is then converted to a visible image. This is the principle of the image intensifier. This image can be observed visually through a mirror system, it can be photographed (this process is called photofluorography), or a television camera can be fitted to an aperture on the intensifier outlet and the image observed on a monitor.

Photofluorography and television observation can take place simultaneously by means of separate apertures, and thus observation of what is being photographed is achieved. The television image can be stored on a videotape recorder and subsequently viewed repeatedly and at leisure. If the image is to be photographed, either a 100-millimetre camera or a 70-millimetre camera is used. The 100-millimetre cameras are commonly used to produce single-frame pictures (for example, of the chest); 70-millimetre cameras are used either for single-frame exposures or for exposures of up to six per second (as in alimentary tract work). By use of a 16- or 35-millimetre motion-picture camera, rapid serial sequences can be made; joint movements, micturition (urination) studies, and the swallowing mechanism are all fields of application, but the most common field is in the study of the heart, using a contrast medium (angiocardiography). The motion pictures are usually taken at 32 frames per second, but recent developments favour high-speed cineradiography (200 frames per second) in the investigation of some cardiac conditions.

Ward and operating-theatre radiography. Portable and mobile X-ray machines of considerable output have now been designed, some of conventional design, some battery-operated, and some using a condenser-discharge mechanism that can be used outside the X-ray department. There is a growing demand for X-ray work in the wards, with the development of intensive care units, and in the operating theatres, where radiographic control of surgical procedures is becoming more exacting.

Clinical applications of X-rays. *The skeletal system.* Among the first applications of X-rays for diagnostic purposes was visualization of bones. In fractures, the actual break in continuity of a bone can be portrayed, as well as its site, extent, and type, together with important features such as shortening or comminution (shattering into small fragments). The satisfactory reduction of the fracture and the progress of bony union can be demonstrated. In disease processes the radiological changes lag behind the clinical state so that the diagnosis usually cannot be made early—for example, in an acute infection such as osteomyelitis (inflammatory infection of bone). In a chronic disease, however, such as tuberculosis or chronic arthritis (inflammation of a joint), in which the onset is usually more gradual and unnoticed and the progress slower but more inexorable, changes in either the bone or the joint can often be demonstrated in minute detail. Normal bone consists of the cortex, which shows as a dense shell on the outside, and the medulla, which shows as fine honeycombed meshwork of bony plates, or trabeculae. Joint cartilage does not cast a shadow; thus a joint appears as a space between two bone ends. Changes in this space and in the underlying bone are an indication of joint disease, and on the character of these changes the type of joint disease can often be diagnosed. Thus, in addition to infections, long-standing conditions, such as rheumatoid arthritis and ankylosing spondylitis; degenerative disease, such as osteoarthritis and Paget's disease; deficiency diseases, such as scurvy and rickets; benign and malignant tumours; congenital malformations and hereditary congenital diseases, such as hemophilia, can all be demonstrated. Occasionally, in order to demonstrate the extent of a disease process or an injury within a joint, a contrast medium is injected into the joint cavity, and films are taken; this is the technique of arthrography.

Observation of fractures and of chronic disease

Usefulness of chest X-rays

The respiratory system. Viewing of the lungs and the other structures within the chest is the most widely used application of X-rays in clinical medicine. Air in the lungs absorbs little or no radiation, and so the blood vessels in the lungs appear on the radiograph as branching linear shadows, and abnormalities in the lungs appear as distinct shadows contrasted against the background of the air content of the surrounding area. Similarly, changes in the shape of mediastinal structures (the structures between the two lungs) appear as abnormal, projecting out against the air contrast of the adjacent lung. Thus, lung diseases (such as pneumonia, tuberculosis, pneumoconiosis, and cancer) can be diagnosed. In many Western countries, in the campaign that has been successfully waged against pulmonary tuberculosis, one of the two most important tools has been chest radiography, the other, of course, being the development of anti-tuberculous drugs. In the lungs, the air content is mostly in the alveolar spaces (the minute air sacs), and it may not be possible to detect changes in the respiratory passages—the bronchi and bronchioles. The technique of bronchography has therefore been devised: a radiopaque contrast medium is run into the bronchial tree, and minute changes in this structure are shown with great anatomical accuracy. Similarly, during a routine chest examination it may not always be possible to define the shape and limits of a lesion because of the superimposition of the shadows of overlying structures. Tomography, which eliminates these unwanted shadows, is widely used in the investigation of intrathoracic disease (disease of structures in the chest).

The cardiovascular system. Perhaps the greatest strides have been made in recent years in the investigation of the system of arteries and veins. The two contributory factors to this progress have been the development of highly sophisticated apparatus enabling the taking of high-speed motion pictures and the introduction of catheter techniques. By the introduction of a catheter, or tube, into a vein in the arm and threading it through until ıts tip reaches the heart chambers, or even until the tip has passed through the heart and into the pulmonary arteries, detailed investigation of the heart can be carried out. Pressures can be determined at different points along this course; blood can be withdrawn at various sites for analysis of oxygen saturation; finally, a contrast medium can be injected to demonstrate the size and shape of heart chambers and valves, and also any defects in chamber walls because of congenital lesions. Expertness in this procedure progressed rapidly after World War II, but the real revolution occurred in the 1950s when a catheter was developed with a shaped or curved end and with a central guide wire that could subsequently be withdrawn in order to allow the injection of contrast medium. Many variations on this have been developed; it is now possible to introduce a catheter into almost any artery or vein of reasonable size in the body by threading it from a puncture site in the leg or arm to the desired point for injection. Thus, it has become possible to examine the chambers in the left side of the heart as well as the right side, and the investigation of cardiac disease has become an extremely refined procedure. Similarly, it is possible to carry out coronary arteriography (X-raying of the coronary arteries), and, as a result, surgery of the coronary arteries is developing. It is possible to show the vessels of the liver or the alimentary tract, the renal circulation, or the pelvic vessels. In the pelvis, the placental circulation in a pregnant woman can be shown and thus the placental site can be localized in suspected cases of placenta praevia (abnormal location of placenta, so that it tends to precede the child during birth). The vessels of the limbs can be shown, and the whole field of vascular surgery, in addition to cardiac surgery, has grown up as a result. Another field that has recently developed in vascular work is the demonstration of the lymph vessels (lymphangiography). In this technique a contrast medium is injected into the .fine peripheral lymph vessels, and its progress through the meshwork of lymph vessels and glands is followed radiographically.

Use of catheter in investigation of blood vessels

The central nervous system. Investigations of the brain and spinal cord have reached an advanced stage; the subspecialty of neuroradiology is now recognized. The bony anatomy of the skull and spine can be demonstrated in considerable detail by conventional techniques. Skull radiography is a precise technique requiring accurate knowledge of the anatomical landmarks on the part of the radiographer. He must take immense care in the positioning of the patient, have angulation of the patient's head and of the X-ray tube precisely accurate, and use an X-ray exposure exactly suited to the investigation. Fractures of the skull and facial bones can be demonstrated, as can their relationship to other structures, such as the air sinuses. Infections in the paranasal and mastoid air sinuses (bony cavities alongside the nose and behind the ear) can be shown, as can destruction of bone in the skull by tumours. Changes due to long-standing increase in intracranial pressure may be seen. Changes of a similar nature may be detected in the spine.

By injection of a contrast medium into an artery in the neck--either the carotid artery or the vertebral artery—part of the cerebral circulation can he shown, and by injection in turn into the arteries on both sides of the neck the total intracranial circulation can be investigated. Such conditions as aneurysms (an aneurysm is a dilation and thinning of a vessel wall at a weak point), degenerative vascular changes, and brain tumours can be diagnosed. Another technique, known as encephalography, is to inject air into the subarachnoid space in the spine (the space between the middle and the inner covering of the spinal cord) and allow it to rise into the cerebral ventricles (the cavities in the brain), thus demonstrating their site and size; displacements due to atrophy or tumours can thus be shown.

Lesions in the spinal canal causing disease due to pressure on either the spinal cord or the nerve roots can be shown by the technique termed myelography; a radiopaque contrast medium is injected into the spinal canal and, under fluoroscopic control, is allowed to run up and down the canal while suitable radiographs are taken.

The alimentary tract. One of the first fields in which contrast media were used was that of examination of the alimentary tract. A barium sulfate mixture is used either by mouth to examine the esophagus (gullet), stomach, and small intestine, or by enema to examine the large intestine. Various refinements have been introduced, such as a combination of barium and air contrast to improve the sharpness of the depiction of anatomical details and so enable the radiologist to detect smaller and earlier evidence of disease. In the hands of an experienced radiologist, diagnosis of diseases of the alimentary tract is now a highly accurate procedure, though it is important to realize its limitations and the great need for close cooperation between physician and radiologist.

The biliary tract. For examination of the gallbladder and biliary tract a contrast medium can be given orally or introduced into a vein, or it can be injected directly into the biliary tract during or after a surgical operation. The contrast medium given orally is an organic iodine compound that is secreted by the liver along with bile. The gallbladder absorbs water from the bile containing the contrast medium; thus, the medium becomes concentrated and casts a radiographic shadow. Stones within the gallbladder can be shown by this method; the ability of the gallbladder to contract, a normal function, can be assessed by having the subject swallow a substance that normally causes such contraction. A clearer demonstration of the system of bile ducts can usually be obtained if an organic iodine compound is injected into a vein; this method is particularly useful if the gallbladder has been removed or has failed to concentrate substances given orally. Finally, when operating to remove the gallbladder, the surgeon may wish to make certain that there are no residual stones remaining in the common bile duct. This is accomplished by injecting a contrast medium into the system of bile ducts under direct vision, a technique called operative cholangiography.

Assessment of gall bladder function

The urinary tract. With the development of techniques for X-ray studies of the blood vessels, it is now possible to demonstrate the arterial, capillary, and venous phases of renal (kidney) blood circulation. Reduction in the toxicity of contrast media and advances in techniques have enabled radiologists to test with certainty the excretory efficiency of the kidneys, thus allowing earlier diagnosis of chronic renal disease and also showing the anatomy of the excretory and collecting system—the renal calyxes and pelvis (the cavities in which urine collects before passing by way of the ureter into the bladder), the ureter, and the bladder. For investigating the excretory efficiency of the kidney, a catheter technique is used, the contrast medium being injected directly into a renal artery. By this means, vascular abnormalities, such as aneurysms and infarcts (sections of dead tissue), can be shown, scarring due to chronic infection can be seen, and renal tumours diagnosed. To show the anatomy of the excretory and collecting system, an organic iodine compound that is selectively concentrated and excreted by the kidney is injected into a vein. By this means, chronic infection, obstructive lesions, stones, and tumours may be diagnosed, and indications suggestive of vascular lesions and some other conditions observed. The lower urinary tract, the urethra, and the bladder can be examined most effectively by the injection of a contrast medium directly into the outer opening of the urethra.

Other applications. Many other aspects of human anatomy and function are routinely examined by radiological means. The female genital tract can be investigated, especially in cases of infertility, by injection of a contrast medium, to show the cavity of the uterus and to investigate whether the Fallopian tubes are open. The salivary glands can be demonstrated by injection of contrast medium into their ducts. The lachrymal system can similarly be defined by injection into the tear duct. Foreign bodies in the eye or any other part of the body can be accurately localized. The mechanism of swallowing can be depicted and analyzed. .The female breast can be examined (the technique is called mammography), and in many cases lumps not perceptible by touch can be detected and diagnosed as cancer; in other instances, masses that can be felt can be diagnosed as cysts or nonmalignant lesions. With the growth of this branch of medicine, greater specialization has occurred. There are now individual departments dealing solely with pediatric radiology, urological radiology, skeletal radiology, and so on, and there are radiologists who devote their time and skill entirely to one or another of these fields.

Scanning techniques. *Isotope scanning.* In isotope scanning, radioactive isotopes are introduced into the body and their distribution is then determined by observations of their radioaction. By a process of comparison of the distributions so determined, it is often possible to recognize the presence, shape, and size of abnormalities. A radioactive material is administered, usually by way of the veins, and becomes concentrated to different amounts in different organs and systems. Different isotopes are used for different organs and are usually emitters of gamma rays only. Technetium-99, for example, is used in one compound (a pertechnate) for brain scans and in another (technetium sulfur colloid) for liver scans. The radiation emitted is detected by a scintillation counter, which is moved to and fro over the organ at the part of the body being scanned; these messages can then be electronically recorded as marks on paper or as a photographic image. Various scanning devices have now been elaborated, the most recent, the gamma camera, having the advantage of producing a picture four times more rapidly than conventional scanners and so being capable of screening a larger number of patients. Within the abdomen, scanning of the spleen can show the exact site and size of that organ. Liver scanning can distinguish localized lesions, such as cancer, from diffuse disease, such as cirrhosis. Similarly, the pancreas and tumours within it can be shown. The flow of blood through the kidney blood vessels can be demonstrated and abnormalities of function detected. Lung scans are useful in the diagnosis of pulmonary infarction. Brain scans may show localized lesions. The investigation of thyroid function by the use of iodine-131 is now routine.

Detecting and recording scintillations

Ultrasonic scanning. In ultrasonic scanning mechanical vibrations with a frequency of several million cycles per second—far above the range of human hearing—are used. These frequencies can be confined within a narrow beam. The diagnostic procedure is based on the reflection of ultrasonic waves at interfaces between different tissues. The echo amplitudes thus created can be detected by a sensitive receiver. Energy not reflected travels on and is available for reflection by deeper interfaces. Because ultrasound is absorbed rapidly by bone, application of the technique lies mainly at soft-tissue interfaces. The echoes collected by the receiver, as the transducer (the source of the sound) is moved across the surface of the subject, may be displayed on the screen of a cathode ray tube, or photographic records may be made.

In clinical diagnosis the technique is being used increasingly. In the brain it is used for localization of midline structures, especially after a head injury. Similarly, it can be used to determine whether a structure is solid or cystic or to localize foreign bodies in the eye. The progress of cirrhosis of the liver can be assessed in this way without the need for an operation. It is of value in obstetrical diagnosis, in diagnosing cysts and tumours in the pelvic cavity, cysts in the kidney or tumours in the bladder, and in assessing the function of heart valves in cardiac disease. It seems likely that the techniques of scanning organs by means of radioactive isotopes and ultrasound will evolve not as competitive but as complementary methods.

Thermographic scanning. Thermography has been widely used in clinical practice, although its precise status is still uncertain. Alterations in the heat of the forehead skin correlate with blood flow in the brain. The skin over and above the middle part of the eye is supplied almost exclusively by an artery that arises as a branch from the main artery to the brain within the cranium. Thus, reduced blood flow through the latter artery is reflected by reduced heat emission in this skin area.

A wide variety of inflammatory and cancerous conditions in bone, muscle, and other parts of the body can be detected by this means. Similarly, incompetent (malfunctioning) perforating veins in the calf of the leg (they are called perforating because they pass through the membrane connecting the two bones of the lower leg) can be recognized; warm blood from the deep veins can flow to superficial veins if the valves of the communicating channels are not efficient; this is useful information to the surgeon contemplating surgery in patients with varicose veins. Breast cancer can be detected and topographically mapped out by this method, and extensions of the tumour can be noted.

THERAPEUTIC RADIOLOGY

Radiotherapy is the treatment of disease by ionizing radiations; by far the most important group of conditions that respond to such treatment are cancers. As radiotherapists became expert in the treatment of malignant disease, it was a logical development for them to embrace other nonoperative techniques, such as hormone therapy and cytotoxic chemotherapy (attacking cancers with chemicals that destroy cells). Radiotherapy, however, is also used in certain skin diseases and in some other conditions.

Techniques. In the treatment of malignant disease both X-radiation and gamma radiation are used. It will be remembered that X-rays are produced when targets are bombarded by electrons, whereas gamma rays are ionizing radiations with similar properties that come from radioactive materials. In recent years equipment has been developed for producing X-rays of great penetrating power, at least equal to and often greater than that of gamma rays. The equipment designed to produce X-rays of great penetrating power does so by generating an intense beam of fast-moving electrons that strike a target; it can be arranged for the electrons to avoid striking the target and to escape from the accelerating chamber through a thin window. They can be used directly as a beam of therapeutic value. Similarly, some radioactive materials give off a spontaneous emission not only of

Types of ionizing radiation

gamma rays but also of alpha particles and beta particles. Alpha particles have little penetrating powei and are easily absorbed, though they have the property of producing intense ionization in theii path. The penetrating power of the beta particle is considerably greater and its ionizing effect less. Beta-radiation emitters have certain therapeutic values.

X-ray production. Grenz rays (X-rays of long wave and less ability to penetrate) are generated by use of a step-up transformer at voltages as low as ten kilovolts but rising occasionally to 40 kilovolts. With this type of radiation, the modern tube is fitted with a beryllium window that allows a broad beam of X-rays to pass through; the beam can be limited by applicators.

Superficial and medium therapy is achieved in the range of 50 to 140 kilovolts, usually by means of a step-up transformer. The amount is usually transformed into direct current by means of vacuum tubes, thus allowing a greater use of the current. The X-ray tube is usually oil-immersed to promote cooling; this cooling may be supplemented by the use of circulating water. The tube has an inherent filtration, absorbing some of the radiation, and additional filters can be used to control the quality of radiation emitted.

Deep therapy is produced usually with a voltage in the region of 240 to 300 kilovolts with a step-up transformer. With a current of 15 milliamperes, this type of unit can produce a dose of 90 roentgens per minute. The tube is, again, usually immersed in oil, with cooling assisted by water circulation. A variety of filters can be attached to the tube window to control the quality of radiation.

The linear accelerator is a supervoltage therapy machine now in common use. In this device, electrons are accelerated in a straight line by using extremely short radio waves of about ten centimetres' wavelength. The electrons are produced from a thermionic emitter (an apparatus that emits electrons and ions from white-hot materials) and are injected into the accelerator chamber in pulses, so that they are in step with that part of the radio wave in which the electric field will accelerate the electron along the chosen axis, and are directed onto a heavy gold target. By this means X-ray beams of energy from four to 15 million electron volts can be generated. In this type of apparatus it is possible to allow the electrons to escape by changing the solid gold target for a thin window, though at this level of megavoltage their therapeutic value is not great.

The betatron is another apparatus in current use in which electrons can be given enormous energy. In this the electrons are made to travel around and around in a circular path inside a suitably shaped vacuum chamber, and during the time they travel they are made to move with continually increasing energy; the energy either is emitted as an electron beam or is allowed to strike a target with an energy equal to 20 to 42 million electron volts. X-rays produced with this energy level have been found to have particularly useful properties in therapy.

Gamma-ray production. Mention has already been made of the fact that gamma rays may occur naturally in some substances, such as radium, or occur as part of the decay process of artificially produced radioactive isotopes. Several of these are used for therapeutic purposes.

Cesium and cobalt isotopes

Radioactive cesium units are available for gamma-ray therapy. The radioactive isotope of cesium—cesium-137—has a half-life of 33 years (that is, it takes 33 years for half of the cesium atoms to disintegrate). The physical properties of this form of radiation give it certain advantages over deep therapy, and the casing for housing can be relatively small, though protective walls must be much heavier than are required with X-ray units. Cesium-137 is present in substantial quantities in the fission products from uranium and is thus a waste product of chain-reacting piles.

Radioactive cobalt units are also now available. The development of nuclear reactors with high-neutron flux (high emission of neutron particles) has made possible the production of the very active isotope cobalt-60. In these units, the source is permanently mounted in the middle of a large sphere of lead that may weigh as much as a ton; consequently, the design of this type of apparatus produces engineering problems, such as the need to provide heavy concrete protective walls. This type of unit is the most usual in use for gamma radiation.

Radium is a naturally occurring element having chemical properties similar to barium and emitting ionizing radiation. It disintegrates with the emission of an alpha particle, forming another element, radon, in the process. Radium has an atomic number of 88 and an atomic weight of 226. The emitted alpha particle carries away from the nucleus a charge of two units and a mass of four units, so that the radon nucleus has a charge of 86 and a mass of 222. Radon, an inert gas similar to neon and argon, is important for therapy. Radon is also radioactive; it decays with a half-life of 3.82 days and gives off an alpha particle. The resultant atom is known as radium-A, itself an alpha-emitting body giving a product known as radium-B. Radium-B and the next product, radium-C, decay rapidly, giving off a beta particle. The series then continues through four more unstable radioactive products; finally, an atom of atomic number 82 and atomic weight 206 results. This is a normal stable lead atom; at this point the radioactive transformations cease. Such a series is known as a radioactive family. Radium itself is the product of a small series of disintegrations starting with uranium as the parent.

Radium sealed in a container so that the radon and subsequent products cannot escape reaches a stage of radioactive equilibrium and can be used for therapeutic purposes.

Radioactive isotopes of other elements also are used for therapeutic purposes. Radioactivity is a property that some isotopes of all elements possess. It can be produced artificially when slight modifications are made to the nuclear structure of the naturally existing isotopes. Strontium-90, for example, which occurs in the products arising from the fission of uranium, has a half-life of 28 years and has therapeutic uses.

Technical aids. It is important for a radiotherapist, who is to use a beam of ionizing radiation, to have adequate information about the distribution of dose that the beam will produce within the body.

Work of radiation therapist

Radiation physics is the science that has grown up in this context. The work of calibrating equipment, mapping the doses of radiation administered, and protecting patients and radiation workers from unwanted radiation is the work of graduate physicists. Modern treatment centres usually require a team of such physicists.

By the use of applicators the field to which a beam is applied can be defined and limited. By the use of filters the intensity and penetrating power of a beam can be varied, and so the depth of a dose assessed. Dose measurement must thus be carefully planned; this will often require a considerable number of measurements and tests before a particular treatment can be carried out. Similarly, routine checks must be carried out from time to time by the physicists after calibrations have been calculated.

This type of physics department requires elaborate laboratories and workshops. Equipment and instruments for the type of delicate measurement undertaken must be readily available. When carrying out measurements the physicist often uses models (which he calls phantoms) of wax or other substances of absorbing and scattering power similar to those of the part of the body to be irradiated. When detailed planning is proceeding it may be necessary to build molds and jigs in order to plan and deliver dose accurately.

Plaster shells may be made to fit on the part of the body under treatment so as to enhance accuracy of beam direction. All such devices are usually manufactured on the spot in associated workshops.

It is the responsibility of the radiotherapist to judge as closely as possible the location and size of the lesion, the amount and delivery rate of radiation likely to be effective, and the tolerance of neighbouring uninvolved tissues. When this has been done, the physicist determines the penetration, intensity, geometry, and aiming of the single or multiple beams best suited to deliver the radiation in the amounts and at the points stipulated. The technician or radiographer then administers the treatment under the supervision of the radiotherapist.

Clinical application. Growth is a fundamental property of living things, the reproduction of cells occurring by the process of division called mitosis. Growth does not cease entirely even in adults, for the normal replacement of tissues is by the reproduction of normal cells. Most such growth is strictly controlled in the interests of the body as a whole; however, some new growths are not thus controlled and are known as tumours, or neoplasms, and these may be benign or malignant. Malignant growths, known as cancers, spread either locally by invading adjacent tissues or to remote parts of the body through the blood or lymphatic channels. These remote lesions are known as secondary growths, or metastases. Malignant growths always inflict some degree of harm on the host, depending on a variety of factors, such as site of origin, type of tissue from which they arise, rate of growth, age of host, spread, and so forth. In the fight against cancer, radiation therapy, either alone or in conjunction with other forms of treatment, including surgery, is in the forefront.

The effect of radiation on tissues, whether normal or cancerous, is essentially the same, but since cancer cells usually grow more rapidly than normal cells, they are more seriously damaged by the beam. Radiation has the effect of arresting cell growth and division, and this effect can be permanently secured throughout a cancerous growth. Some growths and tissues are more sensitive than others to radiation; some are relatively resistant. Inevitably, adjacent normal tissues are also damaged. Some normal tissues, such as the blood-forming tissues and the reproductive organs, are sensitive to radiation. The training and experience of a radiotherapist enable him to assess all these different factors in planning and conducting the treatment and judging its results in the great variety of conditions that come his way.

A great many cancers can now be cured by means of radiation therapy, alone or in conjunction with other types of therapy. In some conditions, though cure may not be effected, arrest of the progression of the growth may be achieved, while in others, though cure or arrest is not achieved, radiation therapy has a palliative effect. Among the more common cancers treated are those of the breast, esophagus, bladder, brain, skin, uterus, larnyx (voice box), and tongue. Cancers of the blood-forming organs and bone also are treated in this way.

Use of oxygen in tumour therapy

The radiotherapist and his other clinical colleagues have a considerable choice of methods, often influenced by the stage that a cancer has reached by the time that they see it. In some, such as in cancer of the breast or some forms of hone cancer, radiation therapy may be used as an adjunct to surgical treatment. Tumours well supplied with blood vessels and well supplied with oxygen tend to respond more readily to radiation; hence in certain instances the oxygen concentration is artificially raised by having the patient inhale a high-oxygen concentration—about three times atmospheric pressure—from specially constructed chambers or tanks. Certain tumours can be treated with hormones with some effect, so that this is another method that the radiotherapist uses as experience dictates. Cancers of the breast, prostate, thyroid, and uterus may be influenced by hormone therapy. Similarly, some chemical agents, called cytotoxic drugs, are known to inhibit growth and to have an effect on tissue cells, especially cancer cells. Most cytotoxic drugs are administered intravenously; they may be used as an alternative to radiation therapy or in conjunction with it; occasionally they are used alternately with radiation therapy.

Radiation therapy is also useful in the treatment of a variety of skin diseases. Grenz rays have been found efficacious, the depth dose being very small. Similarly, some forms of rheumatism, some forms of inflammation, and some gland conditions can be safely and satisfactorily treated with medium or deep therapy.

Radioactive isotopes are also used as therapeutic agents in this field of work. Iodine-131, which is used in tracer doses in the investigation of thyroid disease, is successfully used in the treatment of thyrotoxicosis (the effects of an overactive thyroid). Similarly, phosphorus-32 and gold-198 have clinical uses in treatment. Strontium-90, which is a pure beta emitter, is successful in the treatment of some eye conditions, such as corneal ulcer.

NEW DEVELOPMENTS

Radiology is the most expensive branch of medicine in any institution, and the development stage in the design of any apparatus is always costly. The manufacturers of apparatus, therefore, must be convinced of the likely outcome of any design project before being prepared to tie up capital in its development.

Use of computers

The use of computers as a new tool in this field is still being explored, though used on an increasing scale. In therapeutic radiology, computer-assisted planning and dose calculation are already developing. This has necessitated, so that procedures and results from different centres can be compared, the building up of a computer language or terminology so that the data stored will be in similar terms. In diagnostic radiology, data storage and retrieval are already practiced in a few centres. In some actual radiological reporting can be made directly into an on-line computer and retrieved or displayed at a number of points in a hospital. A remote point, far from the reporting centre, can recall, and display reports whenever required. Similarly, the analysis of work flow in a department by use of a computer model can be carried out and contribute to efficient management. Both branches of radiology now are increasingly using computers for data analysis in large series and with a much larger number of components than was possible by previous methods. All these matters require careful planning and the establishment of a terminology.

In diagnostic radiology, the development of high-speed rotating anodes is just beginning to have effect. This permits higher tube loading and enables the practice of such procedures as high-speed cineradiography. The technique of xeroradiography, which employs charged selenium plates instead of X-ray films, is in the process of being rediscovered and may yet realize its potential both as a simple procedure and as a means of conserving the use of film in a world that is already concerned about the supply of silver.

In the developing countries, maintenance of apparatus, including the replacement of simple component parts, is often difficult and can take many months. The development of simple, hardy apparatus, easy to maintain, for use in such countries is a new development: the World Health Organization has taken an interest in this problem and has drawn up specifications for such apparatus; in the United States, a simple battery-operated machine has recently been designed for use in remote places where the electrical supply is unreliable. Further, in developing countries, the shortage is often not of doctors but of technicians to operate the machines. Efforts have been made recently to establish the basic requirements for entry into training and the minimum requirements in training such operators, so that in poorly served places the maximum use can be made of this essential diagnostic aid without exposing the population to unnecessary radiation hazards. In the field of education, in the programmed training of radiologists, audiovisual programs are being increasingly prepared and used, so that trainees, while working in a teaching environment, may have greater access to self-teaching facilities.

In therapeutic radiology, the main effort in new developments has been applied to the refinement of existing procedures and techniques. The delivery of dose and the regulation of beam direction have now been refined to such an extent that in many cases it is possible to extirpate the tumour with little or no damage to the surrounding tissues. Similarly, research is being done into increasing the sensitivity of the cancer by the use of oxygen under pressure.

Immunosuppressive effect of radiation

One last feature is of considerable interest. Radiation is known to have an immunosuppressive effect, as in reducing the rejection of tissue transplants. During the 1960s the surgery of organ transplantation advanced considerably, but rejection of the transplanted organ remained a problem. The immune response is closely interconnected with the function of lymphocytes in the blood and blood-forming tissues. Radiation is among the immunosuppressive agents that work by the destruction of the lymphocytes and the suppression of new lymphocyte production. The radiation may be administered either as total body radiation or locally. The latter may be either to the site prior to transplant or as external radiation of the transplant itself. Many techniques have been used, with varying degrees of success, in humans, and experimentally in animals, but much work remains to be done in this field.

NEEDED FUTURE DEVELOPMENTS

In diagnostic radiology, improvements in film speed without coarsening film-grain size are needed in order to reduce the patient's exposure to radiation.

Contrast media are needed that are less toxic and that coat the surfaces of hollow structures rather than filling them, since this would contribute to a more precise and detailed examination. Catheter techniques for the exploration of small vessels and for introduction of contrast media into them must be refined. The developing fields of isotope scanning and ultrasound scanning need further development.

In therapeutic radiology, it seems most likely that the application of the electron beam and of beams of fast neutrons will be explored in greater depth. Mention has been made earlier of the fact that electron beams, as produced in the betatron, have been found to have certain therapeutic values. This may open up a whole new area in this field of work. Similarly, work has recently been recommenced on the use of fast-neutron beams. Such a beam is produced in a cyclotron by bombarding a beryllium target. The neutrons thus produced have a wide spectrum of energy up to about 20 million electron volts, with an output of about 40 rads per minute. This type of beam appears to have a diminished effect on the skin, and at the present time it appears that there is a favourable tumour response. Much work remains to be done, and undoubtedly will be done, in this field.

BIBLIOGRAPHY. The historical description of the discovery of X-rays and the actual text (in English) of Rontgen's first papers are given in OTTO GLASSER, Dr. *W.C. Rontgen* (1945). For a detailed and comprehensible presentation of the physics of ionizing radiations and their applications, see W.J. MEREDITH and J.B. MASSEY, *Fundamental Physics of Radiology* (1968). The hazards involved in this field of work and a rational discipline are succinctly described in KATHARINE WILLIAMS, C.L. SMITH, and H.D. CLARKE, *Radiation and Health* (1962). The unbelievably vast scope of diagnostic radiology and its clinical applications are admirably introduced in DAVID SUTTON (ed.), *A Textbook of Radiology* (1969). Greater detail of the field of isotope scanning is given in L.M. FREEMAN and P.M. JOHNSON (eds.), *Clinical Scintillation Scanning* (1969); and of ultrasonic scanning in C.C. GROSSMAN *et al.* (eds.), *Diagnostic Ultrasound* (1966). *A* comprehensive, authoritative account of radiotherapy that is intelligible to the layman is J. WALTER and H. MILLER, *A Short Textbook of Radiotherapy for Technicians and Students,* 3rd ed. (1969).

(J.H.M.)

Radio Sources, Astronomical

Every object in the universe, with the theoretical exception of bodies frozen at absolute zero, emits radiant energy at least feebly. The visible stars emit electromagnetic waves in the optical range, and some emit detectable signals in other ranges that may include radio wavelengths. Radio waves, which may be produced by several processes, are sometimes strong enough to permit study of their emitting objects by radio, as well as optical, telescopes. The existence of some astronomical objects has been detected only by their radio emission. The first detection of radio signals from outside the Earth's atmosphere was made in 1932 by the American radio engineer Karl Jansky, whose discoveries led to the development of radio astronomy. During an investigation of the origin of atmospheric static signals, in addition to the static which originated in distant thunderstorms, he heard a background hiss on the loudspeaker attached to his receiver and antenna system. Significantly, this background hiss, or radio noise, reached a maximum intensity every 24 hours. More detailed study showed that the radio noise occurred whenever the Milky Way passed through the beam of the antenna, and Jansky concluded correctly that radio waves were being generated in the Milky Way and, further, that the most intense emission was coming from the centre of the Milky Way (the galactic centre).

Discovery of extra-terrestrial radio signals

Hearing of these results, Grote Reber, an Illinois engineer with an amateur interest in radio and astronomy, built a 31-foot- (9-metre-) diameter parabolic reflector antenna at his home, searched for the galactic emission at shorter wavelengths, and made the first radio map of the Milky Way; the resolution of his map—that is, the least angular distance apart at which two objects could be distinguished—was 12 degrees.

Although the Sun was the first astronomical object investigated (in 1900 by Sir Oliver Lodge in England) for radio emission, no solar radio signal was found until 1942, when radar systems situated around the English coast picked up an intense new form of interference that was, at first, thought to be enemy jamming but that on further investigation turned out to be radiation from the Sun. In the same year radio emissions from the Sun were detected in the United States.

PRINCIPLES OF RADIO RECEPTION

The range of radio frequencies used in Earth-based studies of the universe is limited only by the screening effect of the atmosphere. Normally, wavelengths from one centimetre to 30 metres can penetrate the atmosphere without undue absorption by the atoms and molecules present in it. The short-wavelength end of the spectrum has been extended down to about one millimetre by the use of high mountain sites, and the long-wavelength end has been successfully extended to 300 metres by observations from artificial Earth satellites.

Most of the radio energy falling on the Earth is in the form of a continuous spectrum—that is, radiation is received over the whole range of wavelengths with the intensity changing gradually across the whole range. Different types of radio sources have different spectra, and the shape of this continuous spectrum gives important clues about the physical conditions in a radio source.

Functions of radio telescopes

Types of receivers. Radio telescopes have two main functions. The first is to collect the weak radio signals, and the second is to separate details of the structure of radio sources. The angular degree to which such details can be distinguished from one another is called the angular resolution of the telescope. Commonly, parabolic reflectors, whose shape ensures that the signals are brought together into a good focus, are used to gather radio waves from a distant radio source together in a focal area to be fed into a radio receiver. The larger the reflector, the more sensitive it becomes-either it can detect more distant radio sources or it can survey the nearer sources more rapidly. The largest radio telescopes operating at metre or decimetre wavelengths have resolutions much worse than the unaided eye, which can normally distinguish between objects 0.1 millimetre (0.004 inch) apart at a distance of 25 centimetres (ten inches), a resolution of a little more than one minute of arc; an optical telescope does better still with a resolution of a fraction of a second of arc. In order to achieve the even better resolutions necessary for proper investigation of many classes of radio sources, the radio outputs from two or more telescopes are combined into what is called an interferometer.

Types of emission. ***Free–free emission.*** Radio emission, like light, is produced when a charged particle, generally an electron, is made to accelerate. One class of astronomical radio sources consists of clouds of hot ionized gas—that is, a gas whose atoms and molecules have absorbed enough energy to lose their electrons and become positively charged ions. Such a cloud emits radio waves by the process known as free–free emission. In this process, the free electrons are attracted toward positively charged ions as they pass each other; the acceleration produces a pulse of radiation. The sum of such pulses originating in a large number of such encounters between electrons and positive ions gives a continuous spectrum in which the power (energy per unit of time) radiated per unit frequency interval is constant with frequency. Radio sources that emit by the free–free process include the outer tenuous layers of the Sun and the ionized hydrogen regions of interstellar space. A similar process is that known as blackbody radiation, in which the emitting region is so compact or deep that the only emission that can escape comes from the near side. In this case, the power radiated increases as the square of the frequency. Measurements of the emitted radio power give a direct estimate of the temperature of the source. The planets belong to this class of source.

Synchrotron radiation. Among emission processes in radio astronomy, the most important is the synchrotron process, in which electrons moving at speeds very close to that of light (called relativistic speed) spiral around magnetic fields emitting radiation that is polarized—that is, which vibrates more strongly in a direction perpendicular to the magnetic field. This radiation was first found in the beams from synchrotron particle accelerators. The electrons may be a part of the energetic cosmic ray background flux in the Milky Way or may be produced in some violent event in the radio source itself. The precise shape of the continuous spectrum of synchrotron radiation depends, among other things, on the energy spectrum of the electrons.

Emission from neutral hydrogen. Line radiation (that is, radiation strongly concentrated toward a particular wavelength) at the 21.1-centimetre wavelength was first detected from clouds of neutral atomic hydrogen in the Milky Way in 1951. This radiation provides a very useful tool for studying the motion of many components of the Milky Way and of external galaxies. Motion in the line of sight toward or away from the observer is calculated from measurements of the Doppler shift—*i.e.*, a change in wavelength caused by the motion of the source—and measured from the 21-centimetre and other lines discovered later. Many interstellar spectral lines have now been detected from atoms and molecules that radiate in the radio range. Of these lines, the 21-centimetre line of neutral hydrogen, which appears in many parts of interstellar space, is very important because hydrogen is widely distributed in the universe as the building material for element formation in the interiors of stars.

THE SUN AS A RADIO SOURCE

The dominant radio source in the sky, the Sun is a huge and complex sphere of hot gas heated from inside by thermonuclear reactions (fusion of atomic nuclei). Lying above the surface that is seen with the naked eye are regions of tenuous gas and magnetic fields in which intense radio signals are generated. These vary from the highly variable emission characteristic of metre wavelengths, in which the intensity changes by factors of up to 10^3 in a few minutes, to a steady signal of about ten centimetres wavelength recorded from the so-called quiet Sun.

Radio emission from the quiet Sun. At times when there are no active areas (sunspots, flares, and prominences on the Sun), radio emission known as the quiet Sun component may be picked up. This emission originates in two layers already known optically: the chromosphere, which lies immediately above the photosphere, the visual surface, and extends for several tens of thousands of kilometres; and the corona, which is at a temperature of 1,000,000" K and extends upward outside the chromosphere to about 1,000,000 kilometres (a height above the surface equal to the Sun's radius). Radio radiation at long wavelengths is unable to escape to the Earth from the cooler chromosphere, and maps of the Sun at wavelengths of about three metres represent the radio appearance of the corona, with a diameter about twice that of the visual solar disk and a brightness temperature of about 10^6° K. At shorter wavelengths, about three centimetres, the radiation can escape from the chromosphere, and the diameter of the radio Sun at three centimetres is only a little larger than the optical disk, the brightness temperature being about 10^4° K. At intermediate wavelengths, the dimensions and brightness temperatures lie within these limits (see Figure 1).

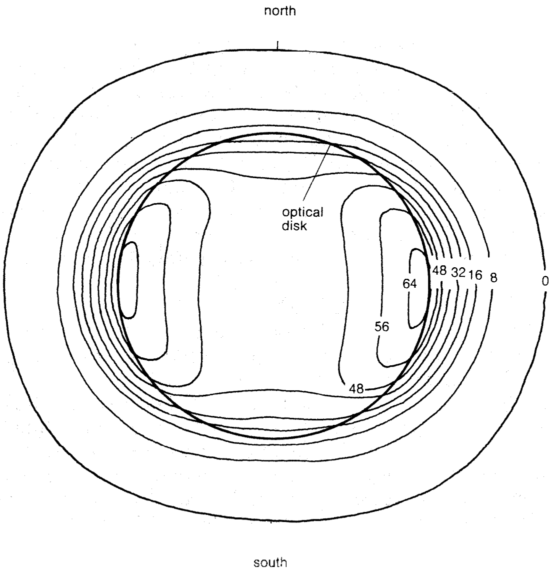


Figure 1: The quiet radio Sun at a wavelength of 20 cm showing contours of equal temperature in units of 1,000" K. Note the characteristic brightening at 20 cm wavelength on the east and west limbs, and the size of the radio Sun, which extends far beyond the optical disk (photosphere) into the chromosphere and corona.

Adapted from H.P. Palmer, *et al.* (eds.), Radro *Astronomy* Today (1963); Harvard University Press and Manchester University Press

Varying radio signals from the Sun. ***The so-called slowly varying component.*** A second part of the Sun's emission is the slowly varying component. This changes over the rotation period of the Sun, approximately 27 days, the rate at which sunspots also rotate. The slowly varying radio emission comes from extended regions larger in section than the sunspots, lying 10,000 to 50,000 kilometres (about 6,000 to 30,000 miles) above the spots and intimately connected with the bright areas called plages. Most of the radio emission is thermal radiation corresponding to the amount and distribution expected at a temperature of 10^6° to 10^7° K. Observations at centimetre wavelengths suggest an origin in magnetic fields with a strength of about 1,000 gauss (a gauss is the unit of magnetic field), which must reach to these heights from the sunspots (for comparative purposes, the strength of the Earth's dipole field is about one gauss as measured at the surface).

Eruptive, sudden, or catastrophic radio signals. Active regions of the Sun, usually near a sunspot, are the sites of solar flares, seen as a brightening in the red light emitted by hydrogen. These active regions are also the breeding grounds of more violent (catastrophic) radio events that are associated in some way with solar flares. Flares are accompanied by the outward ejection of material, and sometimes pre-existing arches of bright solar material are also induced to erupt. The simultaneous emission of X-rays causes immediate excess ionization—that is, breakup of atoms and molecules into charged particles (ions and electrons) in the low ionosphere of the Earth—and solar protons travelling at speeds near the speed of light reach the Earth's surface several hours later. Slower moving streams of particles that are also emitted reach the Earth about 30 hours later and produce auroras and magnetic storms. The total energy released in a solar flare may be as much as 10^{32} ergs (an erg is a unit of energy equal to one dyne centimetre; a man climbing a flight of stairs expends several million ergs) mainly in the form of light, X-rays, and protons. This energy is generally believed to come from the annihilation of part of the magnetic field lying above the sunspot group.

The eruptive component of the Sun's activity varies markedly with wavelength. At metre wavelengths the events are intense; individual bursts may last only seconds or minutes, while noise storms may last from hours to days. At centimetre wavelengths the activity is less intense, and the characteristic durations are from minutes up to an hour. Understanding of the eruptive component has come mainly from measurements of the changes in the spectrum—that is, the distribution over different wavelengths with time.

Sequence of radio outburst events

A large radio outburst will begin at the same time as the flare with a group of the so-called type III bursts; bursts of this type individually last about ten seconds and characteristically show a rapid increase in wavelength through the duration of the burst as the causative disturbance moves out rapidly through the solar atmosphere to levels from which the longer wavelengths can escape. These disturbances typically move at speeds about one-fifth to one-half that of light. Often, type III bursts occur simultaneously at pairs of wavelength bands separated in frequency by a factor of 2. Such a harmonic relationship suggests that this type of burst is generated by the solar plasma (a mixture of electrons and positive ions) by a process in which the plasma oscillates at a frequency that is dependent upon the electron density. Coincident with this group of metre wave bursts is a smooth microwave burst generated through the synchrotron process (see above) by the escaping stream of electrons moving at speeds near that of light in the sunspot magnetic field; this is accompanied by a burst of X-rays. Type III bursts are sometimes followed by a short period of continuum (spread continuously through all radio wavelengths) radiation called type V.

Typically, after a lapse in activity for several minutes, an intense type II burst begins. This type shows sharp band structure and generally has the harmonics attributed to plasma oscillations. The increase in wavelength and, consequently, the outward speed is about 100 times slower than for the type III bursts. The type II burst is believed to be caused by a shock wave originating at the same time as the fast electrons that give the type III burst but moving at only 1,000 kilometres (about 600 miles) per second.

Another type of eruption, called type IV, merges with and follows the type II bursts; it may last for several hours. This emission is quite different in character from the types II and III radiation. The radiation from all levels is spread over a broad band of wavelengths and is believed to be mainly synchrotron emission. It comes from a region several minutes of arc in diameter that moves outward at about 1,000 kilometres per second and may reach heights equivalent to several solar radii. This type IV emission is very complex, and several kinds of emission process are at work. The outward motions of the types II and IV bursts are comparable with those of streams of particles that produce magnetic storms and are probably different aspects of the same phenomenon (see Figure 2).

RADIATION FROM THE MOON AND PLANETS

The Moon. The planets are, on the whole, rather weak radio sources, and intensive studies have been necessary to derive their radio brightnesses. Radio measurements are particularly important in that they give the temperature at depth in the solid surface or in the atmosphere, which cannot be observed at optical wavelengths. The radio emission from the Moon's surface at any wavelength comes from a depth of ten times that wavelength (*e.g.*, from 100 cm at a wavelength of 10 cm). The uppermost surface layers at the centre of the lunar disk vary in temperature from 150" to 370" K (—200" to 200" F) between lunar night and day. At one-centimetre (radio) wavelength, the variation is 200" to 160" K (—100" to —170" F). It is clear, however, that solar heating does not penetrate to a depth of more than one metre (three feet) into the surface material of the Moon, because at 20-centirnetres wavelength, where the radiation from this depth is recorded, it is essentially constant at 230" K (—45" F).

Mercury. Mercury, nearest to the Sun and one of the most difficult planets to observe, has an overall temperature of 400" K (260" F) measured at three-centimetres wavelength. Early radio measurements indicated a change of no more than from 250" to 450" K between night and day on Mercury. This was something of a surprise because it had long been thought that Mercury always turned the same face to the Sun, and that the dark side, since it would never be directly heated, would be at a

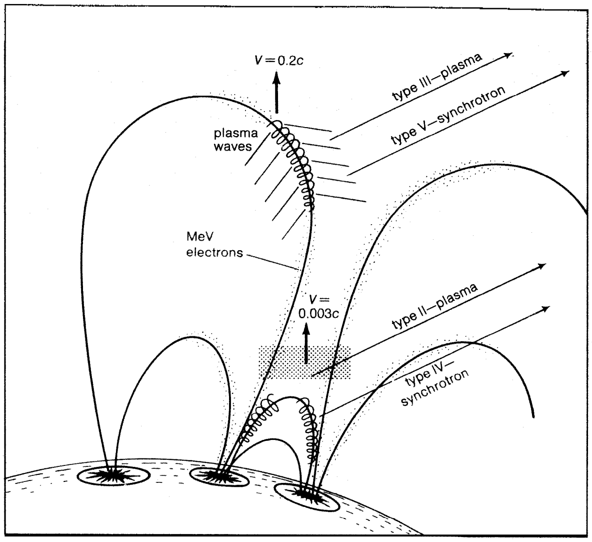


Figure 2: Origins *of* various types of radio *burst* activity on the ***Sun.***
The curves represent typical magnetic lines of force beginning and ending in surface areas (shown dark) related to flares. Type 11, III, and V originates in disturbances moving outward from the optical flare at speed V about 0.003 times the velocity of light, c, for Type II and at 0.2 times the velocity of light for Types III and V. Type IV is generated by relativistic electrons gyrating in the intense magnetic fields near the surface of the Sun.

Adapted from H.P. Palmer, *el al.* (eds.), *Radio Astronomy Today* (1963); Harvard University Press and Manchester University Press

temperature close to 0° K (—459" F). Radar observations resolved the discrepancy by demonstrating that Mercury rotates in 59 (±3) days, which is ⅔ of its orbital period of 88 days. This faster rotation makes the dark and light sides more nearly equal in temperature; at sufficiently long wavelengths, radiation would come from very deep layers, to which the day heating and night cooling does not penetrate, and the measured temperatures would be the same, as they are on the Moon.

Venus. One of the first surprises in radio studies of the planets occurred in the case of Venus and caused an important change in plans for spacecraft exploration. Measurements at wavelengths shorter than one centimetre indicated temperatures of 300" to 400° K (80° to 260" F), significantly above the infrared temperature of 240" K (—30" F). When these measurements were extended to wavelengths longer than three centimetres, the brightness temperature was found to be close to 600° K (620" F) at all wavelengths, indicating that the temperature of the solid surface is near 600" K, and the upper layers of the atmosphere are measured in the infrared, at 240" K. Temperatures between these values, from measurements at millimetre wavelengths, represent intermediate depths within the atmosphere of Venus. Radio measurements during a flyby by a spacecraft have confirmed these results obtained by ground-based radio telescopes. Other measurements from craft entering the Venerean atmosphere have shown that the pressure on the surface of Venus is about 100 times the atmospheric pressure on Earth. Herein lies the reason for the high surface temperature of Venus; the dense atmosphere, and particularly its carbon dioxide content, produces an intense greenhouse effect at the surface, trapping solar heat like a glassed-in enclosure. The observed radio temperatures and the measured density of the atmospheric constituents are consistent with this explanation.

Mars. The planet Mars has a mean infrared temperature of 217" K (—69" F) averaged over the light and dark hemispheres, and radio observations at the three-centimetre wavelength are consistent with this value. Nevertheless, the temperature calculated from measurements at shorter wavelengths is at 170" ±10° K, (—150" ±20° F), significantly lower. There is, as yet, no explanation of this difference. All these measurements refer to the surface layer of Mars and cannot be much affected by

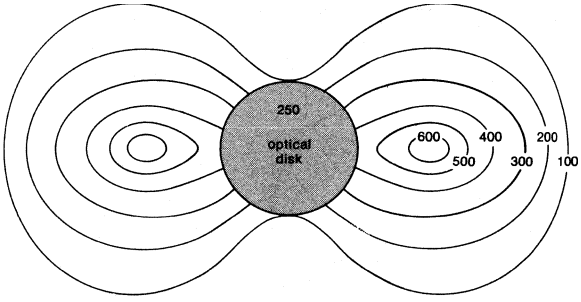


Figure 3: Radio emission (schematic) from Jupiter at 20 cm. At this wavelength, radiation from the Van Allen belts dominates with a maximum brightness temperature of about 600" K. The emission associated with the optical disk contributes less, and the temperature is only 250" K.

the Martian atmosphere as the surface pressure is about 1 percent of that on Earth.

The components of radio emission from Jupiter

Jupiter. Radio emission from Jupiter is interesting in that it has three distinct components. One is the emission from the dense atmosphere that makes up most of the visible disk of the planet. Its temperature, 140° K (—210" F), is the same at all wavelengths. The second component increases in temperature with wavelength and is produced well outside the optical disk (see Figure 3). This emission is strongly linearly polarized (*i.e.*, the radiation vibrates more strongly in some directions than in others) and originates in the strong magnetic field of Jupiter. Electrons of high energy are present within the field as they are in the Van Allen belts of high energy particles surrounding the Earth. These electrons emit radiation by the synchrotron process. Since the axis of Jupiter's magnetic field is inclined at nine degrees to the rotation axis of Jupiter, the intensity and polarization of the emission change as the planet rotates. These changes can be used to derive a very accurate period for the rotation of the planet's solid core, which is presumably responsible for the magnetic field. This value is nine hours, 55 minutes, 29.83 seconds (±0.26 seconds), compared with nine hours, 55 minutes, 40.6 seconds (±0.26 seconds) found for the rotation of the visible markings in the temperate zone. The third type of emission from Jupiter is restricted to wavelengths between eight and 30 metres and consists of sporadic bursts of high intensity. These bursts appear to be fixed at several longitudes on Jupiter for periods of years at a time and have a rotation period of nine hours, 55 minutes, 29.37 seconds. There is evidence for changes in the period of this long-wavelength radiation relative to that of the magnetic field, which indicates that the active centres of metre-wavelength emission are moving slightly in the atmosphere of Jupiter. The emission process of these bursts is not understood, but there is clearly an efficient mechanism that releases about 10^{18} ergs in each pulse. A remarkable modulation effect is connected with Jupiter's satellite Io, which has the power to turn off this emission even though Io is 420,000 kilometres (260,000 miles) from Jupiter. The cause of this effect is not clear, but it is possible that it is due to some interaction between Io and the magnetosphere of Jupiter.

Saturn, Uranus, and Neptune. Optically, Saturn is very similar to Jupiter, but at radio wavelengths the only similarity is the thermal emission from the gaseous disk, for which the temperature measured at a wavelength of several centimetres is 130" K (—225" F). A small but significant increase in emission temperature at longer wavelengths is restricted to the disk and is not like the Van Allen belt emission on Jupiter. No sporadic emission has been detected from Saturn, and no emission has so far been detected from its well-known rings. For comparison, the temperature of Saturn is 93° K (—290" F) measured in the eight-to-14-micron wavelength infrared band. Uranus and Neptune both have temperatures that rise gradually from about 100° K (—280" F) at three-millimetre wavelength to 170° K (—150° F) at three centimetres. In both cases, this radio brightness temperature, even at the shortest wavelength, is greater than the temperature expected from solar heating alone. The longer radio wavelengths come from greater depths within the clouds surrounding Saturn, Uranus, and Neptune, and it is conjectured that temperatures are higher because heat is generated within the planets.

RADIO_OBSERVATIONS_OF_THE_MILKY_WAY

Early radio sky maps

The earliest radio sky maps were relatively crude. They have since been systematically improved to give high-resolution surveys of the entire Milky Way over a wide range of frequencies. The surveys have shown the Presence of discrete radio sources both inside and outside the Milky Way, and a general background of emission covering the whole sky but strongly concentrated to the plane of the Milky Way, also called the galactic plane. The general appearance of the radio sky observed at a frequency of 200 megahertz is illustrated in Figure 4 which has been plotted in galactic coordinates. (Galactic longitude is measured around the galactic equator from the galactic centre, and latitude is measured toward the galactic poles from the equator). The highest brightness temperatures are found in the galactic centre region and the lowest at or near the galactic poles. A ridge of strong emission following the galactic plane has two components—one is thermal emission from ionized hydrogen and the other nonthermal synchrotron emission.

Thermal emission from the galactic plane. The thermal emission from the galactic plane is concentrated in a disk extending out about four kiloparsecs (one kiloparsec equals about 3,260 light-years, or 19.16 × 10^{15} miles) from the galactic centre. It lies in a slab extending about 200 parsecs from the galactic plane. The gas responsible for this emission is hydrogen in a state of high excitation at a temperature of 10,000" K. Excitation at this level is probably produced by a substratum of young stars that heat and ionize the hydrogen. Alternatively, however, it could be produced by the explosive events that occurred in the recent history of the Milky Way; these will be

Adapted from Radio Astronomy by J.D. Kraus (ed.). Copyright 1966: used with permission of McGraw-Hill Book Company

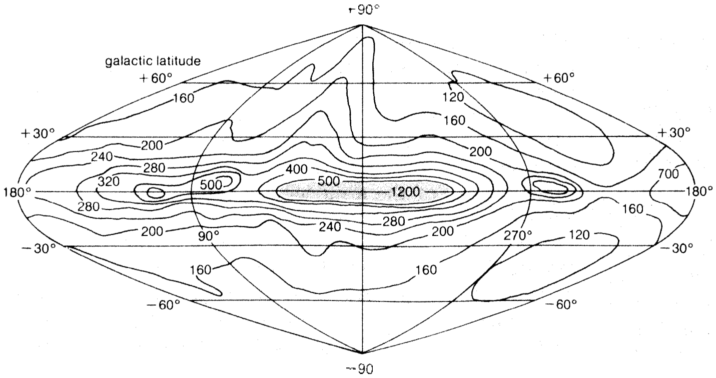


Figure 4: The radio sky at 200 megahertz plotted in galactic coordinates, with galactic equator dissecting the Milky Way. The strongest radiation is concentrated on the galactic equator: the highest radio brightness temperatures (about 1,200" K) are found near the galactic centre.

discussed in the next section. In this region of thermal emission the average ionized hydrogen density is about 0.2 atoms per cubic centimetre. The region contains about 4 × 10' times the mass of the Sun, compared with the total mass of the Milky Way of 1 × 10'' solar masses.

Presence of cosmic rays and magnetic fields

Nonthermal emission from the galactic plane. Studies of the latitude and longitude distribution of the nonthermal radio emission near the galactic equator indicate that it comes from a disk with a thickness perpendicular to the galactic plane of 0.7 kiloparsec and a radius of about nine kiloparsecs; on this scale the Sun lies ten kiloparsecs from the galactic centre. The discovery of this nonthermal emission component provided a new means of studying the cosmic ray electrons and the magnetic fields that thread through the Milky Way and that are responsible for the synchrotron emission from the disk. The radio spectrum of the synchrotron emission is identical with. that expected from the observed cosmic-ray energy spectrum. The observation of this radio emission shows conclusively that the cosmic rays are not a phenomenon existing only in the Sun's region of the Milky Way but are widespread. Radio observations of neighbouring galaxies indicate that cosmic rays are produced there too.

The radiation from the spiral arms. There is strong indication that some synchrotron emission is produced in the galactic spiral arms. This is certainly true of the Andromeda Nebula (a galaxy quite similar to the Milky Way), where the optical and radio spiral arms coincide.

The radiation from the halo. Another component of the nonthermal emission from the Milky Way is called the halo, which is difficult to distinguish from the disk component because of the presence of elongations, called spurs, that extend from many parts of the galactic plane, as can be seen in Figure 4. In all probability there is a weak halo of emission with a radius in the galactic plane of about 15 kiloparsecs and a thickness perpendicular to the plane of half this amount. If this is so, then magnetic fields and cosmic rays extend far beyond the confines of the galactic disk.

The 21-centimetre hydrogen emission. The richest increase in understanding of the Milky Way that has resulted from radio studies has come from measurements of the radiation produced at the 21-centimetre wavelength by cold clouds of interstellar neutral (*i.e.*, not ionized) hydrogen. In 1945, when radio astronomy was still in its infancy, a Dutch astronomer, H. van de Hulst, predicted that this emission should be detectable. Although the transition that produces a 21-centimetre photon from an atom of hydrogen is likely to occur in any particular atom only once every 11,000,000 years, van de Hulst believed that there would be so many hydrogen atoms, between 10^{20} and 10^{22} in a column a centimetre square extending in a line of sight through the Milky Way, that the emission would be detectable. This hope was realized in 1951, when the 21-centimetre line was first observed at a number of radio observatories. Observations pieced together over the whole sky show that the neutral hydrogen producing 21-centimetre radiation is distributed on a remarkably thin layer only several hundred parsecs thick and extending for about 15 kiloparsecs from the galactic centre. The neutral hydrogen is concentrated in long features that appear to be very similar to spiral arms seen in photographs of external galaxies. Optical studies of the spatial distribution of stars and regions of ionized hydrogen suggest that the optically visible constituents of spiral arms agree in position with these neutral hydrogen spiral arms.

Clouds of hydrogen

A wealth of additional structure is found in the neutral hydrogen distribution. High-resolution studies demonstrate that the spiral arms are composed of clouds ranging in size from rather less than one parsec to several tens of parsecs. A typical cloud may have a diameter of ten parsecs, a density of ten hydrogen atoms per cubic centimetre, and a temperature of 100" K (— 280" F). A line of sight through the Milky Way intercepts about ten such clouds per kiloparsec.

Clouds of molecules

Characteristic radio signals from molecules. The first radio observations of interstellar molecules were made in 1963 when the radiation of the hydroxyl (OH) radical was detected at a wavelength of 18 centimetres. Although the OH was found only in absorption against background sources, and the signals were weak compared with those from atomic hydrogen, it was immediately obvious that a new method was now available for studying the interstellar medium. Densities of OH in interstellar clouds were generally between 10^{-4} and 10^{-7} of those of atomic hydrogen. The greater OH abundances were in small clouds near the galactic centre, which also contain remarkably high abundances of other molecules (see below). This region appears to have the most ideal conditions of any discovered for the formation of molecules. Further study of the interstellar medium at 18 centimetres showed hydroxyl emission from dense dust clouds some distance from the galactic plane. These are believed to be regions of high gas density in which atoms and molecules coalesce to produce large aggregates in the form of dust grains. Hydroxyl emission from such clouds is therefore not surprising.

An entirely unexpected development in the hydroxyl story was the discovery (1965) of intense emission with unexpected properties from a new class of hydroxyl sources situated near ionized hydrogen regions, commonly called H II regions, such as the Orion Nebula and the nebula known as NGC 6334. The emission had such unusual features that it was at first called mysterium. Characteristically this anomalous emission has high intensities and spectral lines much narrower than expected; and, further, the ratios of the intensities of the four component lines of the.18-centimetre hydroxyl transition were not in agreement with theory. Additional peculiarities not seen previously in hydrogen-line studies were strong polarization (that is, the strength of the ratio signal depended on direction) and, in some cases, time variability. These characteristics have been investigated in many OH sources. Measurements with interferometers showed that one of these sources near the H II region known as W3 could be resolved into a number of separate bright patches scattered over an area two seconds of arc in diameter. Features of W3 ranging in size from 0.01 to 0.005 seconds of arc were distinguished. These were much smaller than could have been resolved with an optical telescope. The corresponding brightness temperatures of the emitting features are $10^{10\circ}$ to $10^{13\circ}$ K. The overall dimensions of the hydroxyl-emitting region correspond to those expected at a late stage of contraction of a gas cloud on its way to becoming a star.

Maser actions in interstellar gas

The anomalous hydroxyl emission is believed to arise as a result of the so-called maser process (maser stands for microwave amplification by stimulated emission of radiation), which has the property of amplifying background radiation at the wavelength of the masering atom or molecule by very large factors. This property is used in the operations of masers (at radio frequency) and lasers (at optical frequencies). Furthermore, the emitted spectral line is narrower than the normal line emission from the gas. What is not yet clear is how the hydroxyl molecules in interstellar space are "pumped" or "energized" so that there are more molecules in the upper state of the 18-centimetre level than in the lower state, a situation that must arise before the maser can operate. Some excitation of the molecules to a third, much higher state by a radiative or collisional process is thought to be responsible; the molecules would then fall from this "pumped" level, overpopulating the upper of the two levels involved in the 18-centimetre emission. Other groups of radio-frequency lines have been observed from OH that also appear to be a result of masering.

The year 1968 saw the beginning of a rapid increase in the discovery of new interstellar molecules by radio methods following on from the detection of hydroxyl radical (see also 'INTERSTELLAR MEDIUM). The discovery of formaldehyde (H_2CO) in 1969 was surprising, however, because this molecule contains both carbon and oxygen, each of which is about 10^{-4} of the abundance of hydrogen in interstellar space. There was no obvious way of telling which other molecules would be readily detectable, particularly since, if they masered, their emission would be considerably enhanced above that expected on the basis of abundance alone. Six new molecules were discovered in 1970. Each had two or more of the elements carbon, nitrogen, and oxygen. All the molecules detected at radio wavelength in the interstellar medium up to the end of 1970 are listed in the article INTERSTELLAR MEDIUM.

Of the recently discovered molecules, water vapour and formaldehyde have been studied in most detail. Water vapour has only been detected in the masering form; it shows very strong signals that come from small regions comparable in size to those found for the hydroxyl sources. Most water vapour sources come from the regions of anomalous hydroxyl emission. Evidently, the processes that lead to a concentration of hydroxyl also give a high abundance of water molecules, H_2O.

Probable inverse maser action

Formaldehyde is interesting in that it is widely distributed and is always recorded in absorption, even in regions of the sky where there is no continuum source to intercept the radiation and thus produce the absorption. In these regions it is apparently in absorption against the universal 3° K background radiation (see below). This means that formaldehyde can somehow come to a temperature beneath not only that of its surroundings but also beneath the temperature of the radiation field in which it is immersed. It is conjectured that this interstellar formaldehyde refrigerator is an inverse maser in which the lower energy level in the two responsible for the formaldehyde line emission is overpopulated relative to the upper level. The process of this population inversion is not yet clearly understood.

Radio observations of these molecules provide the basis for an entirely new branch of astronomy, interstellar chemistry. Attempts to understand the way in which the molecules are formed and the study of the role that these molecules may have had in providing building blocks for the origin of life on the Earth are important parts of this subject.

Radio observations of discrete sources. ***Radio details at the galactic centre.*** The Galaxy contains a large number of discrete radio sources, emitting both thermal and nonthermal radiation, scattered around the galactic equator and lying mainly within ten degrees of the galactic plane. At the centre of the Milky Way is a complex of many radio sources. The most intense of these, called Sagittarius A because it was the first discrete source discovered in the constellation Sagittarius, is about three minutes of arc in diameter and is believed to be the nucleus of the Milky Way; its position has been used along with other data to define the zero longitude point of the galactic coordinate system mentioned above. Sagittarius A is a nonthermal source presumably emitting by the synchrotron process. At the distance of the centre of the Galaxy, its angular diameter corresponds to a linear dimension of ten parsecs. Around the nucleus are a number of thermally emitting sources over a distance covering about 40 seconds of arc. These are ionized hydrogen with a mass of 10^5 times the mass of the Sun; the gas density is 20 atoms per cubic centimetre. Further out again is a radiating region, a nuclear bulge covering about four degrees in diameter along the galactic plane and one degree perpendicular to the plane. The radio signal from this region is synchrotron emission, and it is the strongest emission from the galactic centre region at longer wavelengths (about three metres).

Mixed with the ionized hydrogen and cosmic ray electrons are pockets of neutral gas. The absorption regions of neutral hydrogen, hydroxyl, and formaldehyde contain the largest known concentrations of these molecules. Most of the molecules so far discovered have also been recorded by their emissions in several regions within one degree of the nucleus. Ammonia, for example, is only found here near the galactic centre. It is evident that this limited volume of the Milky Way is very effective in producing molecules.

The motion of neutral hydrogen gas further outside this nuclear region has a strong circulation (see Figure 5). Out to a distance of two degrees (400 parsecs) there lies a thin disk which rotates in the plane of the Galaxy at a speed of rather more than 200 kilometres per second (120 miles per second). Around this lies a ring with a radius of 700 parsecs rotating at a similar speed. The neutral hydrogen density in these features is from one to three atoms per cubic centimetre, and its total mass is 5 × 10" solar masses. Further out still from the centre, at a distance of four kiloparsecs, is a dominating feature that is expanding outward at 50 kilometres (30 miles) per second, as well as taking part in galactic rotation. The existence of this feature known as the four-kiloparsec arm is a major enigma in the understanding of the Milky Way. It has a mass of 3×10^7 solar masses, and a considerable amount of energy will have been needed to give it an outward speed of 50 kilometres per second; further, its existence implies a large circulation of material through the central regions of the Milky Way. Another phenomenon that some astronomers consider may be related to the outflow of gas in the galactic plane from the centre is the infall of neutral hydrogen clouds towards the galactic plane at velocities of up to 200 kilometres per second in the vicinity of the Sun.

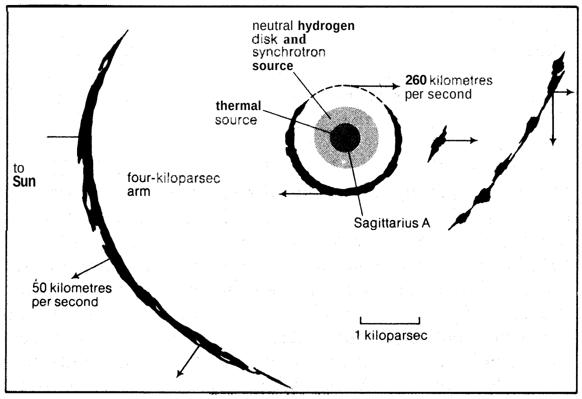


Figure 5: The distribution of neutral hydrogen in the galactic centre region.
The Sun is 10 kiloparsecs to the left of centre. The outer neutral hydrogen features show evidence for expansion from the centre. The central rotating disk contains an extended synchrotron source, a thermal source, and the small-diameter source Sagittarius A.

Many theories have been proposed to explain the phenomena just described in the central regions of the Milky Way. Broadly they fall into two classes: in one, flow of gas is considered as a part of a general circulation of material in the Milky Way, though no convincing source of the material has been proposed; in the other, the outward motion is thought to be the result of the recent explosion or series of explosions in the galactic centre. It is known from other measurements that about 10^9 solar masses of material lie within 100 parsecs of the centre, probably in the form of stars. Considerable activity must have occurred recently, or must still be happening, to produce the electrons moving at nearly the speed of light that are responsible for the two nonthermal sources in the centre. Such fast electrons can exist only for periods of about 10^6 to 10^7 years, a very short time compared with the age of the Milky Way (10^{10} years).

Radio signals from supernovae remnants. One of the two most abundant types of discrete radio source found in the Milky Way are the supernovae remnants. These are the remains of stars that have violently exploded into supernovae, undergoing a sudden tremendous brightness increase and sending out an expanding volume of gas at speeds of several thousand kilometres per second. Embedded in this volume are the very fast moving (relativistic) electrons that emit strong radio signals in the ambient magnetic fields. The first supernova remnant to be identified as a radio source was the Crab Nebula. In the year 1054 a supernova in the position of the Crab Nebula was seen to explode. It became so bright that it could be seen in the daytime. The wisps of glowing gas resulting from this explosion were first noted in 1731, but the true nature of this nebulosity was not realized until the 20th century. The present expansion velocity of the nebula is 1,150 kilometres (715 miles) per second; its

Table 1: Well-known Supernova Remnants

remnant	galactic coordinates		angular size*	distance (kiloparsec)	year of supernova	radio flux at 30 cm†
	long.	lat.				
Kepler's supernova	4.5	+6.9	2.2 × 2.2	11.4	1604	20.0
Cygnus Loop	74.0	–8.6	200 × 160	0.6	approx. 70,000 years ago	180
Cassiopeia A	111.7	–2.1	4.0 X 3.8	2.7	1700	3,000
Tycho's supernova	120.1	+1.4	6.0 × 7.0	4.9	1572	58
S147	180.0	–1.7	180 × 180	0.7	...	120
Crab Nebula	184.3	–5.8	3.0 X 4.2	1.7	1056	1,000
Monoceros nebulosity	205.5	+0.2	210 X 210	0.6	...	150
Puppis A	260.4	–3.3	55 × 55	1.2	...	145
Lupus Loop	330.0	+15.0	270 X 270	0.4	...	340
NGC 6383	355.2	fO.l	7.5 × 9.5	4.9	...	30

*Minutes of arc. †Watts m^{-2}Hz^{-1} × 10^{26}.

diameter is about three parsecs, corresponding to an angular diameter of six minutes of arc at a distance of 1,700 parsecs. Radio emission is generated throughout the Crab Nebula by relativistic electrons (moving at very fast speeds, near that of light) some of which were produced at the time of the supernova explosion and have been continually produced since that time. Relatively strong magnetic fields (compared to the interstellar fields) of about 10^{-4} gauss are deduced from the measured intensity of the optical and radio synchrotron emission.

The brightest radio supernova remnant is Cassiopeia A (Cas A) in the constellation of Cassiopeia. Its shell structure is more typical of supernova remnants than is the centrally filled Crab Nebula. Although the optical object lies behind dense dust clouds, the outlines of a filamentary shell two minutes of arc in radius can be discerned. Its expansion speed is 7,000 kilometres (about 4,000 miles) per second, and estimates of its age calculated from the expansion velocity suggest that the Cas A supernova outburst occurred in AD 1702 (&14). The simple shell structure of Cas A implies that all the material now confined in a thin outer shell was emitted in a single explosion. Because of the gradual expansion of the shell, the radio luminosity is expected to decrease with time. Such a decrease has, in fact, been detected and amounts to 1.1 percent per year.

Estimate of radio wave-length energies in brightest supernova source

The kinetic energy in the shell of the Cas A supernova explosion is estimated at 10^{51} ergs. Kinetic energy is energy of motion. The energy in cosmic rays and electrons is now about 10^{49} ergs, and the total that has been radiated since the initial outburst at all wavelengths from radio to X-rays is approximately 10^{47} ergs. These energies must be compared with the total nuclear energy available from conversion of hydrogen to helium in the original star of about 10'' ergs. Evidently, about 1 percent of this energy has been released into the expanding remnant.

Three other supernova explosions have been seen in the Milky Way: the supernovae of AD 1006, Tycho's supernova (AD 1572), and Kepler's supernova (AD 1604). All are shell-like sources of radio energy. The expansion velocities calculated for Tycho's and Kepler's supernovae exceed 10,000 kilometres (6,000 miles) per second, appreciably greater than those found in the Crab Nebula.

About 100 radio sources believed to be supernovae remnants have been found. Some, like the Cygnus Loop, IC 443, and S147, were later identified with optical nebulosities, but most lie in obscured parts of the Milky Way where no optical identification is possible. The last mentioned radio sources are considered to be supernova remnants because they have a nonthermal, presumably synchrotron, spectrum and, in many cases, a shell-like structure. Most of these supernova remnants are older than the bright compact remnants described in previous paragraphs. The oldest remnants that are still distinguishable at optical and radio wavelengths have ages of about 10" years; after this, the expanding gas becomes tenuous and merges with the rest of the interstellar medium. Some of the better known supernova remnants are listed in Table 1 above.

Analysis of the number and lifetimes of supernova remnants indicates that the rate of supernova explosions in the Milky Way is one every 30 years, a rate comparable to that found in similar external galaxies. A supernova is not seen to erupt once every 30 years, however, because most must occur behind the dense dust clouds in the plane of the Milky Way.

Pulsed radio signals. A class of galactic radio sonrces believed to be closely related to the supernova remnants comprises the pulsars, sources that emit their radio waves in periodic pulses. They appear, like the H II regions and nonthermal disk radiation, to be concentrated to the galactic plane in a layer several hundred parsecs thick. About 50 of these sources are now known (see the article PULSARS).

Radio signals from individual nebulae. Individual emission nebulae, concentrations of ionized hydrogen glowing by their own light, form another group of galactic radio sources. Such objects include the great nebula in Orion (M42), the Rosette Nebula (NGC 2237–46), and the Omega Nebula (M17). All of these ionized hydrogen regions, usually called H II regions, are dense clouds of hydrogen gas in which bright (young) stars have recently formed. These stars emit intense ultraviolet light of sufficiently short wavelength to ionize much of the hydrogen clouds. The ionized clouds then produce thermal radiation at radio wavelengths that can penetrate the complete depth of the Milky Way and give valuable information about regions that are optically invisible.

The Orion Nebula at radio wave-lengths

The Orion Nebula, in the "sword" of the constellation Orion, is one of the most extensively studied H II regions. Although faint optical nebulosity can be traced out 30 minutes of arc from the centre, the most light and radio emission comes from a central region about four minutes of arc in diameter, which corresponds to 0.4 parsec at its distance of 400 parsecs. The average density in the central regions of the Orion Nebula is 2,000 atoms per cubic centimetre, and its total mass is 100 solar masses. This is about 10 percent of the mass of the cluster of young stars called the Trapezium that has condensed from the nebula and now heats it to its equilibrium temperature of 10,000" K (18,000" F). Two bright regions near the centre of the nebula have densities of 10,000 atoms per cubic centimetre. This high central concentration in the nebula and the presence of very young stars at its centre indicate that the Orion Nebula is less than 10^4 years old, very young in the astronomical time scale.

The Rosette Nebula at radio wave-lengths

A much older H II region (the Rosette Nebula) is found in the constellation Monoceros. Its linear diameter is about 16 parsecs (52 light-years), and its mean density is 20 atoms per cubic centimetre. The Rosette Nebula is a shell object having a broad central hole in distinction to the centrally bright Orion Nebula. Estimates of its age range between 10" and 10' years. Radio studies of nebulosities are particularly valuable in giving a clear picture of the distribution of emission throughout the entire depth of the region, in contrast to optical studies, which can be severely limited by obscuration. Most H II regions contain large quantities of dust mixed with the gas. Some of the brighter ionized hydrogen regions are listed in Table 2 below.

The planetary nebulae, often elliptical in shape, sur-

Table 2: Well-known Galactic H II Regions

H II region	galactic coordinates		angular diameter*	distance (kilo-parsec)	radio flux at 6 cmt
	long.	lat.			
M8 (Lagoon Nebula)	6.0	−1.2	10′	1.4	192
M17 (Omega Nebula)	15.0	−0.7	6′	2.4	784
M16 (SC 6611)	16.9	+0.8	15′	2.2	114
NGC 6604	18.5	+2.0	30′	3.2	70
IC 1793	133.7	+1.3	10′	2.7	110
Rosette Nebula	206.3	−1.9	40′	1.4	300
NGC 2024	206.5	−16.4	3′	0.6	65
Orion Nebula	208.9	−19.3	3′	0.4	470
RCW 38	267.9	−1.1	15′	1.0	280

*Minutes of arc. †Watts $m^{-2}Hz^{-1}$ × 10^{26}.

The planetary nebulae at radio wavelengths

rounding extremely hot central stars, form a quite distinct class of H II regions. The nebulosity has been ejected from the star and has a temperature higher than that of ordinary H II regions. The central stars of planetary nebulae are believed to be near the final stages of evolution, whereas the stars associated with normal H II regions are in the earliest stages of evolution.

The final class of galactic radio sources to be described is that of flare stars. These are red dwarf stars just beginning their long lifetimes; from time to time they produce flares that are much more intense than those that may be seen on the Sun and that may increase the star's brightness over a short period (usually minutes or hours) by a factor of two or more. The flares are accompanied by nonthermal radio outbursts that can be detected weakly at the Earth. The prototype star of this group is UV Ceti, about three parsecs distant. Radio bursts that accompany the optical flare can last for several tens of minutes.

RADIO SOURCES BEYOND THE MILKY WAY

In 1951 only about 100 radio sources were known; as of the early 1970s, about 20,000 had been cataloged, most of them lying outside the Milky Way. Of those whose angular sizes have been measured, the majority are found to have diameters less than a few minutes of arc and half have diameters between three and 30 seconds of arc. A few percent of the sources, or at least significant components within them, may have diameters of a few hundredths or thousandths of a second of arc. From a wealth of data some clear-cut patterns have emerged. The extragalactic radio sources can be divided into two classes—those associated with optically recognizable galaxies and those identified with remote starlike (quasi-stellar) objects. At the present time only about 300 of all the extragalactic radio sources are identified with optical objects with measured red shifts (a reddening of light proportional to distance), and it is possible that other classifications may emerge when all the radio sources are investigated optically.

Radio galaxies. Neutral hydrogen has been detected in more than 100 galaxies extending to distances of 40 megaparsecs (40 × 10^6 parsecs). A clear pattern emerges in the relationship between the neutral hydrogen content of a galaxy and its morphological type. Irregular galaxies, such as the Magellanic Clouds, contain several tens of percent of their mass in neutral hydrogen, while spirals contain a few percent and elliptical galaxies less than a few tenths of a percent, if any at all. This pattern is believed to be the result of the different rates of evolution of the different morphological types—the ellipticals evolved rapidly, using up their supply of hydrogen in an early phase of star formation, while the spirals have evolved more slowly and still possess some hydrogen that can be used in future star formation.

The nearby spiral galaxies, all of which have constituents similar to those of the Sun's galaxy (*i.e.*, contain both young and old stars and include at least several percent of their mass in gas and dust), give detectable radio emission, comparable in total intensity to that emitted by the Milky Way and amounting on average to 10^{38} ergs per second integrated over all radio frequencies. These are called "normal" galaxies in distinction to the giant radio galaxies that emit 10^{40} ergs per second and the quasi-stellar radio sources that emit 10^{44} to 10^{46} ergs per second. Among the normal galaxies those with elliptical shapes, which contain little gas and dust and few young stars, are the weakest radio emitters.

Radio studies cf the Andromeda Nebula. The Andromeda Nebula (often referred to by its number in the catalog of the 18th-century French astronomer Charles Messier as M31) is the nearest spiral galaxy and is particularly well placed for detailed study. It has a structure very similar to that of the Milky Way, and many comparisons are useful. Its continuous radio emission has been detected between wavelengths of ten metres and three centimetres, and its integrated emission (that is, the total signal from the whole galaxy) shows a nonthermal spectrum that is presumably generated by fast-moving electrons in an interstellar magnetic field. This nonthermal emission comes from an area measuring six degrees by four degrees with its long axis parallel to the major axis of M31. The outer optical envelope may be traced over an area measuring only four degrees by one degree. The greater extent of the radio source is direct evidence for a halo surrounding M31 and is more convincing than is the case for a halo in the Milky Way described above.

In a high resolution map of M31 at 75-centimetres wavelength details four minutes of arc apart can be distinguished, and the spiral arm radiation can be seen; it is strongest at the position of the dominant optical spiral arms lying between 30 and 80 minutes of arc from the centre along the major axis. A source with a small angular diameter (80 seconds of arc = 200 parsecs) is found at the centre of M31, coincident with its optical nucleus. This source is comparable in intrinsic brightness to the extended component at the nucleus of the Milky Way; but there is no compact source at the nucleus of M31 similar to the bright source Sagittarius A at the centre of the Galaxy.

Like the Milky Way, M31 produces strong 21-centimetre-wavelength emission from its neutral hydrogen. This radiation is strongest in the vicinity of the major spiral arms but lacking in the central regions of the galaxy. The neutral hydrogen extends farther out from the centre than the stars or H II regions. A detailed comparison with the optical data shows that the young stars and their associated H II regions lie in the area of greatest neutral hydrogen density at 40 to 80 minutes of arc (eight to 16 kiloparsecs) from the centre. It appears that the hydrogen in the central regions has already condensed into stars and that the neutral hydrogen density in the outermost regions is too low to permit star formation at the rate found at intermediate positions. Such direct comparisons between the neutral hydrogen and optical data are valuable in understanding the processes by which stars form in the interstellar gas.

The Virgo A and Centaurus A radio galaxies. Only the nearby normal galaxies can be detected in either 21-centimetre-line emission or continuum emission; *i.e.*, radiation spread continuously over a wide range of radio wavelengths. The discovery of anomalously bright radio galaxies extended by a factor or 50 or so the depth to which the universe could be examined at radio wavelengths. Among the first galaxies in this class to be identified were Virgo A, the brightest radio source in the constellation Virgo, and Centaurus A, the brightest in the constellation Centaurus. These sources both happen to be close by and outshine normal galaxies by a factor of 100 or more, in absolute brightness at radio wavelengths. Virgo A has been identified with the giant elliptical galaxy known by its Messier catalog number M87; M87 is the brightest member of the Virgo cluster of galaxies at a distance of 12 megaparsecs. The main radio emission has an extent similar to that of the optical object. Short-exposure photographs show that the object is very unusual in having a bright blue jet that extends a kiloparsec from the nucleus. Radio measurements reveal a compact component, radiating a few percent of the total flux, coincident with the jet. The presence of this jet indicates that the nucleus of M87 is still active. Many astrophysicists have proposed that the main radio source was produced by a similar central explosion but on a vaster scale.

Centaurus A has been mapped over a region ten degrees in diameter, which at its distance corresponds to one megaparsec (10^6 parsecs, or 3.26 × 10' light-years) in overall extent, the largest linear size of any known radio source. It has two outer components of low surface brightness and two intense compact central components embedded in the outer envelope of the galaxy NGC 5128. This is also a peculiar elliptical galaxy, being crossed by a dark band of absorbing material that is a very unusual feature in an elliptical galaxy. The dark material and peculiar velocities found optically in NGC 5128 are again indicative of some central explosion that is presumably also responsible for the ejected radio plasma. In this galaxy two explosions appear to have occurred; one produced the outer diffuse emission, and a more recent explosion produced the compact inner components. Centaurus A was one of the first extragalactic radio sources in which linearly polarized radio emission was detected. Radiation from the outer components of the source is as much as 20 percent polarized. Synchrotron emission is evidently responsible.

Cygnus A

Very distant radio galaxies. By the early 1950s much more distant radio galaxies of even greater intrinsic radio luminosity were discovered. The brightest extragalactic source in the sky, Cygnus A, was identified with a peculiar double galaxy that displayed a red shift in its light equivalent to a speed of recession of 16,000 kilometres (10,000 miles) per second. The presently accepted relationship between red shift and distance is roughly 100 kilometres (60 miles) per second per megaparsec; accordingly, the distance of Cygnus A is 160 megaparsecs. Cygnus A was at first thought to be a pair of galaxies in collision in which the energy of the radio emission came in some way from the energy of motion of the two colliding galaxies. It became evident, however, that, in this system and other double systems, the collisional hypothesis was inadequate to account for the estimated energy and that an explanation in terms of some central explosion was more promising. In Cygnus A the presumed explosion has produced optical radiation in two adjacent volumes, while the two plasma clouds containing the relativistic (fast) electrons moving near the speed of light and magnetic fields responsible for the radio emission were ejected to much greater distances and were now separated by 80 seconds of arc (60 kiloparsecs). The rate of radio emission from Cygnus A is 10^{44} ergs per second, and the estimated value for the energy in the relativistic electrons and magnetic fields to 10^{60} to 10^{61} ergs. This is a minimum estimate of the total energy at the present time stored in this galaxy. This amount of energy is about 10^{-4} of the Einstein rest mass (Mc^2 in which M is the mass and c is the velocity of light) of the entire galaxy—a radio emission so inexplicably large that it poses one of the major problems of astrophysics at the present day.

The Seyfert galaxies

The observed properties of the radio galaxies can be summarized briefly. They are often double or multiple sources in which the radio components are situated on either side of the parent optical galaxy. The associated optical galaxies tend to be surprisingly alike in absolute luminosity, with an average magnitude of —21 (4 × 10^{10} times the intrinsic brightness of the Sun). One group of radio galaxies is identified with the Seyfert galaxies (named after Carl Seyfert, the American astronomer who first studied them), which are spirals with particularly active nuclei. The radio luminosity increases from its lowest value in the irregular galaxies, through the giant ellipticals (*e.g.*, Virgo A) to dumbbells (double galaxies —*e.g.*, Cygnus A) to the D galaxies, which are single galaxies each in an extensive envelope, and N galaxies, which have a bright nucleus in a faint envelope. Many pose the energy problem described for Cygnus A. The most distant radio galaxy discovered so far is 3C295, whose red shift indicates that it is receding at 0.46 times the speed of light. It is ten times further away than Cygnus A but is similar in all other properties.

Quasi-stellar radio sources. Quasi-stellar objects are discussed fully in QUASI-STELLAR SOURCES. A summary of their properties sufficient for the present purpose follows. Quasi-stellar radio sources (quasars) are identified with certain optical objects that produce starlike images on photographs. These are believed to be not stars belonging to the Galaxy but extragalactic objects photographically unresolved. The associated radio sources, on the other hand, are generally not so small and they have sizes and shapes similar to those of radio galaxies. Radio measurements alone, therefore, cannot distinguish whether a particular object is a radio galaxy or quasar. There are, however, some statistical differences between the two groups. The extragalactic nature of the quasi-stellar radio sources was established from observation of the nearest one, number 3C273 in the third Cambridge catalog of radio sources. A compact component in 3C273 coincides with a starlike object, and an extended component 20 seconds of arc away is situated at the end of an optical jet directed away from the "star." Recent radio studies have shown that the compact source (called component B) has several more or less concentric components, the smallest of which has an angular diameter of 0.0003 seconds of arc. This corresponds to a linear diameter of one parsec if it is assumed that the object's red shift is caused by distance and that the usual rules for finding the distances of galaxies apply. The radio emission from component B is variable over periods of years; an examination of photographs made since 1890 demonstrates that the quasi-stellar object has varied also in optical intensity by more than a factor of two with characteristic time scales generally of about ten years but sometimes in a period of a few months. These latter variations imply that some optical components must be smaller than a few light-months (*i.e.*, 0.1 parsec) in diameter. This is of an order of size similar to that of the radio measurements of size mentioned previously. Clearly, vast quantities of energy are being released in 3C273. A typical outburst of this type releases about 10^{58} ergs of energy in the form of relativistic (fast) electrons and magnetic fields.

About 150 quasi-stellar objects associated with radio sources have been studied in some detail, and their red shifts, presumably a measure of their distances, have been determined. The quasi-stellar radio sources as a group show flatter radio spectra than the radio galaxies. Many seem to emit less energy at low frequencies, indicating sufficiently large numbers of fast-moving electrons and magnetic fields to produce synchrotron self-absorption; that is, the energy produced by the accelerating electrons is re-absorbed by other electrons nearer the observer. Those sources with flat spectra have observable components with the smallest angular diameters—in some cases as small as 0.001 second of arc. A few radio galaxies also have this characteristic. Other quasi-stellar radio sources produce extended radio emission spread over ten kiloparsecs or more; this is considered to be the result of some explosion that occurred millions of years ago. In such cases, the related central optical object has continued to emit. Some, like 3C273 with several high and low brightness components, show evidence for multiple explosions. One major difference between radio galaxies and quasi-stellar radio sources is the greater rate of optical and radio emission from the quasi-stellar sources. Since the energy in magnetic fields and fast-moving electrons is comparable in the two types of object, the quasi-stellar radio sources must have shorter lifetimes than galaxies by a factor of ten to 100; typically, their age is several times 10^6 years. Their density in space is considerably less than that of the radio galaxies. Evidence suggests that there are approximately 10' radio galaxies for every quasi-stellar radio source in the nearby region of the universe.

The distribution of very distant radio sources. The methods of measuring distance that have been developed by optical astronomers can be used to derive distances for those radio sources that are identified with a galaxy. Such methods can measure distances to the nearby galaxies with an accuracy of between 10 and 20 percent. In the case of a more distant galaxy, apparent brightness can be used as a distance indicator so long as its absolute brightness can be assessed from its morphological type or spectrum.

In radio and optical astronomy it is now possible to investigate objects at very large distances; consequently, the properties of these objects are assessed as they were a long time ago when light and radio waves set out on their journey to the Earth. Assuming that their red shifts are true distance indicators, many of the quasi-stellar sources are so far away that the emission received from them set out when the universe was less than half its present age. Some astronomers, however, have argued that the red shift may be a gravitational effect in objects at intermediate distances, claiming that the expected correlation between red shift and apparent magnitude for quasi-stellar sources does not exist.

The understanding of the evolution of the universe does not have to rely only on those radio galaxies and quasi-stellar sources whose distances are known. The problem can be approached statistically. If it is supposed that the universe is uniformly filled with a population of radio sources of the same absolute luminosity, then in counting all the sources out to some distance (*R*), at which the apparent luminosity is proportional to R^{-2}, the number of sources will be proportional to R^3. It can be seen then that the number of sources, N, down to some limiting apparent luminosity, S, is proportional to the inverse three-halves power of S—that is, to S-%.This is true even if the sources are not all of the same absolute brightness but are nevertheless mixed in space. Many such counts of radio sources have been made and reliable data is now available down to very low apparent brightness levels. A plot of the logarithm of N versus the logarithm of S taken from the Cambridge surveys is shown in Figure 6. In a uniform universe the slope of this plot should be the power of *S* shown above; that is, —1.5. Figure 6 and similar plots using different sample source counts indicate that the slope for the brighter sources (the bottom right-hand end of plot) is close to —1.8 and is definitely steeper than —1.5, and the results are interpreted as indicating an excess density of apparently weaker and therefore more distant sources. This higher density is then taken to be characteristic of the earlier epochs in the evolution of the universe. Such a decrease of source density from those early stages to the present is precisely what is expected in an expanding universe and is at variance with any simple steady-state model of the universe. At the very faint end of the log N–log *S* plot the number of radio sources falls well below the expected 1.5 slope. This has been explained as the effect of a low rate of formation of radio sources in the first few tens of percent of the age of the universe; at these early epochs conditions were presumably not favourable for the formation of radio sources.

Adapted from *Annual Review of Astronomy and Astrophysics* (1968)

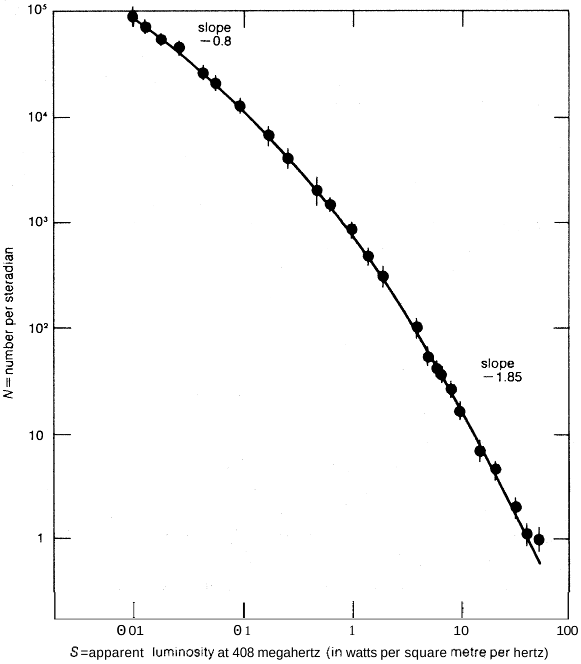


Figure 6: Observed relationship between number of radio sources (N) as a function of limiting apparent luminosity (S) plotted on a logarithmic scale.

The correct interpretation of the log N–log S plot is not entirely clear. There is no reasonable doubt that it can be fitted to a model of an evolving ("big-bang") model as described in outline above. There may be other explanations—for example, in terms of an oscillating model. Indeed, proponents of the steady-state models claim that the data can be fitted to their theories with appropriate modifications to their original concepts. Many astronomers feel uneasy about interpreting log N–log *S* data that include a significant fraction of quasi-stellar sources whose nature is still not clear.

There is one more piece of observational data that radio astronomy has contributed to the cosmological argument: the 3° Kelvin universal blackbody (thermal) radiation. This radiation, which appears to come impartially from all directions, has a brightness temperature at all wavelengths of about 3° K; more accurate recent measurements indicate a value of 2.7" K (—454.8" F). The integrated radiation from the estimated background (of unresolved radio sources) has been allowed for in the measurements. At wavelengths shorter than 10 centimetres this contribution is considerably less than 2.7" K radiation, in any case. In evolutionary cosmologies this radiation is explained as a residual glow from the expanding radiation field generated at the time of the "big bang." Steady-state cosmology has no explanation for it.

BIBLIOGRAPHY. J.S. HEY, *The Evolution of Radio Astronomy* (1973), a history of the field of radio astronomy; also by Hey, *The Radio Universe,* 2nd ed. (1976); and F. GRAHAM SMITH, *Radio Astronomy,* 4th ed. (1974), two introductory surveys of this field; I.S. SHKLOVSKY, *Cosmic Radio Waves* (1960; orig. pub. in Russian, 1956), a description of results obtained up to 1960, providing lucid theoretical background to the subject; J.L. STEINBERG and J. LEQUEUX, *Radioastronomie* (1960; Eng. trans., 1963), a well-illustrated survey of the subject up to 1962; G.L. VERSCHUUR, K.I. KELLERMANN and V. VAN BRUNT, *Galactic and Extra-Galactic Radio Astronomy* (1974), a comprehensive summary of basic theory and research findings; MR. KUNDU, *Solar Radio Astronomy* (1965), a comprehensive description of the observations and theories of the many forms of solar radio emission; B.M. MIDDLEHURST and L.H. ALLER (eds.), *Nebulae and Interstellar Matter* (1968), a textbook containing treatments of the radio as well as the optical data; R.D. DAVIES and F.G. SMITH (eds.), *The Crab Nebula* (1971), proceedings of an IAU symposium devoted entirely to a discussion of the Crab Nebula and its associated pulsar. Review articles on radio sources appear from time to time in the *Annual Review of Astronomy and Asrrophysics;* of particular interest is the article "Counts of Radio Sources," by M. RYLE, 6:249–266 (1968). See also K.I. KELLERMANN, "Extra-Galactic Radio Sources," *Physics Today* 226:38 ff. (October 1973); and R.G. STROM, G.K. MILEY, and J. OORT, "Giant Radio Galaxies," *Sci. Amer.* 233:26–35 (August 1975).

(R.D.D.)

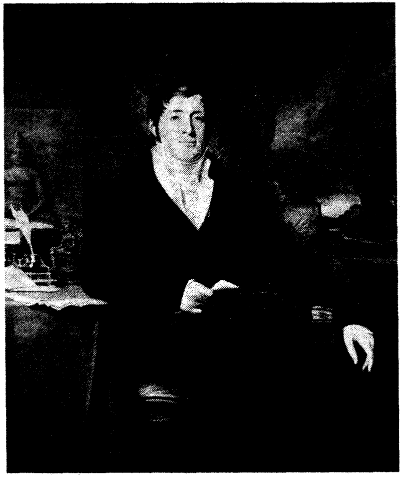
Raffles, Sir Stamford

Sir Thomas Stamford Raffles, as employee of the British East India Company, was one of the founders of Britain's empire in the Far East. In 1819 he founded Singapore, which was to become the vital link in the China trade.

Early life and the beginning of East India service

Born to an improvident. merchant captain and his wife on July 6, 1781, during a homeward voyage from the West Indies, Raffles grew up in an atmosphere of debt. Forced to cut short his schooling at the age of 14, he entered the service of the East India Company as a clerk in order to support his mother and four sisters. Although his formal education was inadequate, he studied the sciences and several languages at his own leisure and conceived an interest in natural history that was to earn him a distinguished reputation. Both in and out of office hours, his industry won him such notice that at the age of 23 he was appointed assistant secretary to the newly formed government of Penang, a hitherto inconspicuous island at the northern entrance to the Strait of Malacca.

In Penang, established in order to give England a foothold in the Dutch-held East Indies, Raffles shaped his career by an intensive exploration into the language, the history, and the culture of the Malayan peoples scattered over the islands of the archipelago. This unique study



Raffles, oil painting by G.F. Joseph, 1817.
In the National Portrait Gallery, London.
By courtesy of the National Portrait Gallery. London

caught the attention of Lord Minto, governor general of India, at a time of crisis, when Napoleon was using Java as a springboard for the destruction of England's slow and lumbering ships, the Indiamen, on the long haul to China. Determined to remove Java from French influence, Minto appointed Raffles his agent to prepare the way for a naval invasion.

Entrusted with an independent authority that aroused jealousy in Penang, Raffles established his headquarters in Malacca. Rewarded for his extraordinary work by an appointment to Minto's staff, Raffles sailed with him to Java, where the expeditionary force landed without mishap on August 6, 1811, and, after a short and sharp engagement with the Dutch–French forces, occupied the island. Minto gave considerable credit for the success to Raffles. Having already described him as "a very clever, able, active and judicious man," he now recognized his intellectual and administrative ability and his humanism and concern for the Javanese, and on September 11 he proclaimed him lieutenant governor of Java. Shortly afterward Minto sailed for Calcutta, leaving Raffles at the age of 30 to rule not only Java but an archipelagic empire of several million inhabitants.

Reforms of the Dutch system in Java

Raffles inaugurated a mass of reforms aimed at transforming the Dutch colonial system and improving the condition of the native population. His reforms, however, proved too costly to a trading company primarily concerned with profit and were short-lived. After four and a half years in Java, suffering from increasing ill health and shattered by the death of his wife, he was recalled. Left vulnerable to personal attack by the death of Minto, he sailed for England on March 25, 1816, thoroughly out of favour with the court of directors of the East India Company.

He never regained their full confidence. Despite a dazzling London success in both fashionable and learned society that culminated in his election as a fellow of the Royal Society and the award of knighthood, he resumed his Eastern service in a situation of reduced and restricted authority, as lieutenant governor of the dilapidated, fever-ridden pepper port of Bengkulu on the west coast of Sumatra. Yet it was from Bengkulu, as he watched the Dutch regain possession of the archipelago and enforce a policy of complete commercial monopoly, that he made his next move to extend British influence in southeast Asia.

In a voyage to Calcutta, which all but ended in shipwreck, he employed his wide knowledge of Eastern affairs and his powers of persuasion to convince Lord Hastings, then governor general of India, that immediate and forceful action was essential to safeguard British trade. On December 7, 1818, he sailed from Calcutta, bearing Hastings' qualified authority to establish a fortified post eastward of the Straits of Malacca and so placed as to wedge open the gateway to the China seas. On the morning of January 29, 1819, he landed on the shore of a sparsely populated island off the southern tip of Malaya and, risking imminent collision with the Dutch, established by treaty the port of Singapore. Although he returned to his post at Bengkulu for three years, he went back to Singapore in October 1822, when he reorganized the various branches of the administration. His regulations of January 1823 stated,

Establishment of the free port of Singapore

> the Port of Singapore is a free Port, and the trade thereof is open to ships and vessels of every nation . . . equally and alike to all.

By a treaty of March 17, 1824, the Dutch relinquished all claim to Singapore. For Raffles, however, this was a time of rapidly deteriorating health, characterized by headaches of increasing ferocity, and he sailed for England, arriving there on August 22, 1824. In London his vast collections illustrating natural history and Malayan lore won him acclaim as an Orientalist, and he assisted in founding the London Zoo, of which he was elected the first president. He died of a brain tumour on July 5, 1826.

BIBLIOGRAPHY. C.E. WURTZBURG, ***Raffles of the Eastern Isles*** (1954), is a comprehensive work of great industry, excellently edited by CLIFFORD WITTING, and includes a comprehensive bibliography; MAURICE COLLIS, ***Raffles*** (1966), is a later and briefer work presenting a clear picture of the man behind the facts.

(H.F.P.)

Railroads and Locomotives

A railroad (or railway) is a mode of land transportation in which freight-goods- and passenger-carrying vehicles, or cars, with flanged wheels move over two parallel steel rails. The guideway, or track, consists of the parallel rails laid on crossties, or sleepers, and anchored in a bed of crushed rock or other ballast. The cars usually are pushed or pulled by a locomotive, although they may be self-propelled. The track gauge (the distance between inside faces of the rails) varies from country to country and sometimes among railroads within the same country. The predominant gauge among railroads of the world, however, is the so-called standard gauge, which is four feet 8.5 inches (1.435 metres).

The development of railroads is one of the great landmarks in the progress of civilization. From early in the 19th century, railroads provided an element that helped greatly to realize the potential of the Industrial Revolution in the form of a reliable, low-cost, high-volume system of land transportation.

Basic principle of the railroad

The railroad's basic principle, flanged steel wheels rolling on steel rails, is what gives this mode its unique capability for heavy-duty transportation. The flanges on the insides of the wheels guide the locomotives and cars, causing them to follow the line of the rails; and the rolling friction of the wheels on the rails is extremely low. In fact, if a 40-ton (36,000-kilogram) railroad freight car of a standard United States type were set rolling on level track at 60 miles (about 100 kilometres) per hour, it would travel five miles (8 kilometres) or more before coming to a stop. By contrast, a motor truck of similar weight set free on a level highway at the same speed would roll only about one mile (1,600 metres).

Because of this self-guiding characteristic and the low rolling friction, a locomotive of relatively modest horsepower can pull a long train of cars. This, basically, is the reason for the economy of railroad transportation. A freight train of 5,000 tons (4,500,000 kilograms) gross weight (*i.e.,* weight of cars and freight) can be hauled with a locomotive of about 5,000 horsepower, depending on terrain and desired operatihg speeds, or approximately one horsepower per gross ton. Typical truck tractor-trailer combinations for intercity highway freight service are powered at about ten horsepower per gross ton. The railroad also has roughly the same 10 to 1 advantage in fuel economy and in employee productivity.

Almost from its inception, the railroad became the dominant all-purpose land carrier of both freight and passengers. It still performs that role in many countries, especially those that are in earlier stages of industrialization. The growth of automotive highway, air, pipeline, and modem water transport, however, has resulted in a changed and more specialized role for the railroads of many countries. In the United States, for example, highway and air services have almost completely taken over the transportation of passengers, except in urban areas and on short- to medium-distance (up to 300 miles [500 kilometres]) intercity runs. In freight transportation, highway trucks perform most of the service at distances of less than 300 miles, and pipelines carry most of the petroleum products, as well as other commodities. Yet United States railroads still produce a substantially greater volume of intercity freight transportation than any other mode; in recent years their freight volume consistently set new records as the economy of the country grew. Similar trends are evident in other highly industrialized nations.

This article is divided into the following sections:

A general survey, from a historical viewpoint, of all forms of transport is presented in the article TRANSPORTATION, HISTORY OF. Forms of urban rail transport are described in TRANSPORTATION, URBAN.

I. History

DEVELOPMENTS TO 1850

Origins. The railroad as it is known today originated in England in the first quarter of the 19th century. Much earlier, however, crude "wagonways" were used in mines both in England and on the European continent. A narrow-gauge railroad in the mines at Leberthal, Alsace, was illustrated in ***Cosmographioe universalis*** by Sebastian Miinster in 1550. In his treatise ***De re nzetallica,*** printed at Basel in 1556, Georgius Agricola (Georg Bauer) described and illustrated a small mining wagon that ran on a wooden track and was guided by an iron prong, or pin, running in a narrow gap between the rails.

Early mine railroads

In the mid-16th century, such mining railroads seem to have been quite generally used in the mines of central Europe. One mining wagon, with flanged wooden wheels and track, is in the Verkehrs und Bau Museum, Berlin. It was used in the gold mines of northwestern Transylvania.

Mining railroads were apparently introduced into England from Germany early in the 17th century. The first written mention of a railroad in Britain is in a manuscript account of an estate near Nottingham, in 1603 or 1604. In the early years of the 17th century, numerous lines, some quite lengthy, were built from mining pits to rivers. **A** notable example was the Tanfield Wagon Way, Durham, England, a portion of which was built possibly as early as 1632. About 1726 this line was considerably expanded; one of its branches required a 103-foot (31-metre) stone-arch bridge. The Tanfield line continued in operation as a mineral railroad until 1964. By the end of the 18th century, there was a considerable network of railroads, or wagonways, extending back into the country from both sides of the Tyne and Wear rivers. These mining railroads were essentially privately operated adjuncts to the mining operations. Horses supplied the motive power.

The early mining railroads used two methods of guiding the vehicles along the track. Earliest practice, apparently, was to use vehicles with special flanged wheels, such as are used today, although the flanges could be on either the outside or inside of the wheel tread, running on "edge rails." Later, "plateways" were introduced, iron angles laid end to end; the vertical flanges on these angles served to guide the wagons, which had ordinary flat-tread wheels. The advantage of the plateway was that wagons did not require special wheels, although the distance between wheels (*i.e.*, the wheel gauge) had to fit that of the plates. Thus, the same wagons could use both ordinary roads and plateways. Many plateways were built, mostly between 1795 and 1815 (one survived as a public plateway until 1917). The plateway, or flanged-rail system, eventually gave way to the flanged-wheel system, at least partly because of the difficulty of devising a workable switch, or turnout, for the former.

Plateways for wagons

Mining railroads were used in Wales and Scotland, as well as in England. The first railroad authorized by an act of Parliament was the Middleton to Leeds colliery line of 1758 (a section of this line is still in operation).

In 1801 the Surrey Iron Railway was incorporated. This was the first public freight-carrying railroad; it opened from Wandsworth to Croydon on July 26, 1803.

Early railroad development. The development of mechanical traction to replace horsepower may be said to mark the emergence of the modern railroad. Thereafter, railroads spread rapidly in Britain and then in other parts of the world. But, because of differing conditions, railroads developed quite differently in Britain, on the continent of Europe, and in North America.

Britain. The "New Castle," a locomotive built by the English engineer Richard Trevithick (Figure 1), ran on a Welsh tramroad in 1804 but, like a number of the early steam locomotives, was too heavy for the rails. The first practical and successful locomotive was built in 1812 to the instructions of John Blenkinsop, an inspector at the Middleton colliery near Leeds. It ran on cast-iron rails and had two vertical cylinders driving two shafts geared to a toothed wheel that engaged a rack rail.

In 1813 the English inventor William Hedley built the "Puffing Billy," a simple adhesion locomotive that relied on friction between the wheels and the rails, dispensing with the toothed rack rail (the rack-rail system is now used only on a few mountain railroads with extremely steep gradients). Like John Blenkinsop's locomotives, "Puffing Billy" was used for hauling coal wagons between a mine and wharves, as was George Stephenson's first locomotive, the "Blücher," completed in 1814.

In 1823 Stephenson, one of the great pioneer railroad and locomotive builders, was invited to build and equip a railroad from Stockton to Darlington. The ceremonial opening of this landmark line took place on September 27, 1825. The Stockton and Darlington Railway was the first public railroad in the world to use locomotive traction and the first built to carry both freight and passengers. At first steam locomotives were used only for freight service; passenger service was provided by a contractor who used horse-drawn coaches. The first locomotive on the Stockton and Darlington was George Stephenson's "Locomotion." It and similar locomotives proved unreliable and expensive to maintain. They were suitable only for hauling low-speed mineral trains; their weight and tractive effort were limited by the relatively weak track. At times the railroad reverted to horses, but the situation was improved in 1827 with the introduction of the "Royal George," a six-coupled locomotive designed by Timothy Hackworth.

First public locomotive traction railroad

But the railroad era really began with the opening, on September 15, 1830, of the Liverpool and Manchester

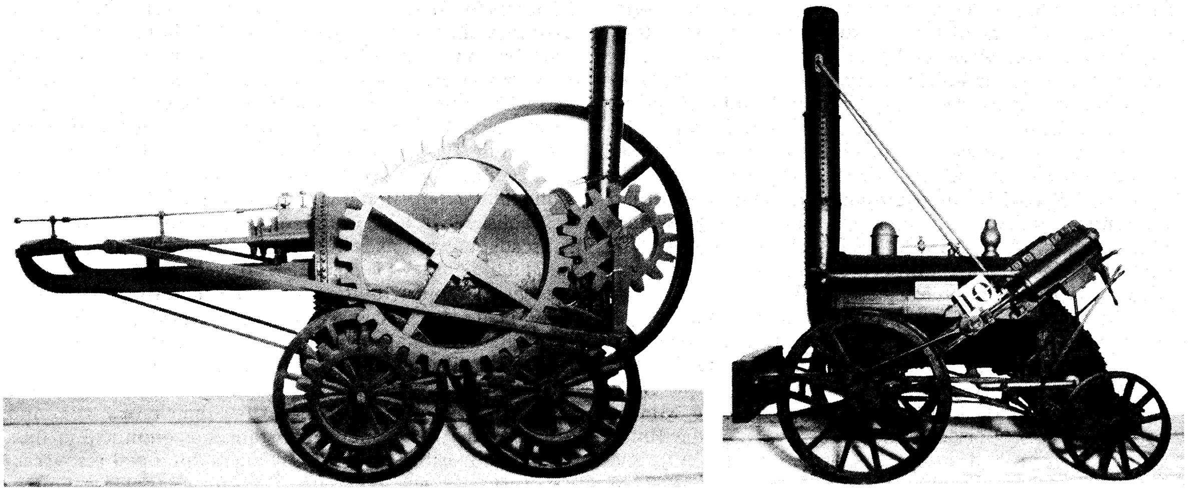


Figure 1. (Left) The "New Castle," built by Richard Trevithick in 1803, the first locomotive to do actual work. (Right) The "Rocket," built by George Stephenson, it won the 1829 trials by the Liverpool and Manchester Railway, England.

By courtesy of the Baltimore and Ohio Railroad Company

Railway. The Liverpool and Manchester incorporated all the features of modern public railroads. It was a public carrier of both passengers and freight. with all business handled directly by the company itself: it used mechanical traction for all traffic.

George Stephenson's "Rocket"

Previous to its opening in 1829, the Liverpool and Manchester held a contest to determine the best type of motive power. The trials took place on the Rainhill level (Lancashire) from October 6 to 14, 1829. Three steam locomotives took part: George Stephenson's "Rocket" (Figure 1), Timothy Hackworth's "Sans Pareil," and the "Novelty," built by John Braithwaite and John Ericsson. On the last day of the trials the "Rocket" was awarded the £500 prize. The "Rocket's" superiority was due mainly to its use of a multiple fire-tube boiler rather than the single-flue boilers previously used. About this time, too, John Birkinshaw developed the fish-bellied (bellying out on the underside), rolled-iron edge rail. This was much stronger than the cast-iron rails previously used and enabled heavier locomotives to be run.

After the Rainhill trials, the steam locomotive became the dominant form of railroad motive power and was to remain so for over a century. The success of the Stockton and Darlington and the Liverpool and Manchester lines touched off widespread railroad building in Great Britain. In 1836 the London and Greenwich Railway opened, bringing the first public passenger service to London. In 1838 the London and Birmingham Railway was opened for its full length, and in 1840 the line from London to Southampton was completed. The Great Western Railway's line from London to Bristol was opened in 1841. By that time there were over 1,300 miles (2,100 kilometres) of rail line in the United Kingdom, and between 1844 and 1846 Parliament authorized the construction of more than 400 new lines, representing the height of the "railway mania" in Great Britain.

Unlike the other early British railroads, which used the four-foot-8.5-inch (1.435-metre) gauge, the Great Western line was built to a seven-foot (2.1-metre) gauge. The directors of the road adopted this broad gauge at the instigation of the line's brilliant chief engineer, Isambard Kingdom Brunel. There were, and still are, sound technical reasons for using a wide gauge; by 1835, however, when the Great Western directors adopted Brunel's seven-foot, it was already too late. The necessity for individual railroads to work together as an integrated network dictated a common gauge, and the 4-foot-8.5-inch width had already become the "standard gauge"; it remains so today. The Great Western continued to use the broad gauge until 1892, when the line was converted to standard gauge.

Continental Europe. Although the beginnings of railroads occurred quite early in most European countries, subsequent extension of rail lines was much slower than in Great Britain. Continental railroad practices soon diverged from those in Britain, though not as sharply as in North America. Standard gauge (or a width so close to it as to be compatible for through running) became dominant in all continental countries except Russia and Finland, where the width is five feet (1.5 metres), and Spain and Portugal, which use five feet six inches (1.680 metres). As compared with British railroads, continental railroads are also built to a much larger loading gauge, or clearance diagram, the dimensions above and outside the rails that rolling stock or loads must not exceed. This means that continental locomotives and cars are much bigger than those in Britain.

There was also more emphasis on state planning and control in the construction of early continental railroads. There, unlike the situation in Britain (and in North America), few competing railroads were built. In some cases the governments themselves built the lines; in other cases private companies received exclusive franchises to serve specific routes.

First railroad in France

The first railroad in France, from Saint-Etienne to Andrtzieux, went into use in 1827. It used horses and carried only freight until 1832, when steam locomotives were adopted. The Saint-Etienne–Lyon line was completed in 1832. The first international line, from Strasbourg to Basel, was completed in 1841, by which time France had 350 miles (about 550 kilometres) of railroads.

The use of the steam locomotive in Germany began on December 7, 1835, with the opening of a railroad between Nürnberg and Fürth. A gauge of six feet (1.8 metres) was selected for the first public railroad in Russia, which was opened, with horse traction, in 1836. The original line in Russian Poland was of standard gauge, but Russia finally standardized on a width of five feet (1.5 metres). By 1850 there were only about 370 miles (600 kilometres) of railroad in all of Russia, but in 1851 the 404-mile (650-kilometre) link between St. Petersburg and Moscow was completed by the American engineer George Washington Whistler.

The first railroad in the Austro-Hungarian Empire, a 90-mile (140-kilometre) horse line between Linz and Budweis, was opened in 1832. By 1837, when the first steam locomotives were introduced, there were some 170 miles (270 kilometres) of horse-traction public railroads in the country.

North America. Interest in railroads in the United States developed almost as soon as in England. One of several horse-drawn tramways built early in the 19th century was Gridley Bryant's Granite Railway in Quincy, Massachusetts. This three-mile (five-kilometre) broad-gauge line carried the granite used in building the Bunker Hill Monument in Boston.

By 1813 the inventor Oliver Evans was proposing a railroad between New York and Philadelphia. Two years later, John Stevens received from the New Jersey legislature the first charter for a railroad ever granted in America. Stevens was ahead of his time: the chartered line, between the Delaware and Raritan rivers, was never built. But in 1825 he built and operated the first locomotive to run on rails in America; it ran on a half-mile circle of track at Stevens' home in Hoboken,·New Jersey.

Success of the Stockton and Darlington in England spurred interest in railroads in the United States. On February 28, 1827, the Baltimore and Ohio Railroad Company was chartered. The line began carrying revenue traffic on January 7, 1830. The first 13 miles (21 kilometres) of line, from Baltimore to Ellicott's Mills (now Ellicott City), opened on May 24, 1830.

Baltimore and Ohio Railroad

The Baltimore and Ohio was the first railroad in the United States to be chartered as a common carrier of freight and passengers. Its promoters, looking beyond local needs, envisaged a line going all the way to the Ohio River to channel the commerce of the growing Middle West through the port of Baltimore. By 1834 the Baltimore and Ohio had built to Harpers Ferry, Virginia (now West Virginia), and on December 24, 1852, it reached the Ohio River at Wheeling. Subsequently, the company expanded, both by new construction and by acquiring other railroads, until it reached Chicago, St. Louis, and the Great Lakes.

Almost simultaneously, several other American railroads came into being. Construction of the five-foot-(1.5-metre-) gauge line from Charleston to Hamburg, South Carolina, by the South Carolina Canal and Rail Road Company began in February 1829. On December 25, 1830, this line became the first in the United States to start scheduled passenger operations using a steam locomotive. When the entire line to Hamburg was completed in 1833, it was the longest (136 miles [219 kilometres]) then operating in the United States. Ultimately, this line became part of the 10,000-mile (16,000-kilometre) Southern Railway System. Several other railroad systems that later grew into giants of the United States railroad industry had their beginnings in the decade of the 1830s. By 1840 there were 2,800 miles (4,500 kilometres) of line in the United States, and the country had entered its first great era of railroad building. Twenty years later, on the eve of the American Civil War, the country had 30,000 route miles (50,000 kilometres) of track.

With few exceptions, early railroads were designed to promote the commercial interests of local communities or areas. As growth progressed, however, many of the small roads were consolidated, forming through routes that served fairly large territories, and new railroad projects became more ambitious. The Pennsylvania Railroad Company (now Penn Central) completed its line from Philadelphia to Pittsburgh in December 1852, using ten inclined planes to climb over the Allegheny Mountains. A little more than a year later, it completed an all-rail route.

The growth of Canadian railroads paralleled somewhat that of the United States. Construction of the first line, the Champlain and St. Lawrence Railroad, between Laprairie and St. John, Quebec, began in 1835; its first train operated on July 21, 1836. By 1860 there were about 2,000 miles (3,000 kilometres) of railroad line in Canada.

Social and economic consequences. The impact of the early railroads upon the societies in which they developed amounted to a social revolution. For the railroad represented the first big leap ahead in man's efforts to reduce the effects of time and distance on his travel and communications. As such, its effect on the society of the early 19th century was as profound in its day as was the effect later on of the development of the automobile, the airplane, and the radio. For the first time, farmers, manufacturers, merchants, and travellers had available a form of overland transportation that was fast, relatively inexpensive, little affected by weather conditions, and capable of moving large volumes of both goods and people.

Railroads compared to older forms of transport

The effect of the railroad upon the older forms of land transportation, the stagecoach and the canalboat, was immediate and often catastrophic. Though diehards opposed the railroads, few could afford not to use them once they were built. Typically, in 1841 a trip from London to Exeter, about 175 miles (280 kilometres), required about 21 hours by stagecoach, including meal stops. By 1846 an express train made the journey in 6½ hours at less cost than by stage. (In 1970 express trains made this trip in two hours.) Newspapers and mail soon began to move by rail with dramatic time savings. On December 11, 1849, to publicize their joint route between London and Paris, the South Eastern and Nord railways of Britain and France dispatched copies of the morning ***Times*** from London at 7:00 AM and had them in Paris by 1:30 PM.

Established freight traffic quickly shifted to the new railroads; but just as railroads produced a vast increase in the total volume of passenger travel, so did they also present new opportunities for manufacturing and trade. They made it unnecessary for some types of factories to be located on canals or rivers; railroads could carry in the raw materials and take the finished products, at low cost, to much more distant markets than previously. Farmers could find a more ready market for their crops; coal and ore could be moved longer distances economically, thus making it feasible to work previously untapped resources.

So pervasive was the railroad's impact that great pressures developed on the part of both industries and communities to be served by rail transportation. Only a few towns and cities opposed the coming of railroads, and often they regretted their obstinacy when they saw growth and prosperity going to neighbouring communities that had welcomed the new form of transport. In both Europe and the United States, community leaders pushed hard to have main lines built through their towns, and if they could not get main-line status, they tried at least to get a branch-line connection.

In the United States, Canada, and other new undeveloped countries, the railroad proved to be the means of opening up new territory. Many towns and cities in the United States and Canada originally came into being as railroad "division points," where locomotives were changed, serviced, and repaired, and where operating crews were changed.

FROM 1850 TO THE PRESENT

Railroad development: 1850–1900. The second half of the 19th century saw the railroad reach maturity and become a worldwide technical, economic, and social phenomenon. In the earlier part of the half century, railroad accidents became frequent and serious, as locomotive and car sizes grew rapidly, overstraining tracks and bridges. The first steel rails were introduced in England in 1857, and the first steel bridge was built, over the Mississippi at St. Louis, in 1867–73. Other major technological advances included automatic electric block signalling, telegraphic train dispatching, automatic coupling, and air brakes. Railroads learned to work together on common standards for brakes, couplers, and wheels, to permit interchange of equipment among different lines.

Britain. As the country where the .railroad originated, Great Britain was the first to experience intensive building. Construction of new lines peaked there during the 1840s, when some 4,500 route miles (7,200 kilometres) were laid down, and remained at a high rate during the next two decades. By 1870 Britain had about 13,500 miles (21,700 kilornetres) of railroad, largely double tracked. In the remaining three decades of the century the pace slowed down, but other important developments took place. Physical plant of the main lines was improved. In Scotland bridges were built over the two great firths of Tay and Forth. The first built over the Tay, on wrought-iron piers, was blown down in a gale (1879) and was replaced in steel. The Forth was bridged by the world's longest span bridge, a giant cantilever (1890). The world's longest underwater tunnel was driven under the Severn for the Great Western Railway (1886). First-class sleeping cars were introduced on trains between London and Scotland in 1873, and dining cars on the Great Northern's London to Leeds line in 1879. On August 22, 1895, competing trains of two lines raced from London to Aberdeen, Scotland, both completing the run at average speeds in excess of 60 miles (100 kilometres) per hour.

Railroad bridges in Scotland

The United States. Railroad construction in the United States was on an even larger scale than in Britain in the

1850s, when more than 21,000 route miles were built. Construction was slowed down by the American Civil War (the first war, incidentally, in which railroads played a major role); but it resumed on a large scale immediately afterward, reaching a peak in the 1880s: in the year 1882 alone 11,500 miles (18,500 kilometres) were built, a record exceeded in 1887 with a total of 12,800 miles (20,600 kilometres). Altogether, more than 70,000 miles were built in the decade. Construction continued at a relatively high level through the 1900–10 decade.

The major thrust of American railroad development was westward. The first locomotive in Chicago, the "Pioneer" of the Galena and Chicago Union Railroad, now the Chicago and North Western Transportation Company, made its first run on October 25, 1848. The first railroad reached the Mississippi from Chicago in 1854. On May 10, 1869, the first transcontinental route was created when, at Promontory, Utah, the Union Pacific Railroad, building west from Omaha, Nebraska, met the Central Pacific Railroad, building east from Sacramento, California. Ultimately there were about nine major routes leading from the Midwest or South to the West Coast.

Government assistance to U.S. railroads

Although the United States railroad system, like that of other countries, was essentially a product of private enterprise, it received important government assistance, in the early years from state and local governments, and after the Civil War from the federal government in the form of land grants. Such grants, totalling about 131,000,000 acres (53,000 hectares) went to 29 railroads for about 18,000 miles of line (about 8 percent of total U.S. mileage). They made it possible for the railroads to push lines across prairies and mountains almost entirely undeveloped and very sparsely settled.

Railroad building becomes worldwide. Construction in western Europe followed that in Britain. In France six principal companies emerged, Nord, Est, Paris-Orléans, Paris-Lyon-Mediterranean, Midi, and Ouest. In 1878 the state took over a group of small companies in western France, creating the *État* (State) system. By 1902 the French rail network totalled 28,400 miles (45,700 kilometres). In Germany both private companies and the individual German states built railroads, but after 1876 the privately owned systems were gradually absorbed by the states. Germany's total route mileage had reached 35,625 (57,321 kilometres) by 1909.

Meanwhile, countries in other parts of Europe and throughout the world were getting their first railroads or expanding their small beginning networks to national proportions. In Canada the Canadian Pacific Railway Company tied Canada together just as the Central and Union Pacific lines had tied the United States together. One of the terms on which British Columbia entered the confederation in 1871 was the promise of this transcontinental route, which was completed in 1885. Two other Canadian transcontinentals, built with government assistance, reached the Pacific Coast in 1914 and 1915. India's first line, the East Indian Railway Company, was registered in 1849; the first line was opened between Bombay and Thana in 1853. By 1910 there were more than 32,000 miles (51,000 kilometres) of line in what are now India, Pakistan, and Bangladesh.

This period also saw the beginnings and rapid growth of railroads in Africa, Australasia, and Japan. Railroad development was handicapped in Australia because each state laid track to the gauge it considered suitable for its own needs; as a result passengers and freight had to be transshipped from one line to another, and the railroad failed to play as strong a role in opening up the country as it did in the United States and Canada.

The Trans-Siberian Railroad

A notable achievement of this period in railroad development was the building by Russia of the Trans-Siberian Railroad, 5,787 miles (9,313 kilometres) long, linking Moscow and Vladivostok. Construction began in 1891, starting simultaneously from the east and west terminals. Originally, passengers could not make the entire journey by rail but had to travel by boat (or in winter, by sleigh) across Lake Baikal. By 1916 a line around the lake allowed the trip to be made without changing cars.

The 20th century. With the 20th century the railroad reached maturity. Railroad building continued on a fairly extensive scale in some parts of the world, notably in Canada, the Soviet Union, and in Africa. But in the more developed countries construction tapered off, except for improvements or refinements of existing systems. The technological emphasis shifted to faster operations, more amenities for passengers, larger and more specialized freight cars, safer and more sophisticated signalling and traffic-control systems, and new types of motive power. Railroads in many of the more advanced countries also found themselves operating in a new climate of intense competition with other forms of transport.

Major new construction. By mid-20th century, railroads in such countries as the United States, Britain, and Ireland, as well as in western Europe generally, were beginning to abandon secondary and branch lines that had become uneconomic. But in the Soviet Union, China, Japan, Australia, and Canada, important new lines were being built. In the Soviet Union the Trans-Siberian line was double tracked and, by 1970, had been largely electrified. At that time, additional construction had raised the Soviet system to a total of 86,691 route miles (139,511 kilometres), making it the largest single railroad system in the world. In China, a large-scale railroad building program was also launched. A recently completed trunk line links Tsining with Ulaanbaatar, the capital of Mongolia, which was already on a branch of the Trans-Siberian Railroad. Several other important Chinese lines have been built or are under construction. Some of the main lines are being double tracked and electrified. Two major bridges across the Yangtze River, one at Hankow and the other at Nanking, have been built; each is more than a mile (1,600 metres) long.

Australia's record straight-line track

In Australia, considerable progress has been made in overcoming the handicap caused by the variety of gauges. The Commonwealth Railways has built several standard-gauge lines, the most important being the transcontinental link between Port Pirie and Kalgoorlie. Crossing the Nullarbor Plain, this line has one stretch of 300 miles (500 kilometres) without a curve, the longest straight track in the world. Standard-gauge rails also now link Perth with Sydney.

As Canada entered a period of rapid economic growth, it built a number of railroad lines to open up new territory. Among them were the Great Slave Lake Railway, running from Peace River, Alberta, 432 miles (695 kilometres) to Hay River and Pine Point, Northwest Territories, and the Quebec North Shore and Labrador Railway Company, a 320-mile (510-kilometre) ore-carrying line from iron ore deposits near Schefferville, Quebec, to the St. Lawrence River at Sept Iles. The Pacific Great Eastern Railway (now the British Columbia Railway), completed from Vancouver, British Columbia, to Prince George in 1956, in 1958 reached Dawson Creek and Fort Saint John, 327 miles (526 kilometres) north of Prince George. In the 1970s it expected to complete a further extension of 250 miles to Fort Nelson, British Columbia, and another line from Fort Saint James to within 140 miles of the Yukon Territory border.

In Japan, a major new railroad was built not to open new country but to enlarge the capacity of rail transportation between Tokyo and Osaka. The New Tōkaidō Line (see below) was built to standard gauge instead of the 3-foot-6-inch (1.07-metre) gauge commonly used in Japan. It has been so successful that a 100-mile (160-kilometre) extension, from Osaka to Okayama, was added in the early 1970s. The Table summarizes the current state of railway systems in selected countries.

Technological development. In the period 1900 to about 1930, few apparent advances were made in railroad technology. Nevertheless, that period saw several developments that were later to have a profound effect on railroads everywhere. One was the first practical and reliable diesel-electric locomotive; another was the perfecting of centralized traffic control, a highly efficient system of controlling train operations; still another was the first use of continuous welded rail, a major contribution to smoother track that costs less to maintain.

Railway Systems of Selected Countries

	service began	ownership[u]	gauge†	mileage (track length)‡
Algeria	1862		standard	3,071
Argentina	1857	state	5 ft 6 in.	15,462
			metre	10,715
			standard	2,210
Australia	1854	state	standard	8,3025
			5 ft 3 in.	5,7771
			3 ft 6 in.	11,1695
Austria	1838	state	standard	6.998
Bangladesh	1861	state	metre	2,517
Belgium	1835	state	standard	7,754
Brazil	1854	state	metre	14,553§
Bulgaria	1866	state	standard	3,5841
Burma	1877	state	metre	2,400
Canada	1836	state	standard	35,147
		private	standard	29,321
Ceylon (Sri Lanka)	1865	state	5 ft 6 in.	1,236
Chile	1851	state	metre	5,844
			5 ft 6 in.	2,349
China	1881	state	standard	23,9001
Czechoslovakia	1839	state	standard	13,226
Denmark	1847	state	standard	3,395
East Africa (Kenya, Uganda, Tanzania)	1897	state	metre	4,370
Egypt	1854	state	standard	4,149
Finland	1862	state	5 ft	5,462
France	1828	state	standard	49,028
Germany (East)	1835	state	standard	18,179
Germany (West)	1835	state	standard	44,999
Greece	1869	state	standard	2,002
Hungary	1846	state	standard	8,442
India	1853	state	5 ft 6 in.	36,353
			metre	21,770
			2 ft 6 in.; 2 ft	3,513
Indonesia	1864	state	3 ft 6 in.	4,852
Iran	1917	state	standard	2,581§
Ireland	1834	state	5 ft 3 in.	2,282
Italy	1839	state	standard	20,649
Japan	1872	state	3 ft 6 in.	25,426
		private	3 ft 6 in.	5,733
Korea (South)	1899	state	standard	3,378
Malaysia	1885	state	metre	1,428
Mexico	1850	state	standard	30,613
Morocco	1911	state	standard	1.632
Mozambique	1886	state	3 ft 6 in.	2,551
The Netherlands	1839	state	standard	4,568
New Zealand	1863	state	3 ft 6 in.	4.499
Nigeria	1901	state	3 ft 6 in.	2,680
Norway	1854	state	standard	3,452
Pakistan	1861	state	5 ft 6 in.	7,664
Poland	1845	state	standard	16,540§
Portugal	1856	state	5 ft 6 in.	2,944
Rhodesia	1897	state	3 ft 6 in.	2,040
Romania	1869	state	standard	6,8385
South Africa	1860	state	3 ft 6 in.	19,648
Spain	1848	state	5 ft 6 in.	13,792
The Sudan	1900	state	3 ft 6 in.	3,379
Sweden	1856	state	standard	11.611
Switzerland	1844	state	standard	4,249
		private	metre	1,529
Taiwan	1891	state	3 ft 6 in.	1,061
Thailand	1893	state	metre	2,733
Tunisia	1876	state	metre	1,743
Turkey	1856	state	standard	5,874
Soviet Union	1837	state	5 ft	86,691§
United Kingdom	1825	state	standard	34,471
United States	1830	private	standard	305.1 79
Yugoslavia	1846	state	standard	9,433
Zaire	1911		3 ft 6 in.	3,702
Zambia	1905	state	3 ft 6 in.	1,000

'Predominant ownership is given; most countries have both private and state owned lines. †Gauge of the principal mileage; most countries have lines of several gauges: standard gauge ≃ 4 ft 8½ in.: metre gauge = 3 ft 3⅜ in. $Data for early 1970s. §Route length.
Source: ***Jane's World Railways*** 1971–72.

After World War II these and other technological improvements began to take hold on many of the world's railroads. The diesel-electric locomotive completely superseded the steam locomotive in the United States in little more than a decade. The vastly improved efficiency of this type of locomotive was' a major reason why railroads in that country were able to continue under private operation.

Another important trend became evident in the 1950s and 1960s: the electrification of major trunk lines in many countries. Electrification, though it requires a very heavy capital investment, is profitable wherever the traffic is extremely dense, as in Japan, parts of the Soviet Union, western Europe, and Great Britain.

II. Modem railroad technology

LOCOMOTIVES

Modern locomotive power sources

In normal practice a locomotive is a separate unit incorporating the machinery to generate (or, in the case of an electric locomotive, to convert) power and transmit it to the driving wheels. Motive power, however, can be incorporated into a car that also has passenger, baggage, or freight accommodations. Today there are three main sources of power for a locomotive: steam, oil, and electricity. Steam, the earliest form of propulsion, was in almost universal use until about the time of World War II; since then it has been largely superseded by the more efficientdiesel and electric traction.

The steam locomotive is a self-sufficient unit, carrying its own water supply for generating the steam and coal, oil, or wood for heating the boiler. The diesel locomotive also carries its own fuel supply, but the diesel-engine output cannot be coupled directly to the wheels; instead, a mechanical, electric, or hydraulic transmission must be used. The electric locomotive normally is not self-sufficient; it picks up current from an overhead trolley wire or a third rail beside the running rails. A few small battery-driven locomotives or railcars are used for special tasks.

The only other category of locomotive in use today is the turbine locomotive; although atomic-powered locomotives have been proposed, none has actually been built.

Steam locomotives. The basic features that made George Stephenson's "Rocket" of 1829 successful—its multitube boiler and its system of exhausting the steam and creating a draft in its firebox—continued to be used in the steam locomotive (Figure 2). The number of coupled drive wheels soon increased. The "Rocket" had only a single pair of driving wheels, but four coupled wheels soon became common, and eventually some locomotives were built with as many as 14 coupled drivers. Smaller "pilot" wheels are often used ahead of the drivers. Their main function is to guide the coupled wheels around curves. Large locomotives may also have smaller wheels behind the drivers. These permit the locomotive to carry a larger, heavier firebox.

From R. Ziel, *The Twilight* of *Steam Locomotives*, Grosset & Dunlap, Inc. Copyright © 1963, 1970 by Ron Ziel

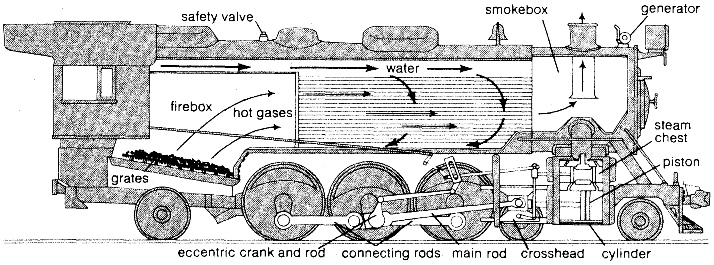


Figure 2: Cutaway diagram of steam locomotive.

Driving wheels

Steam-locomotive driving wheels are of various sizes, usually larger for the faster passenger engines. In Europe, the average was about 66- to 78-inch (1.7- to two-metre) diameter for passenger engines and 54 to 66 inches (1.4 to 1.7 metres) for freight or mixed-traffic types. Typical locomotives built in the United States just before the end of the steam era had drivers ranging from 60 to 84 inches (1.5 to 2.1 metres) in diameter.

Supplies of fuel (usually coal but sometimes oil) and water can be carried on the locomotive frame itself (in which case it is called a tank engine) or can be carried in a separate vehicle, the tender, coupled to the locomotive. The tender of a typical European main-line locomotive had a capacity of ten tons (9,000 kilograms) of coal and 8,000 gallons (30,000 litres) of water. In the Soviet Union and on some African, Asian, and Australian systems, higher capacities were common. The tender of a typical large United States steam locomotive of the World War II period was carried on 14 wheels and had a capacity of 28 tons (25,000 kilograms) of coal and 25,000 gallons (95,000 litres) of water.

In the United States the steam locomotive quickly evolved to the general type that became known as the

American Standard. It had a horizontal fire-tube boiler, a four-wheel pilot truck, and four coupled driving wheels. This type, conventionally designated 4-4-0 to reflect the number of wheels, was used for all kinds of services; it dominated American railroading until well after the Civil War period.

To meet the special needs of heavy freight traffic, other types of locomotives with more (and usually smaller) drive wheels were developed. Among them were the Mogul (2-6-o), the Consolidation (2-8-o), and the Mikado (2-8-2). The Union Pacific used a three-cylinder 4-12-2, which had 12 coupled driving wheels. Still greater tractive effort was obtained by using two separate engine units under a common boiler. The front engine was articulated, or hinge-connected to the frame of the rear engine, so that the very large locomotive could negotiate curves. The articulated locomotive was originally a Swiss invention, with the first built in 1888. The largest ever built was the Union Pacific's "Big Boy," used in mountain freight service in the western United States (Figure 3).

The articulated locomotive

By courtesy of the Unon Pacific Railroad

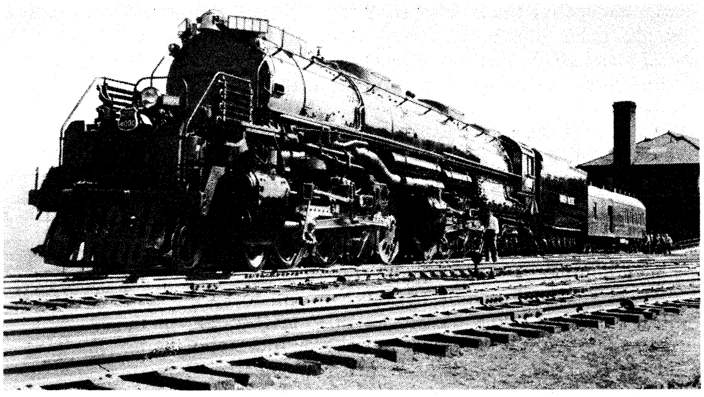


Figure 3: The "Big Boy" of the Union Pacific Railroad, built in 1941. Considered to be the largest steam locomotive ever built, it was used for hauling freight in mountainous areas.

"Big Boy" weighed nearly 600 tons (540,000 kilograms). It could exert 135,400 pounds (61,400 kilograms) of tractive force and developed more than 6,000 horsepower at 75 miles (120 kilometres) per hour.

One of the best known articulated designs is the Beyer-Garrett, which has two frames, each having its own driving wheels and cylinders, surmounted by water tanks. Separating the two chassis is another frame carrying the boiler, cab, and fuel supply. This type of locomotive is valuable on lightly laid track; it can also negotiate sharp curves. It is widely used in Africa.

In simple expansion locomotives (such as "Big Boy"), the steam from the boiler, after being used simultaneously in two (or more) cylinders, is exhausted to the atmosphere. With compound expansion, the steam goes first to one or two small high-pressure cylinders and then to two large low-pressure cylinders before being exhausted. This results in greater thermal efficiency, but with the disadvantage of higher maintenance costs.

Simple expansion locomotives with more than two cylinders were used in a number of countries; three- and four-cylinder locomotives were common in Britain. Compound locomotives were most popular in France, where for many years most locomotive designs were of this type.

Various further refinements gradually improved the reciprocating steam locomotive. Some included higher boiler pressures (up to 290–300 pounds per square inch for some of the last locomotives, compared with around 200 for earliei designs); superheating, feed-water preheating, roller bearings, and the use of poppet (perpendicular) valves rather than sliding piston valves.

Low efficiency of steam locomotives

Still, the thermal efficiency of even the best modern steam locomotives seldom exceeded about 6 percent. Incomplete combustion, heat losses from the firebox, stack, boiler, and cylinders, and other losses dissipated most of the energy of the fuel burned. For this reason the steam locomotive became obsolete, but only slowly, because it had compensating advantages, notably its simplicity and ability to withstand abuse.

By 1971 steam locomotives were virtually extinct in North America, Britain, and western Europe. They were still used to a declining extent in Africa, Asia, and Australasia. More than half the motive power in India (and in some smaller nations) was still steam, however.

Electric locomotives. Efforts to propel railroad vehicles using batteries date from 1835, but the first successful application of electric traction was in 1879, when an electric locomotive ran at an exhibition in Berlin. The first commercial applications of electric traction were for suburban or metropolitan railroads. One of the earliest came in 1895, when the Baltimore and Ohio electrified a stretch of track in Baltimore to avoid smoke and noise problems in a tunnel. One of the first countries to use electric traction for main-line operations was Italy, where a system was inaugurated as early as 1902.

By World War I a number of electrified lines were operating both in Europe and in the United States. Major electrification programs were undertaken after that war in such countries as Sweden, Switzerland, Norway, Germany, and Austria. By the end of the 1920s nearly every European country had at least a small percentage of electrified track. Electric traction was also introduced in Australia (1919), New Zealand (1923), India (1925), Indonesia (1925), and South Africa (1926). A number of metropolitan terminals and suburban services were electrified between 1900 and 1938 in the United States, and there were a few main-line electrifications. The advent of the diesel locomotive inhibited further electrification in the United States after 1938, but following World War II electrification was rapidly extended in Europe; in Switzerland 99 percent of the lines were electrified. There was also expansion of electrified operations in Africa, Asia, and Australasia. By 1972 electrified lines comprised a significant percentage of total route miles in such countries as Sweden (62 percent), Norway (57 percent), Italy (59 percent), France (26 percent), West Germany (33 percent), Soviet Union (22 percent), Japan (45 percent), and Great Britain (19 percent). By contrast, electrified route mileage in the United States was less than 1 percent.

Advantages and disadvantages. Electric traction is generally considered the most economical and efficient means of operating a railroad, provided that cheap electricity is available and that the traffic density justifies the heavy capital cost. Being simply power-converting, rather than power-generating. devices, electric locomotives have several considerable advantages. They can draw on the resources of the central power plant to develop power greatly in excess of their nominal ratings to start a heavy rrain or to surmount a steep grade at high speed. A typical modern electric locomotive rated at 4,000 horsepower has been observed to develop as much as 10,000 horsepower for a short period under these conditions. Moreover, electric locomotives are quieter in operation than other types and produce no smoke or fumes. Electric locomotives require little time in the shop for maintenance, their maintenance costs are low, and they have a longer life than diesels.

Extra power for starting and grades

The greatest drawbacks to electrified operation are the high capital investment and maintenance cost of the fixed plant—the trolley wires and structures and power substations—and the costly changes that may be required in signalling systems to make them compatible with electrified operation. A less important disadvantage is the lack of flexibility of electric locomotives, which cannot operate where there are no trolley wires.

Types of traction systems. Electric-traction systems can be broadly divided into those using alternating current and those using direct current. With direct current, the most popular line (*i.e.*, trolley or third-rail) voltages are 1,500 and 3,000, although there is a large mileage of 600 volts in southern England and several systems in the 600- to 700-volt range around New York City. The disadvantages of direct current are that expensive substations are required at frequent intervals, and the overhead wire or third rail must be relatively large and heavy.

The low-voltage, series-wound, direct-current motor is well suited to railroad traction, being simple to construct and easy to control. It was on a line electrified at 1,500 volts direct current that in 1955 two different French electric locomotives (Figure 4) achieved a speed of 205 miles (330 kilometres) per hour.

By courtesy of the French Embassy, Press and Information Division

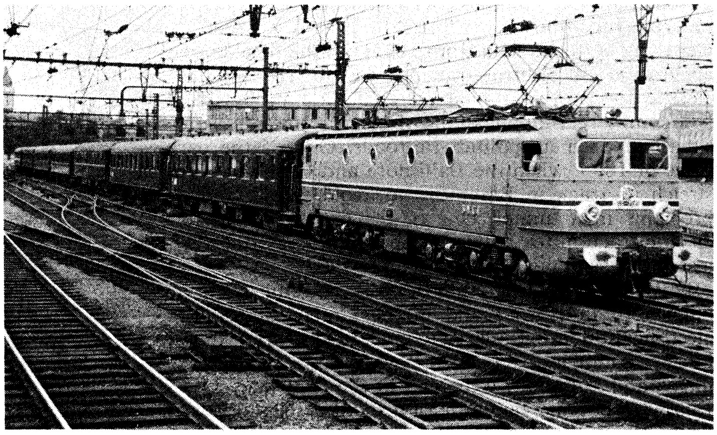


Figure 4: French C-C type electric locomotive, which operates on 1,500 volts direct current. This is one of the locomotives that set a world speed record of 205 miles (330 kilometres) per hour in 1955.

The potential advantages of using alternating instead of direct current prompted early experiments and applications of this system. With alternating current, especially with relatively high trolley-wire voltages (10,000 volts or above), fewer substations are required, and the special equipment needed to produce direct current for the locomotives is eliminated. Available alternating-current motors, however, were not suitable for operation with alternating current of the standard commercial or industrial frequencies (50 hertz [cycles per second] in Europe; 60 hertz in the United States). It was necessary to use a lower frequency (16⅔ hertz is common in Europe; 25 hertz in the United States); this in turn required either special railroad power plants to generate alternating current at the required frequency or frequency-conversion equipment to change the available commercial frequency into the railroad frequency.

Nevertheless, alternating-current trolley wires at 16% hertz are widely used on European railroads, primarily those where electrification began before World War II. Several main-line electrifications in the eastern United States were built using 25-hertz alternating current (these comprise the present Penn Central and Reading railroad systems).

Interest in using commercial-frequency alternating current on the trolley wire continued, however; and in 1933 experiments were carried out in both Hungary and Germany. The German State Railways electrified its Höllenthal branch with a trolley wire at 20,000 volts, 50 hertz.

Electrification on the French railroads

In 1945 Louis Armand, former president of the French railroads, set up a commission to study the 50-hertz operation on the Höllenthal line. As a result, the French railroads went ahead with further development of this system and converted a line between Aix-Les-Bains and La Roche-sur-Foron for the first practical experiments. This was so successful that a network of lines in northeastern France was electrified using 25,000-volt, 50-hertz alternating current.

Subsequently, the 25,000-volt, 50- or 60-hertz system has become virtually the standard for new electrification systems; it is used in a number of countries, including Britain, Turkey, Portugal, the Soviet Union, India, China, Japan, and Argentina. In 1971 several major North American railroads were studying the feasibility of electrifying key main-line segments using alternating current at 50,000 volts, 60 hertz.

With commercial-frequency, alternating-current systems, there are three ways of taking power to the locomotive driving wheels: (1) by a rotary converter or static rectifier on the locomotive to convert the alternating-current supply into direct current at low voltage to drive standard direct-current traction motors; (2) by a converter to produce variable-frequency current to drive alternating-current motors; (3) by direct use of alternating-current traction motors. The first method, using silicon rectifiers or silicon-control rectifiers (thyristers), is by far the most satisfactory. It has the advantage that the locomotive designer, if he wishes, can use the same standard direct-current traction motors that are widely used in diesel-electric locomotives.

In Europe, the problem of international operation of electric locomotives is complicated by the variety of electrification systems in use. The solution is to use locomotives designed to operate on several different voltages or frequencies or both. In the United States, equipment of the former New Haven Railroad (now Penn Central) operating into Grand Central Terminal, New York City, is designed to run on two systems: 660-volt direct current and 11,000-volt, 25-hertz alternating current.

Diesel-electric locomotives. By the end of the 1960s, the diesel-electric locomotive (or simply the diesel) had almost completely superseded the steam locomotive as the standard railroad motive power in most parts of the world, except on the electrified lines. The change came first and most quickly in North America, where, during the 25 years 1935–60 (and especially 1951–60), railroads in the United States completely replaced their steam locomotives.

What caused the diesel to supersede the steam locomotive so rapidly was the pressure of competition from other modes of transport and the continuing rise in wage costs, which forced the railroads to improve their services and adopt every possible measure to increase operating efficiency. Compared with the steam locomotive, the diesel has a number of major advantages:

1. It can operate for long periods with no lost time for maintenance; thus, the diesel can operate through on a run of 2,000 miles (3,000 kilometres) or more and then, after brief servicing, start the return trip. Steam locomotives require extensive servicing after only a few hours' operation.

2. It uses less fuel energy than a steam locomotive, for its thermal efficiency is about four times as great.

3. It can accelerate a train more rapidly and operate at higher sustained speeds with less damage to the track.

In addition, the diesel is superior to the steam locomotive because of its smoother acceleration, greater cleanliness, standardized repair parts, and operating flexibility (a number of diesel units can be combined and run by one man under multiple-unit control). With diesels, too, there is no problem of supplying large quantities of boiler feed water; there is no loss of power capability in cold weather; and there is less standby cost, since the locomotive can be completely shut down when not in use, whereas a steam locomotive must have a fire under its boiler even when on standby at the roundhouse.

Drawbacks of the diesel

The diesel-electric locomotive is, essentially, an electric locomotive that carries its own power plant. Its use, therefore, brings to a railroad many of the advantages of electrification, but without the capital cost of the power distribution and feed-wire system. As compared with an electric locomotive, however, the diesel-electric has two important drawbacks: (1) because it is a more complex mechanism than the electric, it costs more to buy and maintain and (2) since its output is essentially limited to that of its diesel engine, it can develop less horsepower per locomotive unit. Since high horsepower is required for high-speed operation, the diesel is, therefore, less desirable than the electric for high-speed passenger services and very fast freight operations.

Diesel development. Experiments with diesel-engine locomotives and railcars began almost as soon as the diesel engine was patented by the German engineer Rudolf Diesel in 1892. Attempts at building practical locomotives and railcars (for branch-line passenger runs) continued through the 1920s. The first successful diesel switch engine went into service in 1925; "road" locomotives were

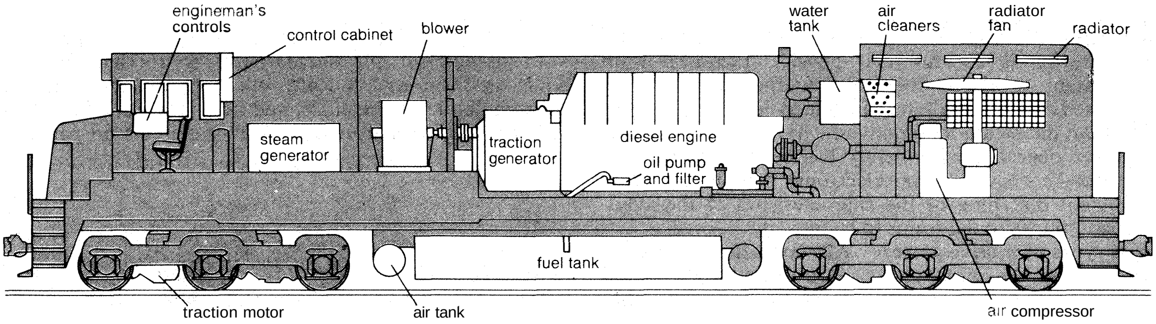


Figure 5: Cutaway of a modern diesel-electric locomotive.

delivered to the Canadian National and New York Central railroads in 1928. The first really striking results with diesel traction were obtained in Germany in 1932. There, a two-car, streamlined, diesel-electric train, with two 400-horsepower engines, began running between Berlin and Hamburg on a schedule that averaged 77 miles (124 kilometres) per hour. Diesel-electrics soon appeared elsewhere, notably in the United States.

The next step was to build a separate diesel-electric locomotive unit that could haul any train. In 1935 one such unit was delivered to the Baltimore and Ohio and two to the Santa Fe Railway Company. These were passenger units; the first road freight locomotive, a four-unit, 5,400-horsepower Electro-Motive Division, General Motors Corporation demonstrator, was not built until 1939.

By the end of World War II, the diesel locomotive had become a proven, standardized type of motive power, and it rapidly began to supersede the steam locomotive in North America. In the United States a fleet of 27,000 diesel locomotives proved fully capable of performing more transportation work than the 40,000 steam locomotives they replaced.

After World War II, the use of diesel traction greatly increased all over the world, though the pace of conversion was generally slower than in the United States. In Britain, steam-locomotive operation ended in 1968; by 1971 steam power had been largely superseded in most developed countries by diesels or by a combination of diesels and electrification.

Elements of the diesel locomotive. Although the diesel engine (Figure 5) has been vastly improved in power and performance, the basic principles remain the same: drawing air into the cylinder, compressing it so that its temperature is raised, and then injecting a small quantity of oil into the cylinder. The oil ignites without a spark because of the high temperature. The diesel engine may operate on the two-stroke or four-stroke cycle and may have cylinders arranged in line, in V-formation, horizontally opposed, or vertically opposed. Rated operating speeds vary from 350 to 2,000 revolutions per minute and rated output may be from ten to 3,600 horsepower. Railroads in the United States use engines in the 1,000-revolutions-per-minute range; in Europe and elsewhere, more compact but higher speed engines are common (see DIESEL ENGINE).

In the United States, early road units and most yard switching engines are equipped with diesels ranging from 600 to 1,500 horsepower. Road units commonly have engines ranging from 2,000 to 3,600 horsepower. Most builders use V-type engines, although in-line types are used on smaller locomotives (up to 1,200 horsepower).

Traction electrical system

With the nearly universal electric transmission, the diesel engine is directly connected to a main direct current generator that converts the mechanical energy produced by the engine into electrical energy. Through the appropriate control equipment, this is then used to drive the traction motors. The traction electrical system operates at a nominal 600-volts direct current, but the voltage varies greatly under operating conditions. The traction motors are of the series-wound type, each geared to the axle it drives (see ELECTRIC MOTOR). Most locomotives have a traction motor on each axle, although some passenger locomotives or units for light branch-line service may have six-wheel trucks (bogies) with the centre axle an idler.

In many recently built diesel-electric units, the engine drives an alternator (producing alternating current) instead of a direct-current generator. Static rectifiers convert the resulting alternating current to direct current for the traction motors. The reason for this design is that an alternator can produce more power and is less costly to maintain than an equivalent direct-current machine.

Other types of transmissions are also used in diesel locomotives. The hydraulic transmission has become quite popular in Germany. It employs a centrifugal pump or impeller driving a turbine in a chamber filled with oil or a similar fluid. The pump, driven by the diesel engine, converts the engine power to kinetic energy in the oil impinging on the turbine blades. The faster the blades move, the less the relative impinging speed of the oil and the faster the locomotive moves.

Mechanical transmission is the simplest type; it is mainly used in low-power locomotives. Basically it is a clutch and gearbox similar to those used in automobiles. A hydraulic coupling, in some cases, is used in place of a friction clutch.

Types of diesel motive power. There are four broad classes of railroad equipment that use diesel engines as prime movers:

1. The light railcar or rail bus (up to 180 horsepower) usually is four-wheeled and has mechanical transmission. It is often powered by a standard highway bus engine; it may be designed to haul a light trailer car.

2. The low-horsepower railcar (up to 1,000 horsepower) may have either mechanical or hydraulic transmission. Most railcars can haul additional trailer cars; some are designed mainly for this purpose, although they also have passenger and baggage accommodations.

3. Train sets (500 to 2,000 horsepower) are formations of more than one vehicle, usually designed to be worked from a single set of controls. The sets include one or more powered vehicles and usually have hydraulic or electric transmission.

4. Locomotives (ten to 6,600 horsepower) may have mechanical, hydraulic, or electric transmission, depending on power output and purpose. They are frequently designed to work in multiple-unit formations. Lower powered units (up to 600 horsepower) are usually designed for switching and light freight service. Medium-powered locomotives (600–1,200 horsepower) are normally used for freight, passenger, or heavy switching duties. Locomotives over 1,200 horsepower are required in main-line service in Europe and North America. In Europe locomotives in the 1,500- to 2,000-horsepower range are the most popular for express passenger work. In North America road freight units may range as high as 6,600 horsepower, and several units totalling as much as 15,000 horsepower may be used in multiple on heavy, fast trains.

Operating methods. Multiple-unit control and the operation of multiple diesel units in a train are almost universal in North America hut much less common elsewhere. A number of railroads in the United States and Canada use "slave" locomotives; these are spotted in the middle of a long freight train and controlled automatically by radio from the locomotive cab at the front end of the train. Radio-controlled slave locomotives permit easier and more efficient handling of a very long freight train, up to as many as 250 cars in regular operations.

The "slave" locomotive

To take advantage of the diesel's special characteristics, many railroads found it necessary to change both their operating techniques and their physical facilities. With diesels, many local and regional roundhouses and shops were closed, since diesels require much less frequent servicing than did steam locomotives. The scheduling of motive-power operations frequently has been centralized at the railroad's headquarters; electronic computers aid in the efficient matching of the characteristics of the available locomotives to the trains to be moved.

Other types of motive power. Although the electric locomotive and the diesel-electric are virtually the world standards, railroads continue to experiment with other types of motive power. In 1969 high-speed, passenger-train sets powered by aircraft-type turbines were placed in service between Montreal and Toronto and between New York and Boston. Though these "TurboTrains" are still in the development stage, the turbines themselves have performed well. These trains have a power unit at each end and use mechanical transmissions.

An experimental turbine-electric car built for the Long Island Rail Road in the U.S. proved sufficiently promising that additional units were planned. They are to be "dual powered" and operate from the third-rail electrification near New York City and under turbine power when outside the electrified zone.

The gas-turbine locomotive

In 1941 a gas-turbine locomotive was developed in Switzerland, and in the 1950s gas-turbine propulsion was tried in Britain and the United States. In France a two-car, gas-turbine-powered passenger train was tested at speeds up to 142 miles (228 kilometres) per hour and introduced in regular service on certain routes in 1970. British Rail was also studying gas turbines for passenger trains to operate at 150 miles (240 kilometres) per hour. The German Federal Railway used a gas turbine as a "booster" in one class of diesel-hydraulic locomotive.

Several attempts have been made to adapt the steam turbine to railroad traction. One of the first such experiments was a Swedish locomotive built in 1921. Other prototypes followed in Europe and the United States. They all functioned, but they made their appearance too late to compete against the diesel and electrification.

CARS

After the first crude beginnings, railroad-car design took divergent courses in North America and Europe, partly because of differing economic conditions and partly because of differing technological developments. Early cars on both continents were largely of two-axle design; but passenger-car builders soon began constructing cars with three and then four axles, the latter arranged in two four-wheel swivel trucks, or bogies. The trucks result in smoother riding qualities and also spread the weight of heavy vehicles over more axles.

Freight cars. European freight cars generally have two axles (four wheels), although there is a trend to two-truck, four-axle vehicles. Conventional British freight cars (goods wagons) have an average capacity of only about 13 tons, though many newer British cars are larger. On the Continent, the capacity of standard freight wagons varies from about 18 to 33 tons. The Soviet railroads use much larger, two-truck cars of up to 100 tons capacity. There is a growing trend toward higher capacity, four-axle cars in Britain and on the Continent, some capable of carrying loads up to 90 or 100 tons.

Standard American freight cars are all of two-truck design, usually with four axles though occasionally with six axles arranged in two three-axle trucks. American freight cars carry anywhere from 50 to 125 tons (45,000 to 113,000 kilograms); most new cars designed for merchandise freight have a capacity of 50 or 60 tons (45,000 to 54,000 kilograms), while new cars designed to carry bulk commodities, such as coal, ore, or chemicals, are mostly rated at 100 tons (90,000 kilograms).

There are three basic types of freight car: the open-top car, the boxcar, or house car, and the flatcar. Car builders and railroads have developed many varieties of these three basic types, plus many special types designed for efficiently carrying specific commodities.

Specialized freight cars

Among the special types of freight cars are double-deck cars for transporting automobiles (Figure 6), open cars with sliding roofs, pressurized tank cars for handling dry solids in bulk, hydraulically operating tipping cars for dumping bulk commodities, and boxcars with wide doors for easy loading and unloading of shipments on pallets.

By courtesy of the Museum of Railway Traffic. Tokyo

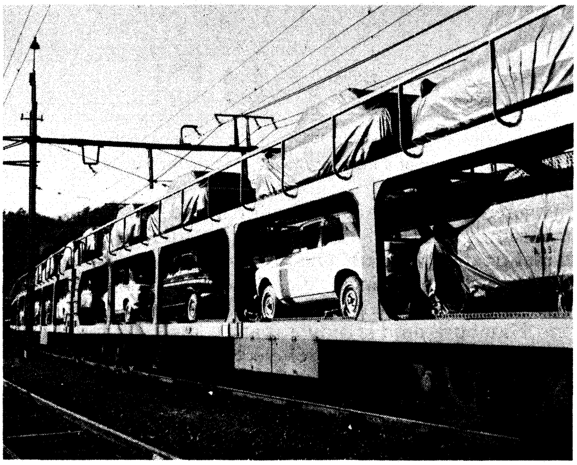


Figure 6: Japanese double-deck car, especially designed for carrying automobiles; it can load up to 12 compact cars.

Because of vertical-clearance limitations, highway trailers cannot be carried piggyback on conventionally designed flatcars on most European railroads. Ingenious special designs have overcome this problem. In France, for example, trailers are positioned on "kangaroo cars," which have pockets into which the trailer wheels are positioned, lowering the overall height. In West Germany highway trailers with their tractors are carried on certain fast intercity runs by special flatcars having small wheels to reduce the overall height.

On American railroads, in addition to the standard boxcar, the flatcar, and open-top "gondolas" and hopper cars (the latter equipped with doors for bottom unloading), the commonest specialized cars are tank cars, refrigerator cars, and livestock cars.

Automobile-parts cars

A growing trend is that of equipping cars with special internal fittings to accommodate specific products without the need for dunnage (*i.e.*, bracing or blocking of the load). Typical of these are the automobile-parts cars, which are often designed to carry certain automobile components, such as bodies, frames, or engines, of a single manufacturer. Among other special types of cars are those with "long travel" cushioned draft gear that absorb jolts and prevent damage to fragile commodities; extra-long flatcars with two- or three-level automobile racks; piggyback and container flatcars; 100- or 125-ton (90,000- or 113,000-kilogram) covered hopper cars for dry-bulk commodities; 100-ton gondolas for unit-train coal service; and 10,000-cubic-foot (280-cubic-metre) boxcars for light but bulky automobile parts.

The growing variety of specialized freight cars has complicated the problem of keeping each railroad shipper supplied with cars of the proper type. On the other hand, specialized cars usually run more miles and earn more revenue than standard cars. In the United States, so-called private car companies have played an increasing role. They acquire the cars and lease them to shippers or to the railroads themselves, a practice long followed with tank and refrigerator cars.

Passenger cars. The first passenger cars were simply road coaches with flanged wheels. Almost from the beginning, railroads in the United States began to use longer, eight-wheel cars riding on two four-wheel trucks, or bogies. In Britain and Europe, however, cars with more than six wheels were not introduced until the 1870s. Most cars of modern design, for both local and long-distance service, have an entrance at one or both ends of the car (some commuter-service cars also have additional centre doors), with centre aisle or a side corridor running the full length of the car. Flexible connections between cars give passengers access to any car of a moving train. In the United States modern passenger cars are usually 85 feet (25 metres) long and weigh from 60 to 80 tons (54,000 to 73,000 kilograms) or more. British and continental European cars are usually shorter and are always of lighter construction than cars in the United States. The standard British coach is 67 feet (20 metres) long; a modern, lightweight coach of the Swiss Federal Railways is 73 feet (22 metres) long. Recent continental designs, however, are up to 86½ feet (26 metres) in length and weigh 50 to 60 tons (45,000 to 54,000 kilograms). Narrow-gauge railroads impose weight and length restrictions. But even so, a 3.5-foot- (1.07-metre-) gauge coach of the South African Railways may be up to 63 feet five inches (19.3 metres) long and weigh up to about 48 tons (about 44,000 kilograms). Recent Australian cars follow American design and appearance.

Coaches. The interior arrangements and amenities on passenger cars vary widely. The most common type of car is the coach. In Europe the most favoured design has six- or eight-seat compartments, with the aisle, or corridor, along one side of the car. In the United States, two rows of double seats on either side of an aisle have long been standard; this design was appearing more frequently in Europe in the 1970s. Some American cars for local commuter runs have one row of triple seats and one row of double seats, with the centre aisle, permitting a 25 percent increase in the number of seated passengers. Several United States and Canadian railroads also use, for commuter runs, cars having an upper "gallery" in addition to the standard seating. These two-level coaches may seat up to 156 passengers. Luxury coaches in many countries have a small number of individual seats.

Sleeping cars. A crude sleeping car was operated in the United States as early as 1837; the first Pullman sleeper, named after its inventor and leased by the railroads, went into service in 1859. Sleeping cars were introduced in Europe and elsewhere beginning in the 1870s. Typical European sleeping cars have compartments with one, two, or more transverse beds (the trend is to more cars with single-berth compartments to meet the needs of business travellers). A typical modern American sleeping car has six bedrooms, each with two beds, and 12 "roomettes," with a single bed, giving a total capacity of 24 persons. Because the low capacity makes high fares necessary and so discourages patronage, the so-called slumbercoach was invented, with eight small double rooms and 24 single rooms, or a total capacity of 40.

"Feature" cars. Dining, or restaurant, cars became a feature of long-distance passenger trains almost from the beginning. A typical modern dining car has a kitchen at one end and tables seating 32 to 40 persons at the other. Such a car requires a large staff of waiters and kitchen personnel; thus, in recent years there has been a trend to "coffee-shop," or "lunch-counter," cars that provide limited, quick service and to food-service cars in which the passenger obtains his own food from vending machines.

The “coffee-shop" car

Another type of car often seen on long-haul trains is the lounge, or observation, car. It has a bar for beverage service and comfortable chairs for relaxation or reading. Dome cars, developed in the United States after World War II and popular on railroads having scenic routes, are also found in Canada and Europe. In a dome car passengers ride under a raised, glassed-in roof section that affords a wide-range view of the countryside.

Special types of cars. Besides the conventional freight and passenger cars already described, the world's railroads operate many special types of rolling equipment. Most numerous are the powered passenger cars and train sets mentioned previously. These may be operated by electricity on electrified railroads or have diesel engines or turbine engines mounted under the car floors. Such cars are usually designed to operate in multiple under the control of one driver and are used extensively for commuter service. The high-speed Metroliner trains (see below) operating on the electrified line between New York and Washington are of this type. Single or multiple diesel-powered passenger cars are used on light traffic lines.

Highly specialized, too, are a small number of freight cars designed to carry unusually heavy or oversized loads, such as large electrical transformers. One such car (Figure 7) carries a transformer suspended between two sets of high-capacity trucks. It has a total capacity of 375 tons (340,000 kilograms), carried on 28 wheels.

UPI Compix

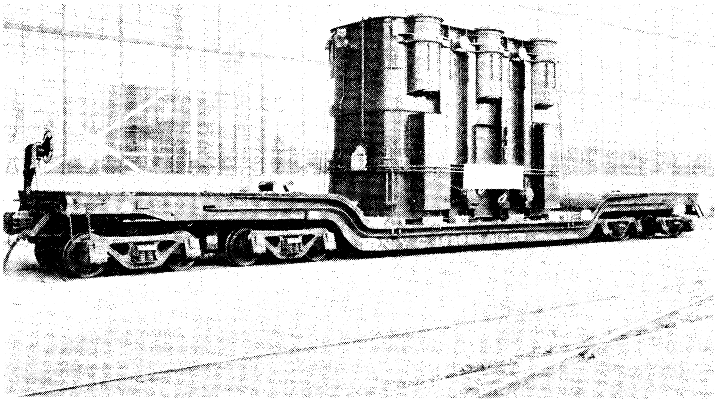


Figure 7: Specialized flatcar with low central section designed to carry electric power transformers while maintaining clearance in tunnels and under low bridges. The car is also unusual in its arrangement of eight axles.

Among the many types of railroad "service vehicles" are the so-called office, or business, cars used by company officials while on inspection tours of the line, track-measuring cars and rail "detector" cars that automatically record the condition of the track surface alignment or locate internal flaws in the rails, and various types of instructional, or personnel-training, cars. "Camp cars" are often used to house the personnel of maintenance forces engaged in major projects, such as renewing rail or constructing a new line.

Rolling-stock standardization. Much of the commercial effectiveness of the railroad has derived from the standardization of basic rolling-stock components. Each railroad builds cars suited to its own needs; but cars can be interchanged freely among railroads (assuming that they use a common track gauge) only if such elements as wheels, wheel bearings, couplers, and brakes are standardized. The knuckle type of automatic coupler was adopted in North America beginning in 1882; a similar coupler is used in several other countries, notably Australia, Japan, and the Soviet Union. Screw-and-buffer couplers are used in western Europe, but a knuckle coupler compatible with that used in the Soviet Union is under development.

There are two general types of continuous braking systems: the vacuum brake and the compressed-air brake. Of the two, the more popular is the air brake, patented in 1869 by George Westinghouse of the United States, where it was made compulsory equipment in 1893. The vacuum brake is used principally in Britain, where continuous brakes have been compulsory on passenger cars since 1889. With either system, application of the brakes on the locomotives applies the brakes throughout the entire train; and should the train become uncoupled, the brakes are automatically applied.

The Westinghouse air brake

The brake itself takes the form of a single or double shoe that presses against the wheel tread. The disk brake has also been widely applied on high-speed passenger trains and some freight cars as well.

RAILROAD TRACK AND ROADWAY

Location and construction. Ideally, a railroad should be built in a straight line, over level ground, between large centres of trade and travel. In practice, this ideal is rarely approached. The location engineer, faced with the terrain to be traversed, must balance the cost of construction against annual maintenance and operating costs, as well as against the probable traffic volume and profit.

Thus, in areas of dense population and heavy industrial activities, the railroads were generally built for heavy duty, with minimum grades and curvature, heavy bridges, and perhaps multiple tracks. Examples include most of the main-line railroads of Britain and the European continent. In North and South America, and elsewhere, the country was sparsely settled, and the railroads had to be built at minimal costs. Thus, the lines were of lighter construction, with sharper grades and curves. As traffic grew, main routes were improved to increase their capacity and to reduce operating costs.

The gauge, or distance between the inside faces of the running rails, is one of the main cost determinants. Generally, the narrower the gauge, the less costly is the line to construct and equip. This is why many of the railroads in underdeveloped, sparsely settled countries have been built to narrow gauges. On a narrow-gauge line, curvature can be more severe, less space is required, construction can be lighter, and rolling stock is less costly. Disadvantages are the limitation of speed because of reduced lateral stability and limitations on the size of locomotives and cars. About 60 percent of the world's railroad mileage is built to standard gauge, four feet 8½ inches (1.435 metres).

Eliminating adverse grades

The advent of modem high-capacity earth-moving machinery, developed mainly for highway construction, has made it economically feasible for many railroads to eliminate former adverse grades and curves through line changes. Graders, bulldozers, and similar equipment make it possible to dig deeper cuts through hillsides and to make higher fills where necessary to smooth out the profile of the track. Modem equipment has also helped improve railroad roadbeds in other ways. A number of railroads carried out ditching programs in which the drainage ditches along the roadbed were deepened. Where the roadbed is unstable, injecting concrete grout into the subgrade under pressure is a widely used technique. In planning roadbed improvements, as well as in new construction, railroads have drawn on modem soil-engineering techniques.

The first step in building a new railroad line, after the route has been surveyed and cleared of brush and trees, is to grade the right-of-way, much as is done when building a highway. Next, the crossties, or sleepers, are distributed and the rails laid and fastened to the ties. Then, ballast (usually crushed rock, slag, or volcanic ash) is applied. Finally, the track is aligned in both the horizontal and vertical planes, and the ballast is tamped, or compacted, around and under the ties.

In Canada, where much new railroad mileage was built after World War II, track-laying machines were often used (Figure 8). The machine is mounted on railcars. It feeds ties and rails ahead of the working crew, moving forward over the new track as soon as it is spiked down.

The inventors of the modern rail

Hail. The modern railroad rail has a flat bottom, and its cross section is much like an inverted T. An English engineer, Charles Vignoles, is credited with the invention of this design of rail in the 1830s. A similar design was also developed by Robert L. Stevens, president of the Camden and Amboy Railroad in the United States.

Present-day rail is, in appearance, very similar to the early designs of Vignoles and Stevens. Acutally, however, it is a highly refined product in terms of both engineering and metallurgy. Much study and research have produced designs that minimize internal stresses under the weight of traffic and thus prolong rail life. Sometimes the rail surface is hardened to reduce the wear of the rail under extremely heavy cars or on sharp curves. After they have been rolled at the steel mills, rails are allowed to cool slowly in special boxes. This controlled cooling minimizes internal shatter cracks, which at one time were a major cause of broken rails in track.

In Europe a standard rail length of 30 metres (98 feet five inches) is common. The weight of rail, for main-line use, is from about 45 kilograms per metre (about 90 pounds per yard) to 75 kilograms per metre (150 pounds per yard). British Railways uses a flat-bottomed rail weighing 55 kilograms per metre (110 pounds per yard).

Railroads in the United States and Canada have used T-rails of hundreds of different cross sections. Many of these different sections are still in use, but there is a strong trend to standardizing on a few sections. In the 1970s most new rail in North America weighed 119 or 136 pounds per yard. The standard American rail section has a length of 39 feet (about 12 metres).

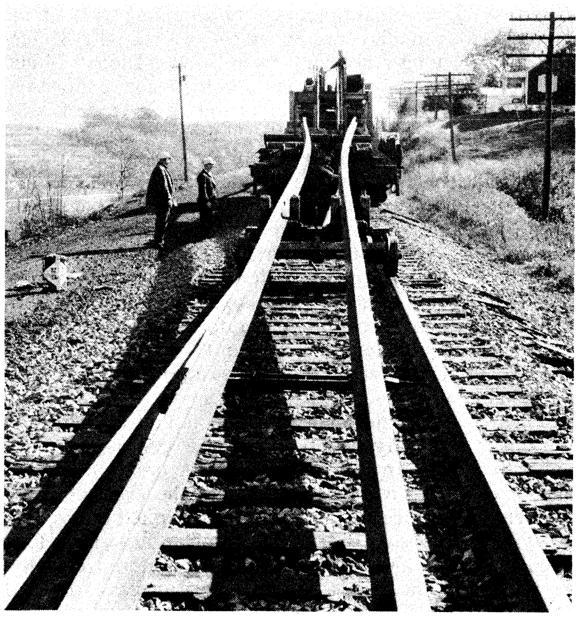


Figure 8: Moving work train laying quarter-mile lengths of continuous welded rail, which is positioned two strings at a time.

By courtesy of The Baltimore and Ohio Railroad Company

One of the most important developments is the welding of standard rails into long lengths. This continuous welded rail results in a smoother track that requires less maintenance. The rail is usually welded into lengths of about one-quarter mile (400 metres). Once laid in track, these quarter-mile lengths are often welded together in turn to form rails several miles long without a break.

Welded rail was tried for the first time in 1933 in the United States. It was not until the decade of the 1950s, however, that railroads turned to welded rail in earnest. By 1971 virtually all new rail, and much old rail taken up and relaid in new locations, was being laid in welded lengths. Welded rail was standard practice, or extensively used, in the United States, Japan, Canada, Germany, France, and Britain.

Controlling the temperature expansion of long welded rails proved not so difficult as first thought. It was found that the problem could be minimized by extensive anchorage of the rails to the ties to prevent them from moving when the temperature changes by the use of a heavy ballast section and by laying the rails when the ambient temperature is close to the mean temperature prevailing in the particular locality.

Fastening rails to crossties

Rail fittings. Whether in standard or long lengths, rails are joined to each other and kept in alignment by fishplates or joint bars. The offset-head spike is the most used and least expensive way of fastening the rails to wooden crossties, but several different types of screw spikes and clips are also used extensively on heavy-traffic lines in many countries. The rails may be attached directly to wooden crossties, but on heavy-traffic lines it is common to seat the rail in a tie plate that distributes the load over a wider area of the tie. A screw or clip fitting must be used to attach rails to concrete ties. A pad of rubber or other resilient material is always used between the rail and a concrete tie.

Crossties (sleepers). Timber has been used for railroad ties almost from the beginning, and it is still the most common material for this purpose. The modern wood crosstie is treated with preservative chemical to improve its life; the average life of crossties on main-line railroads is about 35 years. The cost of wood ties has risen steadily, creating interest in ties of other materials.

Steel ties have long been used in certain European, African, and Asian countries. Concrete ties, usually reinforced with steel rods or wires, have been gaining in popularity, as have ties consisting of concrete blocks joined by steel spacing bars. A combination of concrete ties and long welded rails produces an exceptionally solid and smooth-riding form of track. Concrete ties are extensively used in Britain, Europe, and Japan.

Tıack maintenance. Modern machinery enables a small group of men to maintain a relatively long stretch of railroad track. Machines are available to do all the necessary track maintenance tasks: removing and inserting ties, tamping the ballast, spiking rail, tightening bolts, and aligning the track. Mechanized equipment also can renew rail, either in conventional bolted lengths or with long welded lengths; cranes are used to remove the old rail and lay the new.

Prefabricated track sections

Complete sections of track—rails and crossties—may be prefabricated and laid in the track by mechanical means. Rail-grinding machines run over the track to even out irregularities in the rail surface. Track-measurement cars, under their own power or coupled into regular trains, can record all aspects of track alignment and riding quality on moving charts, so that maintenance forces can pinpoint the specific locations needing corrective work. Detector cars move over the main-line tracks at intervals with electronic-inspection apparatus to locate any internal flaws in the rails.

The mechanization of track maintenance after World War II has constituted a technologic revolution comparable to the development of the diesel locomotive and electrification. In Europe in particular, highly sophisticated maintenance machines have come into use.

Auxiliary plant. Railroad fixed plant consists of much more than the track. The New Sanyo Line, under construction in Japan in the 1970s, runs through tunnels for nearly half its total length. Railroad civil-engineering forces are concerned with constructing and maintaining thousands of buildings, ranging from switch tenders' shanties to huge passenger terminals.

Bridges. The designer of a railroad bridge must allow for forces that result from the concentrated impact that occurs as a train moves onto the bridge; the pounding of wheels, the sidesway of the train, and the drag or push effect as a train is braked or started on a bridge. These factors mean that a railroad bridge must be of heavier construction than a highway bridge of equal length.

As freight-train loads become heavier and train speeds higher, bridges need to be further strengthened. Another major objective in modem railroad-bridge construction is the need to minimize maintenance costs. The use of weathering steel, which needs no painting, all-welded construction, and permanent walkways for maintenance personnel contribute to this end. In the advanced countries there has been a widespread trend toward replacing timber trestles with concrete-slab structures or with concrete or steel-pipe culverts. The railroads also have sought ways to mechanize the maintenance of their bridges (see also BRIDGES, CONSTRUCTION AND HISTORY OF).

Buildings. Railroad buildings in the 20th century have become fewer and more functional. With paved highways running almost everywhere in the developed countries, it has become more economical to concentrate both freight and passenger operations at fewer but larger, strategically located stations. Only a few really modern passenger stations have been built. Notable among them are the highly efficient and functional stations on the New Tōkaidō Line in Japan, the new Euston station in London, and several of the new and rebuilt stations on the European continent, such as the main stations at Rome and Milan.

Diesel maintenance shops

Diesel and electric locomotives require few maintenance shops as compared with steam locomotives. Diesel shops are of three main types: small fueling, sanding, and light-maintenance centres at points where runs end or locomotives are changed; intermediate-maintenance shops, usually serving a region, that perform certain routine inspection and maintenance tasks; and heavy-repair shops, where locomotives undergo extensive repairing or rebuilding. Usually a single railroad requires only one of the latter type of shop.

Car shops, too, have been reduced in number and made more efficient through the use of process-line techniques. Terminal points and major classification yards usually have a shop for light repairs and maintenance of cars; a railroad also usually has one or more heavy-repair shops, where cars are completely overhauled or where new cars are manufactured.

It is usually more efficient to construct new shop buildings rather than convert old ones to handle modem types of rolling stock. Often, prefabricated buildings provide an economical solution to this problem. An important feature of new locomotive shops in particular is provision for collection and disposal of waste oil and other potentially polluting products of the shop or terminal operation.

Tunnels. Although very expensive, tunnelling provides the most economical means for railroads to traverse mountainous terrain or to gain access to the heart of a crowded city. Railroad tunnels, however, confront the construction engineer with some unique problems, particularly in the ventilation of tunnels on lines that are not electrified.

Some examples of famous tunnels and methods of construction are described in the article TUNNELLING AND UNDERGROUND EXCAVATION.

RAILROAD OPERATIONS AND CONTROL

Because a railroad's factory—its plant and train operations—may be spread out over thousands of miles and hundreds of communities, it has operating and service problems in some respects more complex than those of a major manufacturing installation. It is not surprising, therefore, that railroads have been among the pioneers in the use of improved methods of communication and control, from the telegraph to such present-day developments as the electronic computer and automation techniques.

Communications. Railroads were among the first to adopt the electric telegraph and the telephone, both for dispatching trains and for handling other business messages. Today, the railroads are among the larger operators of electronic communications systems.

Radio. Railroads began experimenting with radio at a very early date, but it became practical to use train radio on a large scale only after World War II, when compact and reliable very-high-frequency two-way equipment was developed. In train operations radio permits communication between the front and rear of a long train, between two trains, and between trains and the central dispatcher.

In terminals two-way radio greatly speeds yard-switching work. Through its use, widely separated elements of mechanized track-maintenance gangs can maintain contact with each other and with oncoming trains. Supervisory personnel often use radio in automobiles to maintain contact with the operations under their control.

Microwave communications

As the demand for more railroad communication lines has grown, more and more companies have begun to use broad-band radio beams (microwave) to supplement or replace the traditional lineside telegraph wires. As early as 1959, the Pacific Great Eastern Railway in western Canada began to use microwave radio for all communications, doing away almost entirely with line wires. Other railroads all over the world were turning to microwave in the 1970s.

Computers. A major reason for the growing use of microwave was the tremendously increased demand for circuits that developed from the railroads' widespread use of electronic computers.

Earlier, railroads had been among the leaders in adopting punched-card and other advanced techniques of data processing. In the 1970s there was a strong trend toward "total information" systems built around the computer. In such a system, each field reporting point, usually a freight-yard office or station, is equipped with a computer input device. Through this device, full information about every car movement (or other action) taking place at that point can be placed directly into the central computer, usually located at company headquarters. From data received from all the field reporting points on the railroad, the computer can be programmed to produce a variety of outputs. These may include train-consist reports (listing cars) for the terminal next ahead of a train, car-location reports for the railroad's customer-service offices, car-movement information for the car-records department, revenue information for the accounting department, plus traffic-flow data and commodity statistics useful in market research and data on the freightcar needs at each location to aid in distributing empty cars for loading.

Other pertinent data also may be integrated into the system so that the railroad's managers have a complete, up-to-the-minute picture of almost every phase of its operations. Such complete information and control systems promise to be a powerful tool for optimizing railroad operations and producing better service.

Signalling. Railroad signals are a form of communication designed to inform the train crew, particularly the engine crew, of track conditions ahead and to tell it how to operate the train.

The time-interval signal system

Methods of controlling train operations evolved over many years of trial and error. A common method in the early years was to run trains on a time-interval system; *i.e.*, a train was required to leave a station a certain number of minutes behind an earlier train moving in the same direction. It was common on single-track lines to program all operations in accordance with a timetable, which set up all the places trains were to meet and which could not be varied.

The development of distance-interval systems was a great improvement. In these so-called block systems, a train is prevented from entering a specific section of track until the train already in that section has left it.

The earliest form of railroad signal was simply a flag by day or a lamp at night. The first movable signal was a revolving board, introduced in the 1830s, followed in 1841 by the semaphore signal. One early signal consisted of a large ball that was hoisted to the top of a pole to inform the engineman that he might proceed (hence, the origin of the term highball).

The semaphore signal was nearly universal until the early years of the 20th century, when it began to be superseded, first by the colour-light signal and then by the searchlight type. The colour-light signal uses a separate lens and light bulb for each signal aspect, usually green, yellow, and red. The searchlight signal uses only a single, powerful lens and bulb; the different colours are displayed through the lens by means of roundels, or colour filters, that are rotated in front of the lamp. Two other types of signals are also used to a limited extent: the position light, in which rows of yellow lights duplicate the positions of semaphore arms, and the colour-position light signal, which uses coloured lights arranged in rows.

Most lines in Europe use a manual block system in which operations are controlled from wayside cabins or towers in conjunction with the wayside signals. Each tower controls a section; a train is not permitted to enter a section until the train ahead has left the section. Electric interlocking improves this system by making it impossible to give a "line clear" signal indication if the section is already occupied by a train.

The automatic block system

The basis of much of today's railroad signalling is the automatic block system, introduced in 1872 and one of the first examples of automation. It uses track circuits that are short-circuited by the wheels and axles of a train, putting the signals to the rear of the train and to the front as well as on single track at the danger aspect. A track circuit is made by the two rails of a section of track, insulated at their ends. Electric current, fed into the section at one end, flows through a relay at the opposite end. The wheels of the train will then short-circuit the current supply and de-energize the relay.

Signalling on African, Asian, and Australian systems usually follows European practice in areas of heavy traffic. On light traffic lines, control is often via the telegraph or telephone.

Operation on the basis of a timetable alone, which was common on early lines in the United States, had the disadvantage that if one train were delayed, others would also be delayed, since it was impossible to change the meeting points. By using the telegraph, and later the telephone, the dispatcher could issue orders' to keep trains moving in unusual circumstances or to operate extra trains as required. This "timetable–train order" system is still used on many lines in the United States and Canada. It is often supplemented with automatic block signals to provide an additional safety factor.

The first attempts at interlocking switches and signals were made in France in 1855 and in Britain in 1856. Interlocking at crossings and junctions prevents the signalman from displaying a clear signal for one route when he has already given clearance to a train on a conflicting route. Route-setting or route-interlocking systems are modern extensions of this principle. With them a towerman or dispatcher can set up a complete route through a complicated track area by simply pushing buttons on a control panel. This system allows a large area to be controlled from one point.

A logical development of the route-interlocking principle is centralized traffic control, a system in which trains are controlled entirely from a central point through remote operation of switches and signals (Figure 9). The operator sees the track layout in miniature on his control panel and directs the movement of trains over distances from a few miles to many hundreds of miles. Lights on the panel show the location and progress of all trains at all times.

By courtesy of the British Railways Board

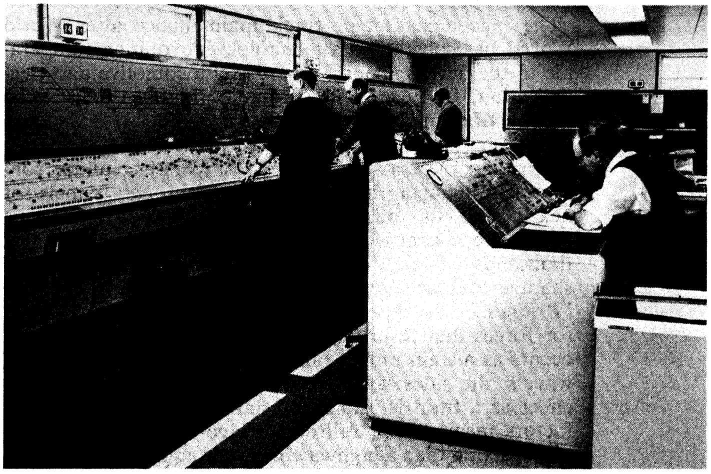


Figure 9: Computerized train control room at Bristol, England. Two panel boxes handle colour light signalling for 300 trains passing daily over the 254 track miles covered by this station.

In centralized traffic control, track circuiting is essential to ensure that the system always knows where each train is. Switches and signals are operated by coded electrical pulses that reduce the wiring required. Over long distances, centralized traffic control substantially increases track capacity by making more effective use of the trackage. Since it eliminates any need for written orders or manual operation of block signals, it permits closing telegraph or signal stations, another major economy.

The trend toward automation. A recent refinement in traffic control is to arrange the system for fully automatic operation. The machine will then set switches and clear signals for each train automatically; the dispatcher need exercise control only in unusual circumstances. This enables one dispatcher to control a still longer section of railroad. Completely automatic signalling activated by electronic program machines is used on some rapid-transit rail lines.

Automatic central traffic control

Automatic train control provides the locomotive engineman with audible (and sometimes visual) information on track conditions. Should he ignore a restrictive signal indication, the brakes are applied automatically to stop the train. A refinement of this system incorporates automatic control of train speed. A miniature signal in the cab repeats the aspects of the wayside signals (or it may take the place of wayside signals). Should train speed exceed that called for by the aspect being displayed, the brakes are applied and the speed reduced to the permissible level.

Only a slight further extension of this technique is needed to permit fully automatic operation of the train. By the early 1970s, a number of mining and industrial railroads were operated with crewless trains under full automation or remote control. On the high-speed New Tōkaidō Line in Japan, all trains operate under computerized automatic control throughout the entire 320-mile (510-kilometre) length of the line. The engineman, however, starts the trains, stops the train at station stops, and opens and closes the train doors. A similar system is used on the London Transport Victoria line, the Patco transit line in Camden, New Jersey, and the Bay Area Rapid Transit system in San Francisco.

Among other automatic aids to railroad operation is the infrared "hotbox detector," which, located at trackside, automatically detects the presence of an overheated wheel hearing and alerts the train crew. Broken flange detectors are used in major terminals to indicate the presence of damaged wheels. Dragging equipment detectors set wayside signals to danger if a car's brake rigging or other component is dragging on the track. Slide detectors warn of rocks or earth that have dropped onto the track from an earth cutting; high-water detectors warn of flood conditions on the track; high-wide detectors alert the train crew of a freight load that may have shifted or of a load that is too high or wide to clear bridges or tunnels.

The automated marshalling yard

A major area for automation techniques in railroading is the large classification or marshalling yard. In such yards, freight cars from many different origins are sorted out and placed in new trains going to the appropriate destinations.

Most large classification yards have a “hump” over which cars are pushed. They then roll down from the hump by gravity and each is routed into a classification or "bowl" track corresponding to its destination.

By 1971, operations in the newer classification yards had reached a high degree of automation. The heart of such a yard is a digital computer, into which is fed information concerning all cars in the yard or en route to it. As the cars are pushed up the hump, electronic scanners confirm their identity by means of a light-reflective label, place the data (car owner, number, and type) in a computer, and then set switches to direct each car into the proper bowl track. Electronic speed-control equipment measures such factors as the weight, speed, and rolling friction of each car and operates electric or electropneumatic "retarders" to control the speed of each car as it rolls down from the hump. Every phase of the yard's operations is monitored by a computerized management control and information system.

Modern classification yards usually also use radio, telephone, teleprinter, pneumatic tube, and closed-circuit television. Repair shops adjacent to the yards are designed for quick, mechanized repair of cars found to be defective as they move over the hump.

Because such electronically equipped yards can sort cars with great efficiency, they eliminate the need to do such work at other, smaller yards. Thus, one large electronic yard may permit the closing or curtailing of a dozen or more other yards. Most modern electronic yards have quickly paid for themselves out of operating savings—and this takes no account of the benefits of improved service to shippers.

REGULATION, ECONOMICS, AND LABOUR RELATIONS

Regulation and public control. Although most of the early railroads were built and operated as private, profit-making businesses, railroads soon came under relatively intense public scrutiny and regulation. This was probably inevitable because of the far-reaching influence railroads quickly developed over the life and the commercial activities of the communities and countries they served. The history of regulation followed the same path with minor variations in all countries; beginning with regulation of railroad construction—that is, deciding whether to permit a new line to be built and approving its location—the government took an increasing interest in details of operation, especially in connection with safety. Railway acts were passed in country after country prescribing signal modes, brakes, track standards, employee training, and other aspects of railroad operation. At the same time, governments took an increasing interest in the financial side of railroads, regulating rates and fares and exercising licensing power over mergers and other financial operations. Degree of public regulation varied, but by the early 1970s in nearly every country in the world, with the notable exception of the United States, regulation had evolved into public ownership and operation. This resulted from the railroads' increasing financial difficulties in the course of the 20th century.

Post-war financial problems

Competition from other modes increased sharply after World War 11 for many of the world's railroads. Because of governmental and labour union restrictions, as well as the large investment required in railroad fixed plant (*i.e.*, track, bridges, buildings, stations), it was difficult for the railroads lo adjust their operations to changing conditions. Unlike other kinds of businesses, a main-line railroad simply cannot cease operating when it gets into financial difficulties; its services are too essential to the economy of the area it serves. (Many secondary and branch-line railroads have, however, been pulled up, especially in Britain and the United States.) Thus, the solution has been to nationalize the railroad—to have the government provide the service. Operating deficits are then met out of tax revenues.

By 1971 only two large countries, the United States and Canada, had major railroads that were privately run. In Canada one large system, Canadian Pacific (CP Rail), remained a private enterprise, while the other line, Canadian National Railways, was a government-owned corporation. France nationalized its remaining privately owned railroads in 1938; Britain nationalized its four major companies (which had been formed out of 123 separate railroads after World War I) in 1948.

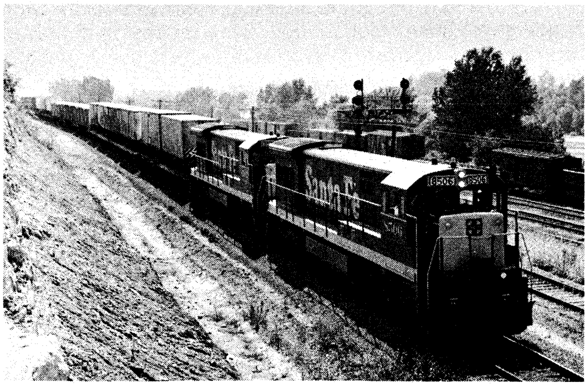
One result of growing financial difficulties in the United States was a wave of mergers during the 1950s and 1960s. Pushed through after protracted regulatory and legal delays were a number of important consolidations. Some of them were the merger of the Erie and the Delaware, Lackawanna and Western railroads to form Erie Lackawanna; Pennsylvania, New York Central, New York, New Haven and Hartford Railroad Company into the Penn Central; merger into the Norfolk and Western Railway of the former New York, Chicago and St. Louis and the Wabash railroads; merger of Atlantic Coast Line and Seaboard Air Line railroads to form Seaboard Coast Line Railroad; merger of Chicago, Burlington and Quincy Railroad, Great Northern Railway, Northern Pacific Railway, and Spokane, Portland and Seattle Railway to form Burlington Northern.

In 1970 the largest railroad in the United States, Penn Central, entered bankruptcy proceedings, and it appeared doubtful that it could be successfully reorganized as private enterprise. Other railroads in that country were also in serious financial difficulties and some form of nationalization appeared increasingly probable.

Innovations by nationalized railroads

While nationalization preserves services felt to be essential, it does not obviate the need for railroads to change, improve, and compete. Some of the most imaginative marketing innovations (for example, Freightliner, Trans-Europ Express, Japan's New Tōkaidō services) have come out of nationally run companies. In France the French National Railways have pioneered in freight-rate innovations. The French National Railways were reorganized to operate much like a privately run business, if possible at a profit. Other nationalized railroads, such as those in Sweden, Switzerland, The Netherlands, and Canada, also operate as much as possible under the incentives of a private business.

Competition and marketing. Planned as money-making enterprises, many early railroads were notably successful financially. Indeed, especially in Britain and the United States, railroad stocks were long highly regarded. Even today, a few of the more strategically located and



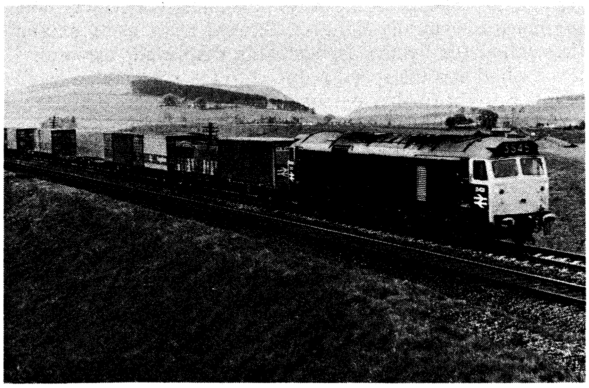


Figure 10: (Left) Santa Fe Railway's new Super C, known as the world's fastest freight train, which goes from Chicago to Los Angeles at top speeds of 79 miles per hour. (Right) British Railways' Freightliner container train, which operates at high speeds throughout the United Kingdom.

By courtesy of (left) Santa Fe Railway, (right) the British Railways Board

well-managed American lines are notably good earners. In recent years, however, the picture has been much different.

The development of the internal-combustion engine and its application to highway vehicles and the invention of the airplane had far-reaching effects on railroad transportation. In the interval between the two world wars, buses, automobiles, and highway trucks were already cutting deeply into traffic that once had been exclusively the railroads'. World War II hastened the technological development of these new competitors as well as of air and pipeline transportation.

By the late 1950s, the inroads of competitors had reached serious proportions, from the railroad viewpoint, in the United States, Britain, and Canada; by 1970 the same situation was developing in western Europe and Japan. The loss of traffic extended to both passengers and freight and was especially pronounced on branch lines. By 1970, railroads in the United States had virtually abandoned the business of carrying intercity passengers, although they still moved much local commuter traffic in several large urban areas. (A quasi-government company, the National Railroad Passenger Corporation [Amtrak], was formed in 1970 to take over and if possible improve the operation of about 200 intercity passenger runs thought to be essential to the national economy.) Railroads in the United States were also suffering from the diversion of profitable freight traffic to trucks and pipelines. Because of the nation's overall economic growth, however, the railroads continued to set records each year in freight volume.

In Britain, western Europe, and Japan, competitors were especially effective in diverting freight traffic from the railroads, although passenger traffic was also affected as the ownership of automobiles grew. Accustomed to competing, if at all, only with each other, the railroads found it hard to adjust their thinking and their operations to meet the challenge of competitors who could offer the public faster, more flexible, and in some cases cheaper service, and who in most cases were burdened much less by government regulation. Many railroads, however, began to fight back against their competitors with considerable success. In the passenger field, typical railroad responses to competition include much faster and more comfortable services, such as the TEE (Trans-Europ Express) services on the European continent, the Inter-City expresses in Britain, and the Metroliners in the United States.

In freight, considerable success has resulted from the "marketing approach," wherein railroads closely tailor their rates, services, and equipment to the particular needs of specific shippers. Typical are the "unit" or "block" trains operated for shippers of bulk commodities, such as coal, oil, ore, and grain. These trains are composed of large, modem cars designed for the commodity to be carried. They operate as a unit on fast schedules between one origin and one destination, bypassing all intermediate yards and terminals en route. With faster operation and larger cars, these trains are so productive that they permit the railroads to offer greatly reduced rates.

Another way in which railroads have responded to new competition is to offer shippers many special types of freight cars designed to load particular commodities quickly and at minimum cost, such as the trilevel autorack car, the 10,000-cubic foot (280-cubic metre) boxcar, and 100-ton (91,000-kilogram) covered hopper cars and gondolas.

The piggyback concept and containerization

Another significant competitive development was piggyback and containerized services (Figure 10). The piggyback idea, which actually dates from the 19th century, combines the flexibility of truck pickup and delivery with the economy of rail movement between cities. Along with piggyback development has come increasing interest among railroads (as well as other modes of transport) in container systems, by which merchandise could be loaded into large standard containers or boxes that could move via highway on a truck chassis, via rail on special container cars, in ships especially equipped to handle them, or even by air. A single shipment might use two or more modes of transport in the course of its trip. In western Europe special TEEM (Trans-Europ Express Merchandise) trains operate between major points, carrying only containers. In Britain, fast container shuttle trains now operate between about two dozen main cities and ports. This Freightliner system (Figure 10) has been highly successful and is said to be competitive with over-the-road trucking, even over relatively short distances. Canadian and Japanese railroads also have extensive container operations, mainly for import-export traffic.

Labour relations. Railroads were among the first industries to feel the effects of trade union activity. Railroad unions in the United States are organized along craft lines; in the early 1970s there were 19 railroad labour organizations. In most countries, however, a single large union embraces all railroad workers. Labour relations are governed by a variety of laws, which, in the democratic countries, guarantee the right of collective bargaining, usually with provision for mediation. Though railroad labour history includes many chapters of violence and bitterness in the past, its more recent era has been relatively peaceful. However, there has been a discernible tendency toward an increase in work stoppages in the 1960s and 1970s.

THE FUTURE OF RAILROADS

The railroad industry can look back over a proud history. Railroads were a vital element in the Industrial Revolution. They helped make Britain an industrial power, played similar roles in countries such as France and Germany, and went on to do much the same in Russia and Japan. Railroads almost literally built the United States and Canada; and they remained the economic backbone of most of the major world powers.

But as the last third of the 20th century began, the railroads in a number of countries, most notably Britain and the United States, were in serious trouble. The railroads' share of the total transportation business was dropping steadily. The railroads in most countries had long since come under state control, and the remaining privately owned lines were finding it difficult to operate at a profit. Some observers believed the end of railroading was in sight. Technology was evolving so rapidly that it was impossible to say, as of the early 1970s, that this could not happen, but it seemed unlikely. It was becoming clear, however, that the railroad of the future would have a different role to play than in the past.

Predictions of the end of railroading

For more than a century the railroad was the dominant form of land transportation in much of the world. It was, and remains, the one land carrier that can carry almost anything, anywhere the tracks go, and do it at a true cost lower than other types of land-air transportation.

Today, however, other modes of transportation have been developed to the point where they can do certain transportation jobs more effectively than the railroads. Pipelines can carry liquids and some solids over long distances economically. Airplanes, with their great speed, can carry some types of light, valuable freight at a saving; and trucks offer speed and flexibility, especially for the shorter hauls. The private automobile, operating over modern highways, and the airplane have taken over much passenger traffic formerly handled on rails. The motorbus is an effective competitor for the short- to medium-distance passenger business. The modern barge, operating on improved inland waterways systems, can move many commodities over specific routes at very low cost.

Undeniably, these competitors of the railroad can do a better job on some types of transportation tasks. The development of these newer modes, therefore, has changed the role of the railroad from that of the general-purpose carrier to that of a more specialized carrier, just as the other modes are specialized.

The future role of the railroad as a specialized carrier will vary in different nations. In general, however, the railroad is particularly strong in these areas:

1. It is especially effective in moving large volumes of bulk commodities, such as coal, ores, chemicals, and grain, over relatively long distances. It can also move large volumes of finished merchandise economically at relatively high speeds over long distances.

2. The railroad can efficiently handle containers in large volumes between major centres, and in some countries, trucks on "piggyback" trains. An efficient railroad container or piggyback shuttle system can be viable even over relatively short distances.

3. The railroad is the best mode for moving large numbers of commuters between big metropolitan centres and the outlying suburban areas.

4. Very high speed intercity passenger services can be successful when operated with modern equipment at distances of up to about 300 miles (500 kilometres).

In short, the railroad under modern conditions is at its best as a high-volume, medium- and long-distance carrier of both passengers and freight. There are, of course, many exceptions to this general rule.

The potential of the modern railroad

In looking at the future place of railroads, three other factors should be noted:

1. A railroad disturbs the natural environment far less than a highway or an air-transport system. It also produces less pollutants per unit of transportation performed than either highway or air transportation. These factors should become more significant as society increasingly concerns itself with the need to preserve the environment and to reduce air, water, and noise pollution.

2. A railroad is far more efficient in its use of fuel than are either highway or air transportation. It is probable that future concern over the best use of these resources will produce more emphasis on rail transportation.

3. While much public money has gone into technological research on the newer forms of transportation and into constructing facilities for them, relatively little has been spent to improve railroad technology. Thus, even the most advanced of today's railroad plants and services, with few exceptions, do not represent anything like the best that is possible from the railroad.

Modern passenger trains. *The New Tōkaidō* Line. In the early 1970s, perhaps the outstanding example of modern railroad technology was the New Tōkaidō Line of the Japanese National Railways (Figure 11). The Tōkaidō (Eastern Sea Route) Line between Tokyo and Osaka serves an area where 40 percent of Japan's population and 70 percent of its industrial output are concentrated. By the early 1960s the original double-track, 3-foot-6-inch- (1.07-metre-) gauge line between those points had reached capacity. The New Tōkaidō Line was opened in 1964 as a completely new, standard-gauge line, entirely separate from the original line, with its own right-of-way and stations, and handling high-speed passenger trains only.

By courtesy of ti-e Museum of Railwav Trafflc, Tokyo

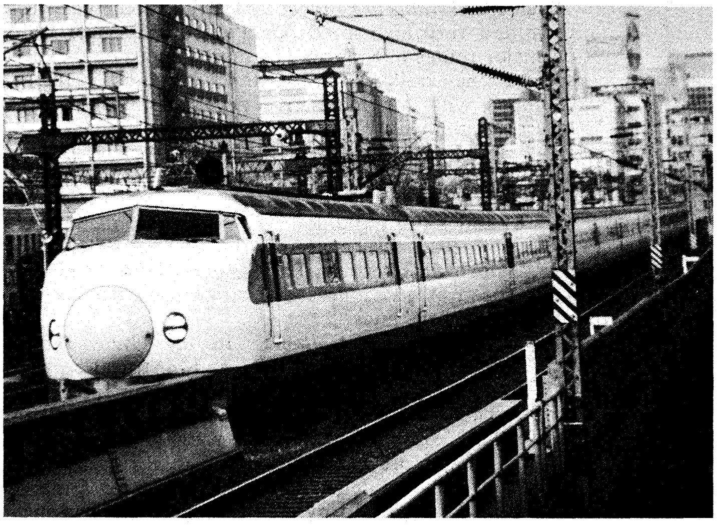


Figure 11: The New Tōkaidō Line train, developed by the Jaoanese National Railwavs. It achieves speeds of uo to 132 miies per hour on its 320-mile run between Osaka and Tokyo. Comprised of six two-car units, the train operates on electric power supplied by overhead wires.

Electrified with a trolley wire voltage of 25,000 volts, 60 hertz, the new line has no grade crossings, and its curves and grades permit operation at 150 miles (240 kilometres) per hour. The track was laid entirely on concrete ties, except at turnouts (switches). Twelve-car electric trains provide the service. The express trains make the 320-mile (510-kilometre) Tokyo–Osaka run in three hours and six minutes, at an average speed of 103 miles (166 kilometres) per hour, stopping at Nagoya-Mura and Kydto. "Limited express" trains make ten intermediate stops and cover the full distance in four hours and ten minutes. The trains reach a maximum speed in normal oyerations of 130 miles (210 kilometres) yer hour.

The New Tōkaidō Line proved outstandingly successful and highly profitable. By the early 1970s it was averaging 233,000 passengers daily, carried in 202 trains (101 each way). Its success prompted the Japanese National Railways to begin constructing a westward extension, the New Sanyo Line, from Osaka toward Hagata, about 350 miles (560 kilometres). This line was expected to open to Okayama, 100 miles (160 kilometres), in 1972; it will handle 16-car trains operating at 155 miles (250 kilometres) per hour. In addition, planning was going ahead for a nationwide network of similar high-speed, standard-gauge trunk lines. Three additional lines—from Tokyo to Morioka, Niigata, and Narita, respectively—were under construction.

Successors to the New Tōkaidō

London–Manchester–Liverpool. Between 1960 and 1966, British Railways carried out a major project to electrify and upgrade its heavy-traffic main lines between London (Euston Station) and Birmingham, Manchester, and Liverpool. This involved electrifying 443 route miles (713 kilometres) or 1,467 track miles (2,360 kiiometres) of line, using a trolley voltage of 25,000, 50 hertz, alternating current. Although this is a modernized existing line, not a completely new railroad as in the case of the New Tōkaidō, it has proved very successful. Passenger carryings on the line doubled in four years after the electrified operation made possible faster, more frequent service. On this line, trains operate regularly at 100 miles (160 kilometres) per hour. Electrification of this line is now being extended all the way to Glasgow.

High-speed trains in France. Electrification has also permitted very high-speed operation of passenger trains on major routes in France. The most famous of the high-speed French trains is the "Mistral," which has now been equipped with the most modern Trans-Europ Express type of equipment. This train makes a daily run between Paris and Nice, 676 miles (1,088 kilometres), at an average speed of 75 miles (120 kilometres) per hour. Another French Trans-Europ Express train, "Le Lyonnais," averages 85 miles (135 kilometres) per hour on its 320-mile (510-kilometre) run between Paris and Lyon. The "Capitole," which runs between Paris and Toulouse, is allowed a top speed of 125 miles (200 kilometres) per hour on some parts of the lines.

The United States Metroliners. In the United States, the strong emphasis on highways and air-travel facilities had, by the 1960s, caused most railroads in the United States to cut their passenger operations drastically. In the Northeast Megalopolis extending roughly from Boston through New York and Washington to Richmond, Virginia, however, the dense population presented a market that could be exploited by a fast, modern rail passenger service. In 1968 the Penn Central railroad, in conjunction with the United States Department of Transportation, began operating an experimental high-speed service on its existing electrified (11,000 volt, 25 hertz) main line between New York and Washington. Comfortable and modern electric multiple-unit train sets provide these services, which in 1971 were taken over by Amtrak (The National Railroad Passenger Corporation). Twelve Metroliner trains daily were operated in each direction as of 1972, with a running time of three hours for the 225-mile (360-kilometre) trip, including five intermediate stops. Average speed is 75 miles (120 kilometres) per hour, but trains are capable of running up to 160 miles (260 kilometres) per hour. The existing line, however, although it has been improved, has a number of highway grade crossings and restrictive curves; top running speed is currently held to 100 miles (160 kilometres) per hour.

Success of the Metroliner

Although the accommodations on the Metroliner trains are as advanced as any, the Metroliners cannot compare in frequency and speed with the high-speed passenger services of other countries described above. Nevertheless, they have reversed a long decline in patronage on the New York–Washington line; the trains consistently operate with large loads of passengers.

Modern freight trains. *Shuttle freight operations.* Perhaps the ideal freight railroad of the future would be one that is essentially a shuttle or "conveyor belt," carrying one commodity continuously between two points. Such a railroad would produce the maximum amount of transportation at the lowest cost. A few such railroads exist; they are mainly mineral railroads carrying coal or ore from mines to points of use.

One such line is an intraplant railroad, double-track, slow-speed, that carries taconite ore from mines 46 miles (74 kilometres) to a boat-loading dock. This relatively short line operates around the clock and carries some 30,000,000 tons of taconite annually; so efficient is the operation that it produces transportation at the rate of about 220,000 gross ton-miles per train-hour (a statistic derived by multiplying the weight of cars and contents by the number of miles hauled, and dividing this figure by the time required) even though half of the train-miles are run empty. This figure is about three times the average of the large general-purpose common carrier railroads in the United States and double the production of the most efficient of the general-purpose railroads.

A one-commodity, shuttle railroad is relatively rare; but similar gains in efficiency can be made on the common carrier railroads by operating unit trains. These are rarely as efficient as the example just cited; but they can be quite effective, especially when operated at fairly high speeds over long distances.

Advantages of the shuttle train

The shuttle-train principle is powerful because it obviates the delay and expense of moving trains through freight yards, since the trains operate directly from shipper to receiver; moreover, it greatly improves the utilization of costly locomotives and cars. Unit trains and similar applications of this principle are being tried by railroads all over the world.

Freightliner and Super C. The container shuttle train holds great promise for similar efficiencies in the hauling of general merchandise. In 1965 the first Freightliner container (Figure 10) shuttle trains began running on British Railways. By 1970 some 28 terminals and container ports were being served; the trains were operating over 80 different routes. More than 500,000 containers were being hauled annually.

The Freightliner system provides an integrated road-rail service direct from shipper to receiver, for both domestic and export–import business. Shippers may use their own containers if they conform to international standards. Space on the fast overnight trains is reserved in advance; the trains operate through to destination without intermediate reclassification in yards. Terminal handling cranes are designed for quick transfer of the containers from rail cars to truck chassis and vice versa. Freightliner service has proved competitive with over-the-road trucking even on many of the shorter runs prevailing in Britain.

Another glimpse of the railroad future is the Santa Fe (United States) railroad's Super C, a freight train carrying only containers or highway trailers. Super C runs 2,200 miles (3,500 kilometres) between Chicago and Los Angeles in less than 40 hours, at an average speed greater than that of the road's fast passenger trains. At this speed, the railroad is more than competitive with express trucks operating over highways.

The examples cited above are representative of some of the better achievements of present-day railroad technology. If they become typical instead of exceptional and if research can further improve existing technology, the railroad of steel rails and flanged wheels will remain viable for many years to come. Passenger train speeds of 125 to 150 miles (200 to 240 kilometres) per hour and freight train speeds up to 100 miles (160 kilometres) per hour can become entirely routine.

Exotic systems. In the 1960s and early 1970s, considerable interest developed in the possibility of building tracked passenger vehicles that could go much faster than conventional trains. Experiments conducted by the Japanese National Railways and others had indicated that the practical upper limit of speed for flanged-wheel railroad vehicles might be in the range of 150 to 200 miles (240 to 320 kilometres) per hour.

A tracked air-cushion vehicle

Several new types of guided vehicle systems have been proposed for possible ultra-high-speed operation. One of the more promising is the tracked air-cushion vehicle, as exemplified by the French Aerotrain (Figure 12). Air-cushion vehicles use a "cushion" of low-pressure air to "float" the vehicle away from the group or the guideway; they have no wheels and, when running, no contact with the guideway.

The Aerotrain has been under development for some years. It is a single vehicle that runs on an elevated beam-

By courtesy of Rohr Corporation

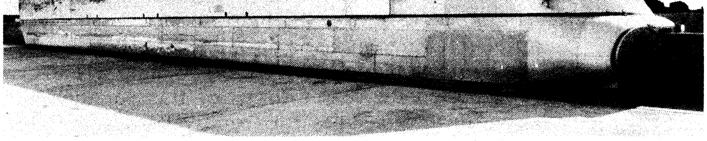


Figure 12: Demonstration model of the French Aerotrain in operation on an 11-mile guideway near Orléans, France. Running at a maximum speed of 170 miles per hour, the train rides on a cushion of air less than one inch above its inverted T-shaped concrete monorail.

way that has a vertical centre guide beam. A prototype 80-passenger vehicle designed for intercity operations used fans to obtain the air for lift and lateral guidance; a propeller driven by two high-powered gas turbines provides propulsion. It has been under test on an elevated guideway near Orléans, France, where it has achieved speeds of up to 180 miles (290 kilometres) per hour. The test section may eventually become part of a projected '70-mile (110-kilometre) Paris–Orléans route.

A tracked air-cushion vehicle project is also under way in Britain. The vehicle, which will be tested at speeds up to 150 miles (240 kilometres) per hour, will be propelled by a linear induction motor. The linear induction motor could also be applied to conventional flanged-wheel vehicles; the United States Federal Railroad Administration is experimenting with this application.

Another proposed type of guided high-speed system is the magnetic-levitation, vacuum-tube system. Vehicles would run at very high speeds in underground tubes from which the air would be exhausted. They would be held away from the tube walls by magnetic force. So far, only small-scale experimental work has been done on this concept.

At their state of development in the early 1970s, none of these proposed systems seemed likely soon to displace the conventional railroad as an all-around high-volume, low-cost mover of both people and goods. They did, however, show promise as high-speed city-to-airport links and perhaps for certain high-speed intercity runs. Much depended on whether the costs of building and operating these still experimental systems could he made comparable to those of the proven flanged wheel, steel-rail system.

BIBLIOGRAPHY

General: *Jane's World Railways* and the *Directory of Railicay Officials and Yearbook* (both annual), comprehensive statistical data on operations of specific railroads (also lists of officers).

United States: A.M. WELLINGTON, *The Economic Theory of the Location of Railways,* 6th ed. (1904), a classic on location and operation principles; W . ~HAY, *Railroad Engineering* (1953) and *Atz Introduction to Transportation Engineering* (1961), on basic principles of design and operation; A.W. BRUCE, *The Steam Locomotive in America* (1952), a history of steam locomotive development in the 20th century: S.H. HOLBROOK, *The Story of American Railroads* (1947), a good popular history of railroads in the United States; R.S. HENRY, *Trains* (1954), a description of the organization and operation of American railroads. *Moden Railroads* (monthly); *Railway Age* (semimonthly); *Trains* (monthly); ASSOCIATION OF AMERICAN RAILROADS, *Railroad Facts* (annual); *Quiz on Railroads and Railroading* (irreg.); *A Review of Railway Operations* (annual), statistics on current American operations; *Moody's Transportation Manual* (annual), comprehensive financial and operating data on United States lines.

Canada: *Canadian Transportation and Distribution Management* (monthly).

Latin America: FERROCARRILES NACIONALES DE MEXICO (National Railways of Mexico), *Ferronales* (monthly); PAN AMERICAN RAILWAYS CONGRESS, *Bulletin* (bi-monthly).

Great Britain: BRITISH RAILWAYS BOARD, *The Reshaping of British Railways* (1963), the "Beeching Report" that influenced present policies; R. BELL, *History of British Railways During the War, 1939–45* (1946), an excellent account of railroads under wartime conditions; *Michael Robbins, The Railway Age in Britain* (1962), a highly readable account of British railroad history and development. *Railway Gazette International* (monthly); *Railway Magazine* (monthly); *Modern Railways* (monthly).

Europe: INTERNATIONAL RAILWAY CONGRESS ASSOCIATION, *Rail International* (monthly); *Revue Générale des Chemins de Fer* (monthly); *Ingegneria Ferroviaria* (monthly); *International Railway Journal* (monthly); *Die Bundesbahn* (semimonthly).

Japan: JAPANESE NATIONAL RAILWAYS, *Fact and Figures* (annual); *Japanese Railway Engineering* (quarterly).

Australasia: *Railway Transportatiott* (monthly).

(T.C.S.)

Rājasthān

Rājasthān, a state of the Indian Union, is situated in the northwest of India. It is bounded on the west and northwest by Pakistan, on the north and northeast by the states of Punjab, Haryana, and Uttar Pradesh, on the east and southeast by Uttar Pradesh and Madhya Pradesh, and on the southwest by Gujarāt. The Tropic of Cancer passes through its southern tip in the Bānswāra district. The state has an area of 132,149 square miles (342.267 square kilometres) and its population at the 1971 census war 25,724,000. The capital city is Jaipur.

"The abode of the rajas"

Riijasthiin, meaning "the abode of the rajas," was formerly called Rājputana, "the country of the Rājputs" (sons of rajas). Before 1947, when India achieved independence from British rule, it comprised 18 princely states, two chiefships, the small British-administered province of Ajmer-Merwara, and a few pockets of territory outside the main boundaries. After 1947 the princely states and chiefships were integrated with India in stages, and the state took the name of Rājasthān. It assumed its present form on November 1, 1956, when the act for the reorganization of the Indian states came into force (for an associated physical feature, see THAR DESERT).

History. Archaeological exploration has shown that early man lived along the banks of the Banās and its tributaries some 100,000 years ago. Harappan and post-Harappan culture (3rd–2nd millennium BC) is traceable at Kalibangan, Ahār, and Gilund. Pottery fragments at Kalibangan are carbon-dated to 2700 BC. The discovery near Bairāt of two rock inscriptions of the emperor Ásoka (*c.* 250 BC) seems to show that his rule extended westward to this part of the state. Later rulers of the whole or parts of the state were the Bactrian Greeks (2nd century BC), the Scythians (Sakas; 2nd to 4th centuries AD), the Gupta dynasty (4th to 6th centuries), the White Huns (6th century), and Harṣavardhana, a Rājput ruler (early 7th century). Between the 7th and 11th centuries several Riijput dynasties arose, including that of the Gurjara-Pratihiiras, who kept the Arab invaders of Sind at bay.

Rājput strength reached its zenith at the beginning of the 16th century under Rānā Sangriim Singh (Sāngā) of Mewār, but he was defeated in a fierce battle by the Mughal invader Bābur, and the brief splendour of a united Rājputana waned rapidly.

When the Mughal emperor, Akbar, came to power in midcentury, the chief of the Kachwāhās at Amber entered the imperial service; Jodhpur and Chitor were subdued. It is noteworthy that the emperors Jahāngīr and Shāh Jahān were both born of Riijput mothers. After the death (1707) of the emperor Aurangzeb, the Riijput state of Bharatpur was developed by a Jat conqueror, but by 1803 most of the rest of Riijputana was under the domain of the Marāthā people of the central Arabian coast. Later in the 19th century the British subdued the Marāthās and with the gradual penetration of British paramountcy the boundaries of the Riijput states were precisely defined. The government of India was represented in Riijputana by a political officer, styled agent to the governor general, who was also chief commissioner of the small British province of Ajmer-Menvara. Under him were residents and political agents who were accredited to the various states.

The birth of nationalism

During this period the idea of Indian nationalism was born. Maharishi Dayanand wrote at Udaipur his *Satydrath Prakāsh*, intended to restore Hinduism to its pristine purity, which created a ferment in Riijputana. Important movements of thought also occurred among the Jain *sādhus* ("holy men") and scholars. Ajmer was the centre of political activity, and nationalist leaders included Arjun Lal Sethi, Manik Lal Varma, Gopal Singh, and Jai Narain Vyas. The people of Rājputana joined the people of the rest of India in nationalist aspirations and were no longer morally dominated by the princes. When the new constitution of India came into force in 1950 Rājasthān became an integral part of India. The Rājput princes surrendered their powers to the central government.

The landscape. The Arāvalli Hills form a line across the state running roughly from Mount Abu, which is 5,650 feet (1,722 metres) high, in the southwest to Khetri in the northeast. About three-fifths of the state lie northwest of this line, leaving two-fifths in the southeast. These are the two natural divisions of Rājasthān. The northwest tract is sandy and unproductive with little water but improves gradually from desert land in the far west and northwest to comparatively fertile and habitable land toward the east. The area includes the Thar (Great Indian) Desert.

The southeastern area is higher (330 to 1,150 feet [100 to 350 metres] above sea level), more fertile, and very diversified in character. In the south lies the hilly tract of Mewār. A large area of the districts of Kota and Biindi forms a tableland; to the northeast of this there is a rugged region following the line of the Chambal River. Farther north the country levels out; the flat plains of Bharatpur district form part of the alluvial basin of the Yamuna River.

The rivers of Rājasthān

The Arāvallis form the most important watershed, the drainage to the east of the range flowing northeast. The Chambal is the only large and perennial river. Its principal tributary, the Banās, rises in the Arāvallis near Kumbhalgarh and collects all the drainage of the Mewiir Plateau. Farther north, the Bāngangā, after rising in Jaipur, flows east to join the Yamuna. In the northwestern sector the Lūni is the only significant river. This rises in the Pushkar Valley of Ajmer and flows 200 miles west-southwest into the Rann of Kutch.

The soils. The soils in the vast sandy northwestern plain of the districts of Jaisalmer, Barmer, Jālor, Sirohi, Jodhpur, Bikaner, Gangānagar, Jhiinjhunu, Sikar, Pāli, and Nāgaur improve in fertility toward the east and northeast. These soils are predominantly saline or alkaline. Water is scarce but is found at a depth of 100 to 200 feet. The soil and sand are calcareous (chalky). Nitrates in the soil increase its fertility.

The soils in the Ajmer district in central Rājasthān are sandy; clay content varies between three percent and nine percent. The salt content is low. In Jaipur and Alwar districts in the east soils vary from sandy loam to loamy sand. In the Kota, Biindi, and Jhālawār tract, they are in general black and deep and are well drained. In Udaipur, Chitorgarh, Diingarpur, Biinswiira, and Bhīlwāra the eastern area has mixed red and black and the western area red and yellow soils.

Climate. There is a wide range of climate varying from extremely arid to humid, the humid zone comprising the Sirohi, Bānswāra, and Jhālawār districts in the south. Except in the hills, the heat in summer is great everywhere, with a mean maximum of 108" F (42° C). Hot winds and dust storms occur, especially in the desert tract. Winter temperatures vary from 68° F to 76° F (20" C to 24.5" C). The western desert has little rain (annual average ten centimetres [four inches]), but in the southwest rainfall is copious. The southwest benefits from both the Arabian Sea and Bay of Bengaı summer monsoon winds, which bring 90 percent of the rainfall.

Vegetation. The main floral feature is scrub jungle. Toward the west, plants characteristic of the arid zone occur, such as tamarisk and false tamarisk. Trees are scarce, found only sparingly in the Ariivallis and in eastern Rājasthān.

Wild life. Tigers are found in the Ariivallis and in several districts. Leopards, sloth bears, sambur (dark brown Indian deer), and chital (a kind of deer) occur in the hills and forests, where *nīlgais* (blue bulls) are also to be found in parts; black buck and ravine deer are numerous in the plains. Snipe, quail, partridge, and wild duck occur everywhere except in the desert. The Bikaner region is well known for several species of sand grouse.

Tribal peoples

Population. Aboriginal tribes in Alwar, Jaipur, Bharatpur, and Dholpur areas include the Mīnās; the Meos; the Banjārās, who are travelling tradesmen and artisans; and the Gadia Lohārs, another itinerant tribe, who make and repair agricultural and household implements. The Bhils, one of the oldest tribes in India, inhabit the former princely states of Diingarpur and Bānswāra and parts of Partāpgarh, Mewār, and Sirohi. Typical Bhils are small, dark, and broad-nosed. They are famous for their skill in archery. The Grasias and Kathodis live in the Mewiir Hills. The Grasias are tall, with good physique; unflinching loyalty to tribe, friends, or employers is a feature of their character. The Kathodis are nomads, Sahariyas are found in the Kota district, and the Rebaris of Mārwār are cattle breeders.

The Rājputs form, though representing only a small percentage of the population, the most important section of the population in Rājastān. They are proud of their warlike reputation and of their ancestry. The Brahmin class is subdivided into many *gotras,* while the Mahājan (the trading class) is subdivided into a bewildering number of groups. Some of these groups are Jains, some Hindus. In the north and west the Jāṭs and Gujars are among the largest agricultural communities.

The four main dialects

Linguistic patterns. The principal language of the state is Rajasthani, comprising a group of Indo-Aryan dialects derived from Dingal, a tongue in which bards once sang of the glories of their masters. The four main dialects are Marwari (in western Rājasthān), Jaipuri or Dhundari (in the east and southeast), Malwi (in the southeast), and, in Alwar, Mewati that shades off into Braj Bhasa in Bharatpur district. The use of Rajasthani is declining with the spread of modem education, its place being taken by Hindi.

Religious affiliations. Hinduism is the religion of much of the population and takes the form of the worship of Brahmā, Siva, Sakti, Viṣṇu, and other gods and goddesses. There are also followers of the Arya Samaj, a reforming sect of modern Hinduism, and of other forms of that religion. Jainism (*q.v.*) is also important; it has not been the religion of the rulers of Rājasthān but has its followers among the trading class and the wealthy section of society. Mahāvīrji, Ranakpur, Dhulev, and Karera are the chief centres of Jain pilgrimage. Another important religious sect is formed by the *Dādūpanthīs*, the followers of Dādū (d. 1603) who preached the equality of all men, strict vegetarianism, total abstinence from intoxicating liquor, and life-long celibacy.

Islām expanded in Rājasthān with the conquest of Ajmer by Muslim invaders in the late 12th century. Khwājah Mu'in-ud-Din Chishti, the Muslim missionary, had his headquarters at Ajmer; and Muslim traders, craftsmen, and soldiers settled there. There are also a few Christians and Sikhs in the state.

Rural and urban settlements. Rājasthān is one of the least densely populated states in India. There are 195 persons per square mile. Of the 32,000 villages and 145 towns and cities, the greater number lie east of the Arāvallis. The urban population has been growing faster than the rural, but even so there are only seven towns with more than 100,000 inhabitants: Jaipur, Ajmer, Jodhpur, Udaipur, Kota, Bikaner, and Alwar. There are no industrial complexes.

Rural houses are huts with mud walls and roofs thatched with straw. They have a single door, but no windows or ventilators. The houses of well-to-do farmers and artisans in larger villages have more than one room. They are roofed with tiles and have a veranda and large courtyard, whose main door will admit a loaded bull cart. The earthen floors are coated with mud and dung.

Administration **and** social conditions. Rājasthān state is headed by a governor, appointed by the president of the Indian Union, whom he represents, for a five-year term; he exercises administrative, legislative, financial, and judicial powers. The powers of the state government are controlled by the Legislative Assembly, which is unicameral and consists of 184 elected members; 52 seats are reserved for representatives of the scheduled castes and tribes. The state is divided into 26 districts: Ajmer, Alwar, Biinswiira, Barmer, Bharatpur, Bhīlwāra, Bīkaner, Biindi, Chitorgarh, Churu, Dūngarpur, Gangānagar, Jaipur, Jaisalmer, Jālor, Jhālawār, Jhiinjhunu, Jodhpur, Kota, Nāgaur, Pāli, Sawai Mādhopur, Sikar, Sirohi, Tonk, and Udaipur.

In each district the collector, who is also the district magistrate, is the principal representative of the administration. He functions in close cooperation with the superintendent of police to maintain law and order in the district; he is the principal revenue officer; he coordinates the activities of the other departments in the district; and acts as a link between the state government and people.

Rule by *pañcāyat*

RBjasthBn was the first state to try the experiment of *pañcāyat rāj* (rule by *pañcāyat*, or village committee); the system has elicited a response among the villages. In 1968–69 there were about 33,000 educational establishments in Rājasthān, including the three universities of Jaipur, Udaipur, and Jodhpur and the Birla Institute of Science and Technology at Pilani. There were in all about 3,000,000 students and 89,000 teachers. There are 400 state hospitals and dispensaries. There are also many Ayurvedic and Unani (a medicinal system using prescribed herbs and shrubs) institutions. The state incurs heavy expenditure on education, on maternity and child welfare, on rural and urban water supplies, and on the welfare of the backward classes.

The economy. By the end of the third five-year plan (1966) Rs. 3,650,000,000 (Rs. 7.58 = $1 U.S.; Rs. 18.19 = £1 sterling) has been spent on development schemes. The techno-economic survey of 1962 envisaged per capita income nearly doubling in the period from 1960–61 to the early 1970s.

Agriculture. Rājasthān is a predominantly agricultural and pastoral state. Against all probability, food production was increased by about 54 percent during the 12 years 1955–56 to 1967–68 and instead of being a food deficit state as hitherto, Rājasthān began to export food grains and vegetables. Despite a low and erratic rainfall, nearly all types of crops are grown; in the desert area, *bājrā* (millet); in Kota *jowār* (sorghum); in Udaipur mainly maize. Wheat and barley are fairly well distributed, except in the desert area, as are pulses (the edible seeds of leguminous plants, such as peas, beans, and lentils) and oil seeds. Improved varieties of rice have been introduced, and this crop was expected to be expanded in the areas of the Chambal and Rājasthān Canal projects. Cotton is an important cash crop in the south and north. The use of improved agricultural implements is increasing.

Animal products

Although more than half its area is arid or semi-arid, RBjasthBn has a large livestock population in comparison with the rest of India and is the largest wool-producing state. It has a monopoly in camels and in draught animals of various breeds.

Irrigation. Having much arid land, RBjasthBn needs extensive irrigation works. By the early 1970s about 11 percent of the cultivated area was irrigated, more than half by wells and about one-third by canals. RBjasthBn already received 2,870,000 acre-feet of water from the Punjab rivers and was to receive more. It was also claiming a share from the Gurgaon (Haryana) and Agra canals (Uttar Pradesh) and from the Narmada River to the south. These resources, properly utilized, would bring one-third of the cultivated land under irrigation. There are thousands of tanks (village ponds or lakes), but they suffer from drought and silt. In about 1969 the Rājasthān Underground Water Board found a large water supply near Jaisalmer. Rājasthān shares the Bhākra Nangal project with the Punjab and the Chambal Valley project with Madhya Pradesh; both are used to supply water for irrigation as well as for drinking purposes. The Rājasthān Canal project is the largest in the state.

Industries. When the state was formed in 1949 there were only about 900 registered industrial units, but after the establishment of the Directorate of the Economic and Industrial Survey and the Directorate of the Small Industries Service Institute the number of factories rose to more than 2,000 (1969). The main industries are based on textiles, oil, wool, minerals, and chemicals, while the handicrafts formerly patronized by the princely courts have been earning much foreign exchange. Various industrial concerns received substantial loans and subsidies from the government and from the Rājasthān Finance Corporation, a semi-governmental agency.

Kota, which by the early 1970s was becoming the industrial capital of the state, has a nylon factory and a precision-instruments factory, as well as plants for the manufacture of calcium carbide, caustic soda, and rayon tire cord. There is a zinc smelter plant near Udaipur.

Consumption of electricity rose from 274,350,000 kilowatt-hours in 1963–64 to 417,608,000 kilowatt-hours in 1967–68. Electricity supplies are obtained from neighbouring states as well as from the Chambal Valley project. The atomic-energy plant at Rāwatbhāta was due to be commissioned in the early 1970s.

Exploration for oil

Exploration for oil was taking place in the Jaisalmer district in the early 1970s. New mineral deposits recently discovered include copper, zinc, limestone, gypsum, lignite, and marble.

Transportation and communications. The total length of roads of all categories in 1968–69 was 19,500 miles (31,380 kilometres) of which 3,750 miles (6,030 kilometres) were fair-weather roads. Strategic roads were being built in the western sector. There were 81,000 motor vehicles in the state in 1969.

There are 3,868 miles of railway track in the state, of which 3,615 miles are meter gauge. Air services connect Jaipur, Jodhpur, Kota, and Udaipur with Bombay, Delhi, and Agra. Telephone and telegraph facilities are provided at all subdivisional and *tahsil* (district administration) headquarters and at about 100 other centres with a population of 5,000 or more.

Cultural life. Hardly a month passes in Rājasthān without a religious festival. The most remarkable and typical is the festival called Gangor, when clay images of Mahādevī and Pārvatī (representing the benevolent aspects of the Hindu mother goddess) are worshipped by women of all castes for 15 days and are then taken out to be immersed in water. Their procession is joined by priests and officers and is led to the water by trumpeters and drummers. Hindus and Muslims join in each others' festivals; general enthusiasm and gaiety prevail on these occasions.

Another important festival is held at Pushkar near Ajmer, taking the form of a mixed religious festival and livestock fair. It is visited by farmers from all over the state, bringing their camels and cattle, and by pilgrims seeking salvation. The average number of visitors ranges between 100,000 and 300,000.

The typical folk dance of Rājasthān is the *ghoomar*, which is performed on festive occasions only by women. The *geer* dance (performed by men and women), the *panihari* (a graceful dance for women), and the *kacchi ghori* (in which male dancers ride dummy horses) are also popular. The most famous song is the *Kurja*, which tells the story of a woman who wishes to send a message to her absent husband by the *kurja* (a bird), who is promised a priceless reward for his service. RBjasthBn has made its contribution to Indian art, and there is a rich literary tradition, especially of bardic poetry. Chand Bardai's poem, *Prithvi Raj Rasak*, the earliest manuscript of which dates back to the 16th century, is particularly notable. A popular source of entertainment is the *khyāl*, a dance drama composed in verse with gay, historical, or romantic themes.

Rājasthān abounds in objects of antiquarian interest: early Buddhist rock inscriptions, Jain temples, forts, splendid princely palaces, and Muslim mosques and tombs.

BIBLIOGRAPHY. General works include V.C. MISRA, *Geography of Rajasthan* (1967); DHARAM PAL, *Rajasthan* (1968); and the *Imperial Gazetteer of India, Provincial Series, Rajasthan* (1968). S.P. RAYCHAUDHURI *et al.*, *Soils of India* (1963), contains mineralogical studies and detailed soil profiles for various districts of the state of Rājasthān. For history and social conditions, see J. TOD, *Annals and Antiquities of Rajasthan*, 3 vol. (1829–32; 2 volume edition, 1914); R.E.U. BISHESHWAR, *Glories of Marwar and the Glorious Rathors* (1943); G.N. SHARMA, *Social Life in Mediaeval Rajasthan, 1500–1800 A.D.* (1968); V.P. MENON, *The Story of the Integration of the Indian States* (1956); IQBAL NARAIN, *Panchayati Raj: Planning and Democracy in Pakistan* (1969); and the Rajasthan Planning Department's *Five-Year Plans*. See also A.K. COOMARSWAMI, *Rajput Painting* (1916).

(I.P.)

Raleigh, Sir Walter

Sir Walter Raleigh (Ralegh), an English soldier, seaman, courtier, writer, explorer, and favourite of Queen Elizabeth I, was also an early English colonizer in America.

A younger son of Sir Walter Raleigh (died 1581) of

GUALTHERI RALEGH EQV: AUR:
VERA EFFIGIES CLARISS: MI VIRI DOM: NI

Raleigh, engraving by Simon Pass from the title page of the first edition of Raleigh's History of the World, 1614.

By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.

Fardell in Devon, by his third wife, Katherine Gilbert (*née* Champernowne), Raleigh was born in 1554 at Hayes Barton near Budleigh Salterton, Devon. In 1569 Raleigh fought on the Huguenot (French Protestant) side in the Wars of Religion in France and later is known to have been at Oriel College, Oxford (1572), and at the Middle Temple (law college, 1575). In 1580 he fought against the Irish rebels in Munster, and his outspoken criticism of the way English policy was being handled in Ireland brought him to the attention of Queen Elizabeth. By 1582 he was the reigning favourite and began to acquire lucrative monopolies, properties, and influential positions. His Irish service was rewarded by vast estates in Munster, where his name is associated with Youghal, Cork, and Lismore. In 1583 the Queen secured him a lease of part of Durham House in the Strand, London, where he had a monopoly of wine licenses (1583) and of the export of broadcloth (1585), and he became warden of the stannaries (the Cornish tin mines), lieutenant of Cornwall, vice admiral of Devon and Cornwall, and frequently sat as a member of Parliament. In 1585 Raleigh was knighted and in 1587 became captain of the queen's guard. His last appointment under the crown was as governor of Jersey (one of the Channel Islands) in 1600.

Effectof his marriage on the Queen

In 1592 Raleigh acquired the manor of Sherborne in Dorset. He wanted to settle and found a family. His marriage to Elizabeth, daughter of Sir Nicholas Throckmorton, possibly as early as 1588, had been kept a secret from the jealous queen. In 1592 the birth of a son betrayed him, and he and his wife were both imprisoned in the Tower of London. Raleigh bought his release with profits from a privateering voyage in which he had invested, but he never regained his ascendancy at court. The child did not survive; a second son, Walter, was born in 1593, and a third son, Carew, in 1604 or 1605.

As a favourite, Raleigh was not popular. His pride and extravagant spending were notorious, and he was attacked for unorthodox thought. A Jesuit pamphlet in 1592 accused him of keeping a "School of Atheism," but he was not an atheist in the modern sense. He was a bold talker, interested in skeptical philosophy, and a serious student of mathematics as an aid to navigation. He also studied chemistry and compounded medical formulae. Some scholars have linked "School of Atheism" with a vaguely similar phrase, the "School of Night," which occurs in *Love's Labour's Lost,* and have read Shakespeare's play as a satire on Raleigh.

Raleigh's breach with the Queen widened his personal sphere of action. Between 1584 and 1589 he had tried to establish a colony near Roanoke Island (in present North Carolina), which he named Virginia; but he never set foot there himself. In 1595, he led an expedition to what is now Guyana, in South America, sailing up the Orinoco River in the heart of Spain's colonial empire. He described the expedition in his book *The Discoverie of Guiana* (1596). Spanish documents and stories told by Indians had convinced him of the existence of Eldorado, a fabulous city of gold in the interior of South America. He did locate some gold mines, but no one supported his project for colonizing the area. In 1596 he went with Robert Devereux, earl of Essex, on an unsuccessful expedition against the Spanish city of Cádiz, and he was Essex' rear admiral on the Islands voyage, in 1597, an expedition to the Azores in the West Indies.

Raleigh's aggressive policies toward Spain did not recommend him to the pacific King James I (reigned 1603–25), Elizabeth's successor. His enemies worked to bring about his ruin, and in 1603 he and others were accused of plotting to dethrone the King. Raleigh was convicted on the written evidence of Henry Brooke, Lord Cobham, and, after a last-minute reprieve from the death sentence, was consigned to the Tower. He fought to save Sherborne, which he had conveyed in trust for his son, but a clerical error invalidated the deed. In 1616 he was released but not pardoned. He still hoped to exploit the wealth of Guyana, arguing that the country had been ceded to England by its native chiefs in 1595. With the King's permission, he financed and led a second expedition there, promising to open a gold mine without offending Spain. A severe fever prevented his leading his men upriver. His lieutenant, Lawrence Kemys, burned a Spanish settlement but found no gold. Raleigh's son Walter died in the action. King James invoked the suspended sentence of 1603, and on October 29, 1618, after writing a spirited defense of his acts, Raleigh was executed.

Writings

Popular feeling had been on Raleigh's side ever since 1603. After 1618 his occasional writings were collected and published, often with little discrimination. The authenticity of some minor works attributed to him is still unsure. Some 560 lines of verse in his hand are preserved. They address the queen as Cynthia and complain of her unkindness, probably with reference to his imprisonment of 1592. His best known prose works in addition to *The Discoverie of Guiana,* are *A Report of the Truth of the fight about the Iles of Açores this last Sommer* (1591; generally known as *The Last Fight of the Revenge)* and *The History of the World* (1614). The last work, undertaken in the Tower, proceeds from the Creation to the 2nd century BC. In the work history is shown as a record of God's providence, a doctrine that pleased contemporaries and counteracted the charge of atheism. King James was meant to note the many warnings that the injusticeof kingsis always punished.

Raleigh survives as an interesting and enigmatic personality rather than as a force in history. He can be presented either as a hero or as a scoundrel. His vaulting imagination, which could envisage both North and South America as English territory, was supported by considerable practical ability and a persuasive pen, but some discrepancy between the vision and the deed made him less effective than his gifts had promised.

BIBLIOGRAPHY

Editions: The Works of Sir Walter Ralegh, 8 vol. (1829); *The Discoverie . . . of Guiana,* ed. by V.T. HARLOW (1928); *The Poems of Sir Walter Ralegh,* ed. by AGNES LATHAM, 2nd ed. (1951); *The History of the World,* ed. by C.A. PATRIDES (1971), representative selections; see also T.N. BRUSHFIELD, *A Bibliography of Sir Walter Raleigh,* 2nd ed. (1908).

Biographies: WILLIAM OLDYS, *The Life of Sir Walter Ralegh* (1736, reprinted 1829), was the first reputable biography; EDWARD EDWARDS, *Life . . . ,* 2 vol. (1868), is well-documented and includes a large collection of letters. See also WILLIAM STEBBING, *Sir Walter Ralegh,* 2nd ed. (1899), a most valuable study; W.M. WALLACE, *Sir Walter Raleigh* (1959), a sober and factual account; A.L. ROWSE, *Ralegh and the Throckmortons* (1962), which established the birth date and legitimacy of the first son; J.H. ADAMSON and H.F. FOLLAND, *The Shepherd of the Ocean: An Account of Sir Walter Ralegh and His Times* (1969), a recent interpretation, lively and sympathetic; and PHILIP AHIER, *The Governorship of Sir Walter Ralegh in Jersey, 1600–1603* (1971), a learned biographical study.

Special aspects: For colonial exploits, see V.T. HARLOW, *Ralegh's Last Voyage* (1932); and D.B. QUINN, *Raleigh and* *the British Empire* (1947). Raleigh's thought is discussed by E.A. STRATHMANN in *Sir Walter Ralegh: A Study in Elizabethan Skepticism* (1951); and by CHRISTOPHER HILL in *Intellectual Origins of the English Revolution* (1965). For Raleigh as historian, see C.H. FIRTH in *Essays Historical and Literary* (1938); and F. SMITH FUSSNER in *The Historical Revolution* (1962). PIERRE LEFRANC, *Sir Walter Ralegh écrivain: l'oeuvre et les idées* (1968), gives an exhaustive account of Raleigh as writer and thinker, with a full and up-to-date bibliography.

(A.M.C.L.)

Rāmānuja

Early life and development

Rāmānuja, also Rāmānujācārya (Tamil, Iḷeya Perumāḷ), a south Indian brahmin, theologian, and philosopher, was the single most influential thinker of devotional Hinduism. Information on his life consists only of the accounts given in the legendary biographies about him, in which a pious imagination has embroidered historical details. Rāmānuja was born according to tradition about 1017 in Śrīperumbūdūr near Kāñcī (Kānchipuram) in southern India, in what is now Tamil Nadu (formerly Madras) state. He showed early signs of theological acumen and was sent to Kāñcī for schooling, under the teacher Yādavaprakāśa, who was a follower of the monistic system of Vedanta of Sankara, the famous 8th-century philosopher. Rāmānuja's profoundly religious nature was soon at odds with a doctrine that offered no room for a personal god. After falling out with his teacher he had a vision of the god Viṣṇu and his consort Śrī, or Lakṣmī, and instituted a daily worship ritual at the place where he beheld them. He became a temple priest at the Varadarāja temple at Kāñcī, where he began to expound the doctrine that the goal of those who aspire to final release from transmigration is not the impersonal Brahman but rather Brahman as identified with the personal god Viṣṇu. In Kāñcī, as well as Srirangam, where he was to become head priest at the Raṅganātha temple, he developed the teaching that the worship of a personal God and the soul's union with him is an essential part of the doctrines of the *Upaniṣads* (ancient speculative texts that are part of Hindu sacred scriptures) on which the system of Vedanta is built; therefore, the teachings of the Vaiṣṇavas and Bhāgavatas (worshippers and ardent devotees of Viṣṇu) are not heterodox. In this he continued the teachings of Yāmuna (Yāmunācārya; 10th century), his predecessor at Śrīraṅgam, to whom he was related on his mother's side. He set forth his doctrine in three major works, the *Veddrtha-samgraha, Śrī-bhāṣya,* and *Bhagavadgitd-bhāṣya.*

By courtesy of the Institut Francais d'Indologie, Pondicherry

Rāmānuja, bronze sculpture, 12th century. From a Vishnu temple in Tanjore district, India

Like many Hindu thinkers, he made an extended pilgrimage, circumambulating India from Rāmeśvaram (Adams Bridge), along the west coast to Badrīnāth, the source of the holy river Ganges, and returning along the east coast. Tradition has it that later he suffered from the zeal of King Kulottunga of the Cōla dynasty, who adhered to the god Śiva, and withdrew to Mysore, in the west. There he converted numbers of Jains (adherents of a dualistic, ascetic sect), as well as King Bittideva of the Hoyśala dynasty; this led to the founding in 1099 of the town Milukote (Melcote, Mysore state) and the dedication of a temple to Śelva Piḷḷai (Sanskrit, Saṃpatkumāra, the name of a form of Viṣṇu). He returned after 20 years to Śrīraṅgam, where he organized the temple worship, and, reputedly, he founded 74 centres to disseminate his doctrine. After a life of 120 years, according to the tradition, he passed away in 1137.

Rāmānuja's chief contribution to philosophy was his emphasis that discursive thought is necessary in man's search for the ultimate verities, that the phenomenal world is real and provides real knowledge, and that the exigencies of daily life are not detrimental or even contrary to the life of the spirit. In this emphasis he is the antithesis of Sankara, of whom he was sharply critical and whose interpretation of the scriptures he disputed throughout his life. Like other adherents of the Vedānta system, Rāmānuja accepted that any Vedanta system must base itself on the three "points of departure," viz., the *Upaniṣads*, the *Brahma-sūtras* (brief exposition of the major tenets of the *Upaniṣads*), and the *Bhagavadgitd,* the colloquy of the god Kṛṣṇa and his friend Arjuna. He wrote no commentary on any single *Upaniṣad* but explained in detail the method of understanding the *Upaniṣads* in his first major work, the *Vedārtha-saṃgraha* ("Summary of the Meaning of the Veda"). Much of this was incorporated in his commentary on the *Brahma-sūtras*, the *Śrībhāṣya*, which presents his fully developed views. His commentary on the *Bhagavadgītā*, the *Bhagavadgītā-bhāṣya*, dates from a later age.

Influence on Hinduism

Although Rāmānuja's contribution to Vedānta thought was highly significant, his influence on the course of Hinduism as a religion has been even greater. By allowing the urge for devotional worship *(bhakti)* into his doctrine of salvation, he aligned the popular religion with the pursuits of philosophy and gave *bhakti* an intellectual basis. Ever since, *bhakti* has remained the major force in the religions of Hinduism. His emphasis on the necessity of religious worship as a means of salvation continued in a more systematic context the devotional effusions of the Āḷvārs, the 7th–8th century poet-mystics of southern India, whose verse became incorporated into temple worship. This *bhakti* devotionalism, guided by Rāmānuja, made its way into northern India, where its influence on religious thought and practice has been profound.

Rāmānuja's world view accepts the ontological reality of three distinct orders: matter, soul, and God. Like Sankara and earlier Vedanta, he admits that there is nonduality *(advaita),* an ultimate identity of the three orders, but this nonduality for him is asserted of God, who is modified (*viśiṣṭa*) by the orders of matter and soul; hence his doctrine is known as Viśiṣṭādvaita ("modified nonduality") as opposed to the unqualified nonduality of Śaṅkara. Central to his organic conception of the universe is the analogy of body and soul: just as the body modifies the soul, has no separate existence from it, and yet is different from it, just so the orders of matter and soul constitute God's "body," modifying it, yet having no separate existence from it. The goal of the human soul, therefore, is to serve God just as the body serves the soul. Anything different from God is but a *śeṣa* of him, a spilling from the plenitude of his being. All the phenomenal world is a manifestation of the glory of God (*vibhūti*), and to detract from its reality is to detract from his glory. Rāmānuja transformed the practice of ritual action into the practice of divine worship and the way of meditation into a continuous loving pondering of God's qualities; both in turn are subservient to *bhakti,* the fully realized devotion that finds God. Thus, release is not merely a shedding of the bonds of transmigration but a positive quest for the contemplation of God, who is pictured as enthroned in his heaven, called Vaikuṇṭha, with his consorts and attendants.

Rāmānuja's doctrine, which was passed on and augmented by later generations, still identifies a caste of Brahmins in southern India, the Śrīvaiṣṇavas. They became divided into two subcastes, the northern, or Vadakalai, and the southern, or Teṉkalai. At issue between the two schools is the question of God's grace. According to the Vadakalai, who in this seem to follow Rāmānuja's intention more closely, God's grace is certainly active in man's quest for him but does not supplant the necessity of man's acting toward God. The Teṉkalai, on the other hand, hold that God's grace is paramount and that the only gesture needed from man is his total submission to God ***(prapatti)***.

The site of Rāmānuja's birthplace in Śrīperumbūdtūr is now commemorated by a temple and an active Viśiṣṭādvaita school. The doctrines he promulgated still inspire a lively intellectual tradition, and the religious practices he emphasized are still carried on in the two most important Vaiṣṇava centres in southern India, the Raṅganātha temple in Śrīraṅgam, and the Veṅkateśvara temple in Tirupati, both in Tamil Nadu.

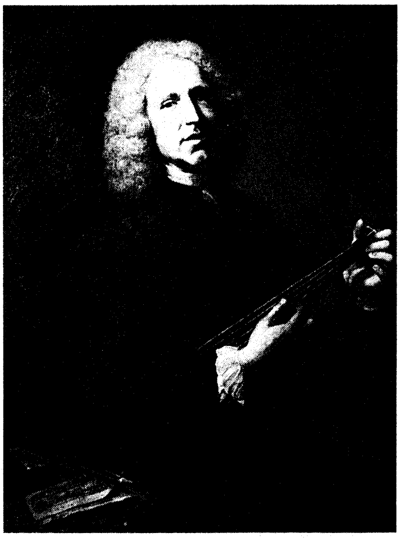
BIBLIOGRAPHY. There are several traditional accounts of Rāmānuja's life. The oldest, on which the others largely rest, is that by ANANTACARYA, entitled the *Prapannāmṛta*, which by some is dated to shortly after Rāmānuja's lifetime. Based on the *Prapannāmṛta* in large parts are the accounts by S.N. DAS GUPTA in his ***History of** Indian Philosophy*, vol. *3 **(1940);***and P.N. SRINIVASACHARI in ***Philosophy*** of *Viśiṣṭādvaita*, 2nd ed. ***(1946).***

(J.A.B.v.B.)

Rameau, Jean-Philippe

The 18th-century French composer Jean-Philippe Rameau is best known today for his harpsichord music, but in his lifetime he was famous as a musical theorist and, above all, as a composer of opera. He is often ranked with the great trio of late-Baroque composers—Johann Sebastian Bach, George Frideric Handel, and Domenico Scarlatti—who were his contemporaries. Rameau is also often paired with François Couperin, who died in 1733, the year that Rameau produced the first of his impressive series of 30 operas and opéra ballets—a medium never attempted by any of the Couperin family. Rameau is recognized as the greatest musical dramatist France has produced.

Snark *International*



Rameau, portrait by Jacques-Andre-Joseph Aved (1702–66). In the Musée des Beaux-Arts, Dijon, France.

Rameau was baptized at Dijon on September 25, 1683. His father, Jean, played the organ for 42 years in various churches in Dijon and hoped one day to see his son on a lawyer's, rather than an organist's, bench. These hopes were dashed by the boy's deplorable performance in school. At the age of 17 he is said to have fallen in love with a young widow who laughed at the errors of grammar and spelling in his letters to her. He tried to refine his language, but, to judge by the prolixity of his later theoretical writings, his efforts resulted in no permanent improvement. At the age of 18, after deciding to pursue a musical career, he travelled to Italy but seems to have gotten no farther than Milan. The following year, he received the first of a series of appointments as organist in various cities of central France: Avignon, Clermont, Dijon, Lyon. There was a brief interlude in the capital, but apparently Paris did not take an immediate fancy to the provincial organist, in spite of his having published there a fine suite of harpsichord pieces in A minor, ***Premier Livre de pièces de clavecin*** (1706). These works show the beneficial influence of Louis Marchand, a famous organist-harpsichordist of the day whose playing Rameau greatly admired.

Organist in the provinces

Back in Clermont by 1715, he rashly signed a contract to be cathedral organist for 29 years. He then settled down to investigate, in an exhaustive and highly original manner, the foundations of musical harmony. He attacked traditional theory on the ground that "The Ancients," who to Rameau included such relatively recent writers as the 16th-century Italian Gioseffo Zarlino, ". . . based the rules of harmony on melody, instead of beginning with harmony, which comes first." Intuitively basing his studies on the natural overtone series, he arrived at a system of harmony that is the basis of most 20th-century harmony textbooks. Finally published in Paris in 1722, his impressive ***Traitk de l'harmonie*** brought him fame at last and a yearning to return to the capital.

Authorities in Clermont were loath to let him go, and the story of his release reveals, as do his own writings and other evidence, something of his thorny personality, his persistence, and his single-mindedness. At an evening service he showed his displeasure with the church authorities by pulling out all the most unpleasing stops and by adding the most rending discords so that "connoisseurs confessed only Rameau could play so unpleasingly." But, after his release from the contract, he played with "so much delicacy, brilliance, force and harmony, that he aroused in the souls of the congregation all the sentiments he wished, thereby sharpening the regret with which all felt the loss they were about to sustain."

Upon his return to Paris, where he was to remain for the rest of his life, Rameau began a new and active life. A second volume of harpsichord pieces, ***Pièces de clavecin avec une méthode pour la mècanique des doigts*** (1724), met with considerably more success than the first, and he became a fashionable teacher of the instrument. A commission to write incidental music for the Fair theatres planted the seeds of his development as a dramatic composer, and the display of two Louisiana Indians at one of these theatres in 1725 inspired the composition of one of his best and most celebrated pieces, ***Les Sauvages,*** later used in his optra ballet ***Les Indes galantes*** (first performed, 1735). The following year, at the age of 42, he married a 19-year-old singer, who was to appear in several of his operas and who was to bear him four children.

His most influential contact at this time was Le Riche de la Pouplinibre, one of the wealthiest men in France and one of the greatest musical patrons of all time. Rameau was put in charge of La Pouplinikre's excellent private orchestra, a post he held for 22 years. He also taught the financier's brilliant and musical wife. The composer's family eventually moved into La Pouplinikre's town mansion and spent summers at their château in Passy. This idyllic relationship between patron and composer gradually came to an end after La Pouplinikre separated from his wife, and Rameau was replaced by the younger, avant-garde composer Karl Stamitz. In the meantime, however, admittance to La Pouplinibre's circle had brought Rameau into contact with various literary lights. Abbé Pellegrin, whose biblical opera ***Jephtk*** had been successfully set to music by Rameau's rival Michel Pinolet de Montéclair in 1732, was to become Rameau's librettist for his first and in many ways finest opera, ***Hippolyte et Aricie.*** It was first performed in the spring of 1733, at La Pouplinibre's house, then, in the autumn, at the Opéra, and the following year it was performed at court. André Campra, perhaps the most celebrated French composer living at that time, remarked to the Prince de Conti: "My Lord, there is enough music in this

Rameau's first opera

opera to make ten of them; this man will eclipse us all."

To some ears there was, indeed, too much music. Those who had grown up with the operas of Jean-Baptiste Lully were baffled by the complexity of Rameau's orchestration, the intensity of his accompanied recitatives (speech-like sections), and the rich and often dissonant diversity of his harmonies. Rameau himself, however, professed his admiration for his predecessor in the preface to *Les Indes galantes*, in which he praised the "beautiful declamation and handsome turns of phrase in the recitative of the great Lully," and stated that he had sought to imitate it, though not as a "servile copyist." Indeed, almost everything in Rameau's operas has, at least technically, a precedent in Lully. Yet the content of his works, the rich dramatic contrasts, the brilliant orchestral sections, and, above all. the permeating sensuous melancholy and languorous pastoral sighings, put him in a different world: in short, the Rococo world of Louis XV.

Collaboration with Voltaire

Among those at the first performance of *Hippolyte* was the great Voltaire, who quipped that Rameau "is a man who has the misfortune to know more music than Lully." But he soon came around to Rameau's side and wrote for him a fine libretto, *Samson*, which was banned ostensibly for religious reasons but really because of a cabal against Voltaire: the music was lost. Their later collaboration on two frothy court entertainments is preserved, however: *La Princesse de Navarre* and *Le Temple de la Gloire* (both 1745). The former was condensed and revised as *Les Fêtes de Ramire* (1745) by Jean-Jacques Rousseau.

Rousseau, Jean Le Rond d'Alembert, and other writers associated with Denis Diderot's *Encyclopédie* began as ardent Rameau enthusiasts, but, by the mid-1750s, as they warmed more and more to Italian music, they gradually turned against him. Rameau appreciated the new Italian music as much as anyone, but the works he consciously composed in this style, such as the overtures to *Les Fêtes de Polymnie* (1745) and to his final work *Abaris ou les Boréades* (1764), are not very individual. He died on September 12, 1764, in Paris.

The zenith of his career may be said to have encompassed the brief span from 1748, when he tossed off the masterpiece *Pygmalion* in eight days and had six other operas on the boards, through 1754, when he wrote La *Naissance d'Osiris* for the birth of the future Louis XVI. Thereafter, his fame diminished, as the prevailing musical style became what is now generally called "Classical." The public preferred catchy tunes with simple harmonies to Rameau's profound emotion and rich, late-Baroque harmony.

MAJOR WORKS

DRAMATIC WORKS: 35 including *Hippolyte et Aricie* (first performed 1733), *Les Indes galantes* (1735), *Castor et Pollux* (1737), *Dardanus* (1739), *Platée* (1745), and *Zoroastre* (1749).

CHAMBER MUSIC: *Pièces de clavecin en concerts* (published 1741).

HARPSICHORD: 3 volumes of pieces under varying titles. Most have either dance titles such as "Allemande" and "Gigue," or illustrative ones such as "Les Tendres Plaintes" and "La Poule."

BIBLIOGRAPHY. JEAN-PHILIPPE RAMEAU, *Traité de l'harmonie* (1722; Eng. trans., *Treatise on Harmony*, with an Introduction and Notes by PHILIP GOSSETT, 1971), a careful translation of Rameau's first great theoretical work; CUTHBERT M. GIRDLESTONE, *Jean-Philippe Rameau: sa vie, son oeuvre* (1962; Eng. trans., *Jean-Philippe Rameau: His Life and Work*, rev. ed., 1969), by far the finest study of the man and his music yet to have appeared; and "Rameau's Self-Borrowings," *Music and Letters*, 39:52–56 (1958), a brief account of a fascinating topic; KATHLEEN DALE, "The Keyboard Music of Rameau," *Monthly Musical Record*, pp. 127–131 (1946), a somewhat out-of-date but elegant introduction to the *Pièces de clavecin;* JOAN FERRIS, "The Evolution of Rameau's Harmonic Theories," *Journal of Music Theory*, 3:231–256 (1959), a highly technical but succinct summary of all Rameau's theoretical writings; ERWIN R. JACOBI, "Rameau and Padre Martini," *Musical Quarterly*, 50:452–475 (1964), a discussion of how Rameau's theories were received in Italy, includes facsimile of autograph letters; SISTER MICHAELA MARIA KEANE, *The Theoretical Writings of Jean-Philippe Rameau* (1961), a thorough discussion of all the theoretical works, including translations of many excerpts and a brief biographical sketch; PIERRE LASSERRE, *L'Esprit de la musique française* (1917; Eng. trans., *The Spirit of French Music*, 1921), an old-fashioned but sympathetic defense of Rameau from his enemies over the ages and an account of his music, especially *Castor;* PAUL-MARIE MASSON, *L'Opera de Rameau* (1930), the most detailed study (in French) of Rameau's music; and "Rameau and Wagner," *Musical Quarterly*, 25:466–478 (1939), what Wagner chose to admire in Gluck is shown to have been foreshadowed by Rameau; WILFRID MELLERS, "Rameau and the Opera," *The Score*, no. 4, pp. 26–51 (1951), an excellent brief introduction to the music of Rameau with special emphasis on his use of harmony in the operas; F. BRUGGEN, S. and W. KUIJKEN, and G. LEONHARDT, *Rameau: Pièces de clavecin en concerts* (Telefunken SAWT 9578–B, 1971), the only recording to combine the use of completely authentic instruments with admirable virtuosity and style.

(A.S.Cu.)

Ramses II the Great

Until the discovery of the tomb of King Tutankhamen in 1922, Ramses II was the only ancient Egyptian king whose name, apart from those of the builders of the Pyramids of Giza, was familiar to the nonarchaeological public, and, to this day, its bearer remains a prominent figure in the long history of ancient Egypt.

By courtesy of the trustees of the British Museum

Ramses II, upper portion of a granite figure from Thebes, 1250 BC. In the British Museum.

King Ramses II was the third king of the 19th dynasty of Egypt; his reign from 1304 to 1237 BC, was the second longest reign in Egyptian history. His family, of nonroyal origin, came to power after the reign of the religious reformer, Akhenaton (Amenhotep IV, 1379–62 BC), and set about restoring Egyptian power in Asia, which had declined under Akhenaton and his successor, Tutankhamen. Ramses' father, Seti I, subdued a number of rebellious princes in Palestine and southern Syria and waged war on the Hittites of Anatolia in order to recover those provinces in the north that during the recent troubles had passed from Egyptian to Hittite control. Sethos achieved some success against the Hittites at first, but his gains were only temporary, for at the end of his reign the enemy were firmly established at Kadesh, on the Orontes River, a strong fortress defended by the river, which became the key to their southern frontier. Early in his reign Seti made the crown prince Ramses, the future Ramses II, coregent with him, giving him a kingly household and harem, and the young prince accompanied his father on his campaigns, so that when he came to sole rule he had already had experience of kingship and of war. It would appear, however, that Ramses was not the eldest son, for in a relief at Karnak of his father's Libyan war, the figure of a prince whose name is not preserved was inserted into the scene after it had been completed, but the figure was later erased and that of Ramses substituted for it. What lay behind these events is not known, but it is noteworthy that Ramses was

Coregency

crowned coregent at an unusually early age, as if to ensure that he would in fact succeed to the throne. He ranked as a captain of the army while still only ten years old; at that age his rank must surely have been honorific, though he may well have been receiving military training.

Because his family's home was in the Nile Delta and in order to have a convenient base for campaigns in Asia, Ramses built for himself a full-scale residence city called Per-Ramesse ("House of Ramses"; Biblical Raamses), which was famous for its beautiful layout, with gardens, orchards, and pleasant waters. Each of its four quarters had its own presiding deity: Amon in the west, Seth in the south, the royal cobra goddess, Buto, in the north, and, significantly, the Syrian goddess Astarte in the east. A vogue for Asian deities had grown up in Egypt, and Ramses himself had distinct leanings in that direction. The first public act of Ramses after his accession to sole rule was to visit Thebes, the southern capital, for the great religious festival of Taurt (Opet), when the god Amon of Karnak made a state visit in his ceremonial barge to the temple of Luxor. When returning to his home in the north, the King broke his journey at Abydos to worship Osiris and to arrange for the resumption of work on the great temple founded there by his father, which had been interrupted by the old king's death. He also took the opportunity to appoint as the new high priest of Amon at Thebes a man named Nebwenenef, high priest of Anhur at nearby Thinis. It seems that, apart from his extensive building activities and his famous residence city, Ramses' reputation as a great king in the eyes of his subjects rested largely on his fame as a soldier.

The Battle of Kadesh

Military exploits. In the fourth year of his reign, he led an army north to recover the lost provinces his father had been unable to conquer permanently. The first expedition was to subdue rebellious local dynasts in southern Syria, to ensure a secure springboard for further advances. He halted at the Nahr al-Kalb near Beirut, where he set up an inscription to record the events of the campaign; today nothing remains of it except his name and the date; all the rest has weathered away. The next year the main expedition set out. Its objective was the Hittite stronghold at Kadesh. Following the coastal load through Palestine and Lebanon, the army halted on reaching the south of the land of Amor, perhaps in the neighbourhood of Tripolis. Here Ramses detached a special task force, the duty of which seems to have been to secure the seaport of Simyra and thence to march up the valley of the Eleutherus River (Nahr el-Kebir) to rejoin the main army at Kadesh. The main force then resumed its march to the River Orontes, the army being organized in four divisions of chariotry and infantry, each consisting of perhaps 5,000 men. Crossing the river from east to west at the ford of Shabtuna, about eight miles from Kadesh, the army passed through a wood to emerge on the plain in front of the city. Two captured Hittite spies gave Ramses the false information that the main Hittite army was at Aleppo, some distance to the north, so that it appeared to the King as if he had only the garrison of Kadesh to deal with. It was not until the army had begun to arrive at the camping site before Kadesh that Ramses learned that the main Hittite army was in fact concealed behind the city. Ramses at once sent off messengers to hasten the remainder of his forces, but before any further action could be taken, the Hittites struck with a force of 2,500 chariots, with three men to a chariot as against the Egyptian two. The leading divisions, taken entirely by surprise, broke and fled in disorder, leaving Ramses and his small corps of household chariotry entirely surrounded by the enemy and fighting desperately. Fortunately for the King, at the crisis of the battle, the Simyra task force appeared on the scene to make its junction with the main army and thus saved the situation. The result of the battle was a tactical victory for the Egyptians, in that they remained masters of the stricken field, but a strategic defeat in that they did not and could not take Kadesh. Neither army was in a fit state to continue action the next day, so an armistice was agreed and the Egyptians returned home. This battle is one of the very few from Pharaonic times of which there are real details, and that is because of the King's pride in his stand against great odds; pictures and accounts of the campaign, both an official record and a long poem on the subject, were carved on temple walls in Egypt and Nubia, and the poem is also extant on papyrus.

The failure to capture Kadesh had devastating effects on Egyptian prestige abroad, and many of the petty states of South Syria and Palestine under Egyptian suzerainty rebelled, so that Ramses had to start again from the beginning. In the sixth or seventh year of his reign, he stormed Ashkelon; the following year he took a number of towns in Galilee and Amor, and the next year he was again on the Nahr al-Kalb. It may have been in the tenth year that he broke through the Hittite defenses and conquered Katna and Tunip—where, in a surprise attack by the Hittites, he went into battle without his armour—and held them long enough for a statue of himself as overlord to be erected in Tunip. In a further advance he invaded Kode, perhaps the region between Alexandretta and Carchemish. Nevertheless, like his father before him, he found that he could not permanently hold territory so far from base against continual Hittite pressure, and, after 16 years of intermittent hostilities, a treaty of peace was concluded in 1283 BC, as between equal great powers, and its provisions were reciprocal. The wars once over, the two nations established friendly ties. Letters on diplomatic matters were regularly exchanged; in 1270 Ramses contracted a marriage with the eldest daughter of the Hittite king, and it is possible that at a later date he married a second Hittite princess. Apart from the struggle against the Hittites, there were punitive expeditions against Edom, Moab, and Negeb and a more serious war against the Libyans, who were constantly trying to invade and settle in the Delta; it is probable that Ramses took a personal part in the Libyan war but not in the minor expeditions. The latter part of the reign seems to have been free from wars.

The temples of Abu Simbel

Prosperity during his reign. One measure of Egypt's prosperity is the amount of temple building the kings could afford and carry out, and on that basis the reign of Ramses II is the most notable in history, even making allowance for its great length. It was that, combined with his prowess in war as depicted in the temples, that led the Egyptologists of the 19th century to dub him "the Great," and that, in effect, is how his subjects and posterity viewed him; to them he was the king par *excellence.* All the nine kings of the 20th dynasty called themselves by his name; even in the period of decline that followed, it was an honour to be able to claim descent from him, and his subjects called him by the affectionate abbreviation Sese. In Egypt he completed the great hypostyle hall at Karnak (Thebes) and the temple built by Seti I at Abydos, both of which were left incomplete at the latter's death. Ramses also completed his father's funerary temple on the west bank of the Nile at Luxor (Thebes) and built one for himself, which is now known as the Ramesseum. At Abydos he built a temple of his own not far from that of his father; there were also the four major temples in his residence city, not to mention lesser shrines. In Nubia (Nilotic Sudan) he constructed no fewer than six temples, of which the two carved out of a cliffsideat Abu Simbel, with their four colossal statues of the king, are the most magnificent and the best known. The larger of the two was begun under Seti I but was largely executed by Ramses, while the other was entirely due to Ramses. In the Wadi Tumilat, one of the eastern entries into Egypt, he built the town of Per-Atum (Biblical Pithom), which the Bible calls a store city (Ex. 1:11), but which probably was a fortified frontier town and customs station. In fact, there can have been few sites of any importance that originally did not exhibit at least the name of Ramses, for apart from his own work, he did not hesitate to inscribe it on the monuments of his predecessors. Apart from the construction of Per-Ramesse and Pithom, his most notable secular work, so far as is known, was the sinking of a well in the eastern desert on the route to the Nubian gold mines. Of Ramses' personal life virtually nothing is known. His first and perhaps favourite queen was Nefretari; the fact that, at Abu Simbel, the smaller temple was dedicated to her and to the goddess of love points to real affection between them. She seems to have died comparatively early in the reign, and her fine tomb in the Valley of the Tombs of the Queens at Thebes is well known. Other queens whose names are preserved were Isinofre, who bore the king four sons, among whom was Ramses' eventual successor, Merneptah; Merytamun and Matnefrure, the Hittite princess. In addition to the official queen or queens, the King, as was customary, possessed a large harem, and he took pride in his great family of well over 100 children. The best portrait of Ramses II is a fine statue of him as a young man, now in the Turin museum; his mummy, preserved in a mausoleum at Cairo, is that of a very old man with a long narrow face, prominent nose, and massive jaw.

The reign of Ramses II marks the last peak of Egypt's imperial power. After his death Egypt was forced on the defensive but managed to maintain its suzerainty over Palestine and the adjacent territories until the later part of the 20th dynasty, when, under the weak kings who followed Ramses III, internal decay ended its power beyond its borders. Ramses II must have been a good soldier, despite the fiasco of Kadesh, or else he would not have been able to penetrate so far into the Hittite Empire as he did in the following years; he appears to have been a competent administrator, since the country was prosperous, and he was certainly a popular king. Some of his fame, however, must surely be put down to his flair for publicity: his name and the record of his feats on the field of battle were found everywhere in Egypt and Nubia. It is easy to see why, in the eyes both of his subjects and of later generations, he was looked on as a model of what a king should be.

BIBLIOGRAPHY

General works: SIR ALAN H. GARDINER, *Egypt of the Pharaohs* (1961), the most recent general history, based primarily on the written documents; R.O. FAULKNER, "Egypt: From the Inception of the Nineteenth Dynasty to the Death of Ramesses III," *Cambridge Ancient History,* 2nd ed. vol. 2, ch. 23 (1966), with bibliography.

Specialized works: R.O. FAULKNER, "The Battle of Kadesh," *Mitteilungen des derrtschen archaologischen Instituts Abteilung Kairo,* 16:93–111 (1958); SIR ALAN H. GARDINER, The *Kndesh Inscriptions of Ramesses II* (1960), and with S. LANGDON, "The Treaty of Alliance between Hattušili, King of the Hittites, and the Pharaoh Ramesses II," *Journal of Egyptian Archaeology,* 6:179–205 (1920); K.A. KITCHEN "Some New Light on the Asiatic Wars of Ramesses II," *ibid.,* 50:47–70 (1964); M.C. KUENTZ, "La Stile du mariage de Ramsès II," *Annales du Service des Antiquités de l'Égypte,* 25:181–238 (1925); K.C. SEELE, *The Coregency of Ramses II with Seti I and the Date of the Great Hypostyle Hall at Karnak* (1940).

(R.O.F.)

Rangoon

Rangoon, the capital of the Union of Burma, is located 25 miles from the sea on the left bank of the Rangoon (or Hlaing) River, the eastern mouth of the Irrawaddy. With its population of about 1,900,000, with an international airfield and modern port, and with air, rail, road, and river connections with all of Burma, it is the largest as well as the most important city commercially and industrially in the country. Its area is nearly 200 square miles (500 square kilometres).

The focal point of the city is the ancient Shwc Dagon pagoda, the centre of Burmese religious life, which attracts pilgrims from all parts of the country. The pagoda is a solid brick stupa in the form of a cone raised above a relic chamber and is completely covered with gold. It has been rebuilt and its height increased many times; it was raised, by the Burmese king Tharrawaddy in 1841, to its present height of 326 feet, crowning a hill that rose 168 feet above the level of the city.

History. The Shwe Dagon pagoda has been a place of pilgrimage for many centuries. The settlement that apparently had existed there for a long time became known as Dagon in the medieval period, and its status was raised to that of a town by the Mon kings in the early 15th century. King Alaungpaya (founder of the last dynasty of Burmese kings) conquered Lower Burma in 1755 and developed the town as a port to replace the port of Syriam. He renamed it Yangon ("the end of strife"; transliterated as Rangoon). Surrounded by a stockade and with creeks on three sides and the river on the south, its area was only about one-eighth of a square mile. It had, however, three wharves and a thriving shipbuilding industry, using local teak, and it was made the administrative capital of Lower Burma.

Yangon, "the end of strife"

Rangoon was taken by the British at the outbreak of the First Anglo-Burmese War in 1824 but was restored in 1826. The city was again taken in the Second Anglo-Burmese War of 1852 and became the administrative centre of British Lower Burma. After the annexation of the whole country, it became the capital city and grew in importance. The city has remained central to Burmese life since independence in 1948.

The contemporary city. *City site and environment.* The site of the city is a low ridge, with surrounding delta alluvium. The original settlements were located on the ridge; the new satellite towns were built on delta land, where the elevation is only about ten to 15 feet. Alaungpaya's Rangoon was also built on alluvium, and, when Lieutenant A. Fraser planned British Rangoon in 1853, it was necessary to drain the creeks around the old settlement and fill up the depressions.

The climate is warm and humid. Average temperatures are about 80–85° F (27–29" C), and mean monthly temperatures vary by only about 10" F (6° C). Minimum temperatures in January are, however, about 60–65" F (about 16–18" C), and maximum daytime temperatures in April are about 95–100° F (about 35–38" C). Rains begin in May and last till the middle of October. The total rainfall is more than 100 inches a year. The original vegetation, that of tropical semi-evergreen forest, has been completely removed over the centuries. The fertile alluvial lands were transformed into paddy fields, some of which now underlie the new satellite towns set up after 1958.

City plan and growth. Fraser's plans called for a city with a rectangular pattern of streets covering an area of 500 acres to accommodate a population of 36,000. With the great increase of population, settlements were built up in a U shape around the planned city, now known as Cantonment, toward the west along the river and to the east to Pazundaung Creek. To the north, a spacious residential area developed, while farther north, there were gardens until, as population increased, the gardens made way for more settlements. Today the built-up area extends to Insein and Mingaladon, nine miles and 12 miles north, respectively. In the meantime, in order to accommodate a larger population, important institutions within the city limits were moved out. A much bigger Cantonment area was developed at Mingaladon, the university was moved five miles to the north, and a new racecourse was set up at Kyaikkasau five miles to the northeast.

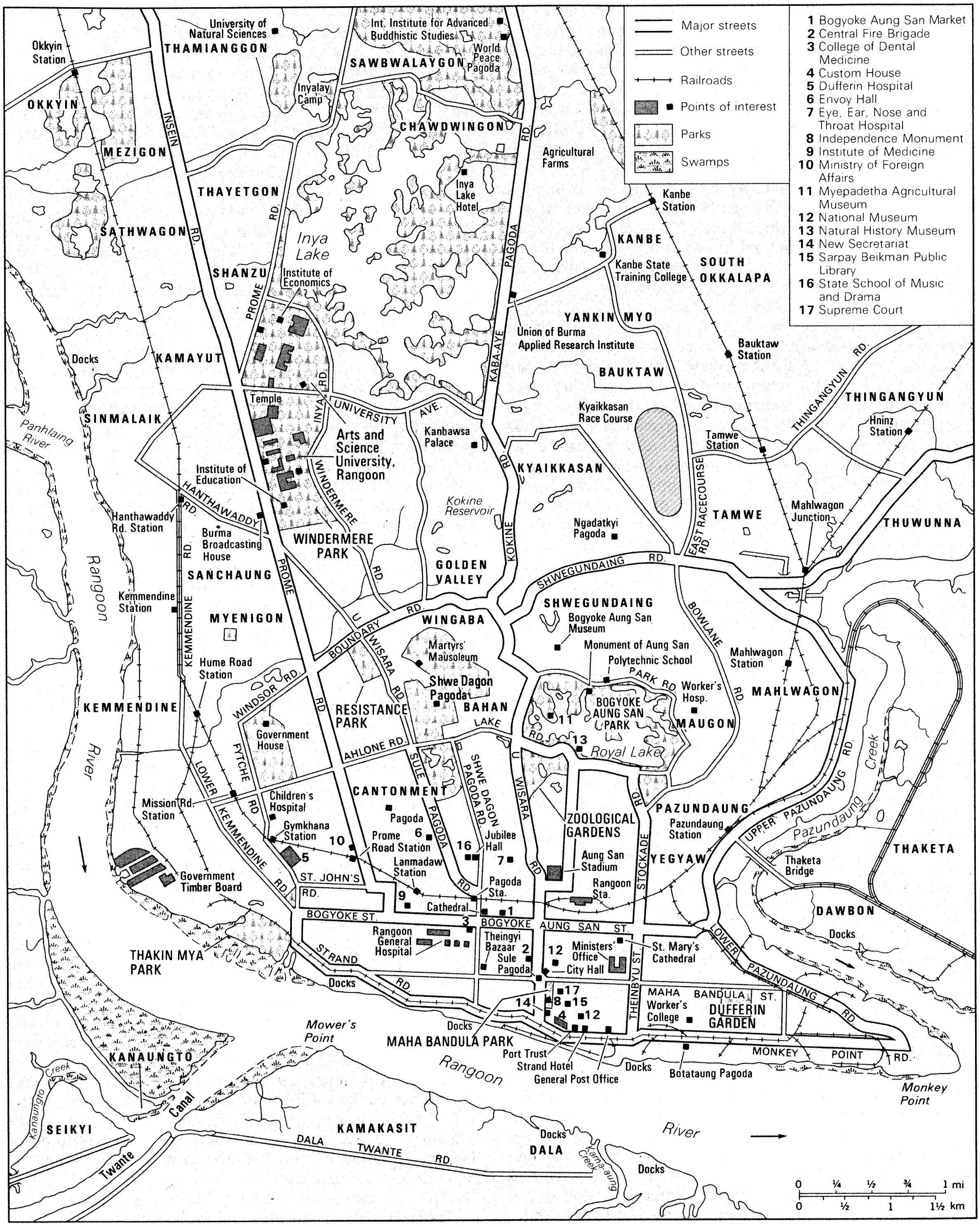
The city plan

The city still maintains the rectangular plan of streets and blocks in the built-up areas of its suburbs and in the new satellite towns of Thaketa and North and South Okkalapa on the east, set up in 1958–59 to accommodate the squatters within the city. A new town, Thuwunna, is also being developed on the east.

Lack of bridging limits the city to the area enclosed by the Rangoon River and Pazundaung Creek; hence, expansion is continuing northward, resulting in lengthening of communication lines and in the necessity of locating cemeteries and factories within the confines of the city.

The city is well served by bus lines, and a railway, using diesel engines, connects all the important suburbs with the centre of the city. In addition there are many local trains.

Demography. In 1971 the population of the city was estimated at 1,900,000. The growth of population has been phenomenal: in 1881 it was 134,176; by 1901 it had risen to 234,881; in 1931 it was 400,415; in 1941, 500,800; and in 1953, 737,079. The present population of nearly 1,900,000 is for the whole of the division covering nearly 200 square miles. In the centre of the city, the density of population is over 100,000 per square mile while in the outskirts it falls to as low as 1,500 per square mile, at Mingaladon.



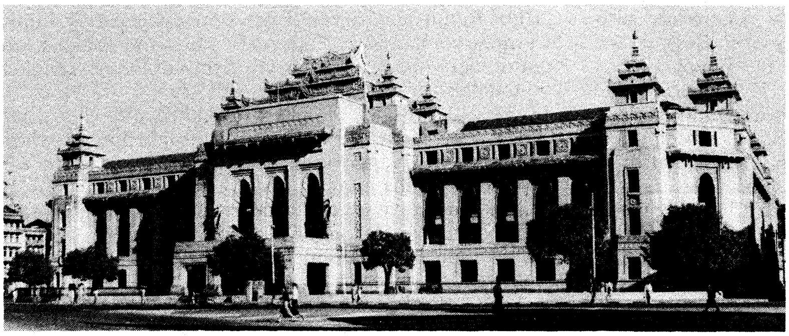
The city of Rangoon and vicinity.

Over the years the ethnic composition of the city also has changed. When first instituted in 1755, the population was made up mainly of Mons with some Burmese and a sprinkling of foreigners. Under the British, the foreign population grew rapidly. In 1901 the Burmese constituted 33% of the population, and Indians 48%. By 1931 the Burmese formed 32%, and the Indians 53%. There was a great exodus of Indians and Europeans during World War II. By 1952 the structure had changed to 60% Burmese, 20.5% Indians, 8% Chinese, 5% Anglo-Burmans, and the rest foreigners, including Europeans, Armenians, Jews, and others. Today 90% of the population is Burmese, and the Chinese outnumber the Indians; a contributing factor is that many Indians and Chinese have adopted Burmese citizenship.

Housing and arclzitecture. The most notable building in Rangoon is the great golden stupa in the Shwe Dagon pagoda. The place of worship is the surrounding tiled terrace with a perimeter of 1,420 feet. In 1952, to celebrate the 2,500th year of the passing of the Buddha, the World Peace Pagoda was built a few miles to the north; around it has been built a complex of buildings, among which a great artificial cave with assembly hall and the International Institute for Advanced Buddhistic Studies are notable. Other religious edifices of note are the Sule pagoda, located in the centre of the city, and Botataung pagoda, demolished by bombing during World War II and completely rebuilt in 1953, near the river to the east.

The Shwe Dagon pagoda

The city centre is made up of brick buildings, mostly three to four stories high, while in the suburbs traditional



City Hall on Sule Pagoda Road, Rangoon.
Richard Allen Thompson

wooden structures are common. Old colonial structures of red brick include the Office of Ministers (formerly the Old Secretariat), the Law Courts, the Rangoon General Hospital, the customhouse, and Government House. New and modern architectural styles are seen in the new secretariat building, the department stores on Shwe Dagon Pagoda Road, the new Polytechnic School, the Burma Broadcasting Service Station, the Institute of Medicine, and the Rangoon Institute of Technology. In housing, too, new styles have appeared.

Economic life. Under the British the most important industries were those using local produce. Hence Rangoon has the biggest rice mills and saw mills in the country, strung out along the riverside. Since independence there has been a great increase in industrialization, including textile mills and soap, rubber, aluminum, and food-processing factories. State-owned industries include pharmaceuticals, soap, and textiles and an iron- and steel-rolling plant. An industrial belt stretches from Kamayut to Insein and beyond to the north and on the east to Thingangyun and beyond.

In the commercial section of the city, located in the central area, are the banks, offices of central trade, and various trade corporations, as well as shops, brokerage houses, and bazaars. In addition there are many local bazaars in various suburbs and, with the program of decentralization, retail shops. Many of the retail shops, formerly state-owned, have been changing over to operation by consumer cooperatives. Private stores, service activities, and workshops are also present in every quarter of the city.

The port of Rangoon

Rangoon commands routes by river, road, and railway to the rest of the country. The Union of Burma Airways connects internally more than 30 airfields in Burma and operates external connections with Bangkok, Hong Kong, Dacca, Calcutta, and Kāthmāndu. The international airport at Mingaladon, built in 1952, accommodates several international airlines. Before World War II, the port of Rangoon handled 84 percent of the country's overseas commerce. During the war it was repeatedly bombed and badly damaged by both the Japanese and the Allies. After independence, with the help of foreign aid, the port was completely rehabilitated. The port now handles all imports that reach the country and 85 percent of all exports, but the overall volume has remained far below prewar totals.

Administration. The whole city area is administered by the Municipal Corporation of Rangoon, established in 1874. The development of the city, however, was left to the Rangoon Development Trust, set up in 1920, and is now under the National Housing Board. Since Rangoon division was instituted in 1964, much of the administration has been carried out by local security and administrative committees. The city is now divided into 13 regions and 28 smaller subregions, constituted as townships in 1971. Being the capital, Rangoon has government offices as well as diplomatic missions.

Public utilities. Rangoon draws its water supply from Hlawga Lake and the Gyobyu reservoir, but artesian wells have been drilled to supplement the supply, and the planned construction of a new reservoir at Hpugyi, 35 miles to the north, scheduled for completion in about 1980. The city has had a sewerage system since 1892, but extensive additions are needed for the new towns. Since 1960, Rangoon has drawn its electricity supply from the Balu Chaung hydroelectric project in Kayah State.

Health andsafety. Before World War II, Rangoon had only the Rangoon General Hospital, the Dufferin Hospital for women, and smaller hospitals in the various quarters. Four large new general hospitals have been built, as well as the Workers' Hospital, Children's Hospital, and the Eye, Ear, Nose, and Throat Hospital. The city also has health service stations in all quarters of the city. Medical treatment is free in Burma. Rangoon also has more than a dozen day nurseries in which children of working families are cared for; more are planned.

In addition to the Central Fire Brigade, there are fire stations in the various suburbs, those at Kemmendine and Insein being the largest. Rangoon is troubled by fires fairly often, though on a small scale. Rangoon Division Police Services control 46 police stations.

Institutions of higher education

Education. There are almost 100 high schools, two technical high schools, and several hundred secondary and primary schools, all government supported. The University of Rangoon, established in 1920, was reconstituted in 1964 into separate institutes of arts and science, economics, education, technology, medicine (two institutes), veterinary and animal husbandry, dentistry, and paramedical sciences, and a Workers College. The Inter- [illegible]al Institute for Advanced Buddhistic Studies was established in the 195[illegible] Alt[illegible]th[illegible] R[illegible] [illegible] between 40,000 and 50,000 students, half of whom are in the Arts and Science University. The various institutes and the university maintain hostels for the students. In the early 1970s, a new University of Natural Sciences was being constructed at Thamaing. Rangoon also has a Teachers Training Institute (for secondary and elementary level) and a Technical Institute at Insein.

Museums and libraries

Cultural life. The government of Burma maintains schools of dancing, music, drama, and art in Rangoon and a cultural troupe, which, besides entertaining the public, also preserves ancient dancing and music. An open-air theatre on the old racecourse near Lanmadaw and the former Jubilee Hall serve as theatres. There are about 20 cinema halls. Rangoon has five museums, the National Museum (opened in 1952), Bogyoke Aung San Museum, the Museum of Military Research, the Museum of Buddhistic Studies at the International Institute for Advanced Buddhistic Studies, and the Museum of Natural History. Libraries include the National Library (with more than 100,000 books and volumes of palm leaf and parabaiks), the Central Universities Library, and the libraries of the Burma Translation Society (Sarpay Beikman Public Library) and the Information Department.

Rangoon has many presses publishing books, magazines, and journals, the largest being the government's central press, Sarpay Beikman Press, Myawaddy Press, Buddhist Society Press, and the University Press. Seven daily newspapers are published, five in the vernacular and two in English. There is no television, but the state-owned radio broadcasting station operates from its modern building opened in 1960.

Some 16 parks are maintained in the city, the largest of which are the Maha Bandula Park (the old Fytch Square

Garden) and the Park of Revolution. There are also spacious zoological and botanical gardens. Besides the old racecourse near Lanmadaw, where the Envoy Hall is often open for exhibitions, the old Kyaikkasan race course also serves for exhibitions and as a gathering place on important national days. (Horse racing has been abolished since 1962.) There is a fine 18-hole golf course at Mingaladon, and another one near Insein. There are 19 stadiums for sports and athletic meets, the best known is the Aung San Sports Stadium, which can accommodate 50,000 spectators and has an indoor stadium and lighting facilities for sports at night.

BIBLIOGRAPHY. O.H.K. SPATE and L.W. TRUEBLOOD, "Rangoon: A Study in Urban Geography," *Georgl. Rev.*, 32:56–73 (1942), treats in detail the historical and geographical background of the study and the growth of the city and its population, with maps. The classic historical study is B.R. PEARN, *History of Rangoon* (1939).

(T.Ky.)

Ranjit Singh

Ranjit Singh, Sikh maharaja of the Punjab (Paiijrib), was the first Indian in 1,000 years to turn the tide of invasion back into the homelands of the traditional conquerors of India, the Pathans and the Afghans, and hence came to be known as the Lion of the Punjab. His domains extended from the Khyber Pass in the northwest to the Sutlej River in the east, and from Kashmir down to the deserts of Sindh. Although unlettered, he was a shrewd judge of men and events, free from religious bigotry, and mild in his treatment of adversaries. An ugly little man, blind in one eye and with a face pitted with pockmarks, he liked to surround himself with handsome men and women; a *bon viveur*, he had a passion for hunting, horses, and strong liquor.

Early life and military conquests

Ranjit Singh was born on November 13, 1780, the only child of Maha Singh, on whose death in 1792 he became chief of the Śukerchakīās, a Sikh group. His inheritance included Gujrānwāla town and the surrounding villages, now in Pakistan. At 15 he married the daughter of a chieftain of the Kanhayas, and for many years his affairs were directed by his scheming and ambitious mother-in-law, the widow Sada Kaur. A second marriage to a girl of the Nakkais made Ranjit Singh pre-eminent among the clans of the Sikh confederacy. In July 1799 he seized Lahore, the capital of the Punjab. The Afghan king, Shrih Zamān, confirmed Ranjit Singh as governor of the city; on April 12, 1801, however, Ranjit Singh proclaimed himself maharaja of the Punjab. He had coins struck in the name of the Sikh Gurūs, the revered line of Sikh leaders, and proceeded to administer the state in the name of the Sikh Commonwealth. A year later he captured Amritsar, the most important commercial entrepôt in northern India and sacred city of the Sikhs. Thereafter he proceeded to mop up smaller Sikh and Pathan principalities scattered over the Punjab. His forays eastward were checked by the English. By a treaty signed in 1806, he agreed to expel a Marāthā force that had sought refuge in the Punjab. His ambition to bring together all the Sikh territories extending up to the vicinity of Delhi was thwarted by the English, and he was compelled to sign another treaty with them on April 25, 1809, relinquishing claims to territories east of the Sutlej River.

Ranjit Singh turned his ambition in other directions. In December 1809 he went to the aid of Raja Sansar Chand of Kangra in the Punjab Hills and, after defeating an advancing Gurkha force, acquired Kangra for himself. In 1813 he joined a Barakzrii Afghan expedition into Kashmir. While the Barakzāis betrayed him by keeping Kashmir for themselves, he more than settled scores with them by rescuing Shiih Shojāʿ, brother of Shāh Zamān, who had fled the Barakzāis, and by occupying the fort of Attock on the Indus. Shrih Shojāʿ was brought to Lahore and pressured into parting with the famous *Koh-i-noor* diamond.

In the summer of 1818 Ranjit Singh's troops captured the city of Multān and six months later entered the Pathan citadel, Peshāwar. In July 1819 he finally expelled the Afghans from the Vale of Kashmir.

All of Ranjit Singh's conquests were achieved by Punjabi armies composed of Sikhs, Muslims. and Hindus. His commanders were also drawn from different religious communities, as were his cabinet ministers.

Modernization of the Punjabi army

In 1820 Ranjit Singh began to modernize his army. Almost 50 foreign officers, a good proportion of whom had served in Napoleon Bonaparte's armies, were employed to train the infantry and artillery. The modernized Punjabi army gave a good account of itself in campaigns in the North-West Frontier (on the Afghanistan border). In 1823 it defeated a combined Afghan and Pathan force at Naushahra. In 1831 it successfully quelled a rising of the Frontier tribesmen roused to a holy war (*jihād*) by the Muslim fanatic Sayyed Aḥmad, and again at Naushahra, in 1837, it defeated a Pathan–Afghan alliance.

In October 1831 the British viceroy, Lord William Bentinck, received Ranjit Singh at Rūpar on the eastern bank of the Sutlej River. The British, who had already begun to navigate the Indus River and were eager to keep Sindh Province for themselves, prevailed upon Ranjit Singh to fall in line with their plan. Ranjit Singh was chagrined by the British design to put a cordon round him. He opened negotiations with the Afghans and sanctioned an expedition led by the Dogra commander Zorawar Singh that extended his northern territories into Ladākh (a region of eastern Kashmir) and brought the Punjabis in confrontation with the Chinese.

In 1838 the British viceroy Lord Auckland met Ranjit Singh at Firozpur and persuaded him to engage in a joint expedition into Afghanistan to place the British nominee, Shrih Shojāʿ, on the throne of Kābul. In pursuance of this agreement, the British Army of the Indus entered Afghanistan from the south, while Ranjit Singh's troops went through the Khyber Pass and took part in the victory parade in Kābul.

Ranjit Singh was taken ill during the victory festivities at Firozpur. He died at Lahore on June 27, 1839, exactly forty years after he had entered the city as a conqueror. At his cremation four maharanis (wives) and seven maidservants immolated themselves on his funeral pyre.

BIBLIOGRAPHY. LEPEL H. GRIFFIN, *Ranjit Sittgh* (1892); N.K. SINHA, *Ranjit Singh* (1933); and KHUSHWANT SINGH, *Ranjit Sitrgh, Maharajah of the Panjab* (1962), are three conventional biographies. For an eyewitness account of the personality and court of Ranjit Singh, see EMILY EDEN, *Up the Country: Letters Written to Her Sister frotn the Upper Provinces of India*, 2 vol. (1866); VICTOR JACQUEMONT, *Correspondance de Victor Jacquemottt avec sa famille et plusieurs de ses amis, pendant son voyage dans l'Inde (1828–1832)*, 2 vol. (1833; Eng. trans., *Letters from India Describing a Journey in the British Dominions of India, Tibet, Lahore and Cashmere, During the Years 1828–1831*, 2 vol., 1834); and W.G. OSBORNE, *The Court and Catnp of Runjeet Sing* (1840).

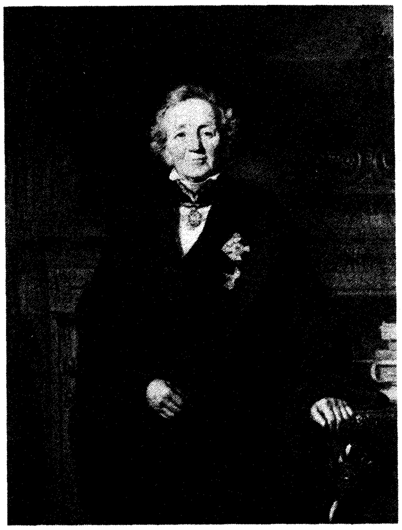
(K.S.)

Ranke, Leopold von

Leopold von Ranke, the leading German historian of the 19th century, has been called the father of modern historiography. Distinguished by his insistence on careful and methodical research, his passion for objectivity, and his belief in the unique importance of historical studies, he transmitted these ideas to several generations of future historians and had profound influence on historiography in the 19th and 20th centuries.

Education

Ranke was born into a devout family of Lutheran pastors and lawyers on December 21, 1795, in Wiehe, Thuringia (now in East Germany). After attending the renowned Protestant boarding school of Schulpforta, he entered the University of Leipzig. He studied theology and the classics, concentrating on philological work and the translation and exposition of texts. This approach he later developed into a highly influential technique of philological and historical textual criticism. His predilection for history arose from his studies of the ancient writers, his indifference to the rationalistic theology still in vogue in Leipzig, and his intense interest in Luther as a historical character. But he decided in favour of history only in Frankfurt an der Oder, where he was a secondary school teacher from 1818 to 1825. Apart from the contemporary patriotic enthusiasm for German history, his decision was



Ranke, oil painting by J. Schrader, 1868. In the National-Galerie, East Berlin.
By courtesy of the Staatliche Museen zu Berlin

influenced by Barthold Georg Niebuhr's Roman history (which inaugurated the modern scientific historical method), the historiographers of the Middle Ages, and Sir Walter Scott's historical novels, as well as by the German Romantic poet and philosopher Johann Gottfried von Herder, who regarded history as a chronicle of human progress. Yet Ranke's strongest motive was a religious one: influenced by the philosophy of Friedrich Schelling, he sought to comprehend God's actions in history. Attempting to establish that God's omnipresence revealed itself in the "context of great historical events," Ranke the historian became both priest and teacher.

Early career. The typical features of Ranke's historiographical work were his concern for universality and his research into particular limited periods. In 1824 he produced his maiden work, the *Geschichte der romanischen und germanischen Volker von 1494 bis 1514 (History of the Latin and Teutonic Nations from 1494 to 1514,* 1846), which treats the struggle waged between the French and the Habsburgs for Italy as the phase that ushered in the new era. The appended treatise, *Zur Kritik neuerer Geschichtsschreiber*, in which he showed that the critical analysis of tradition is the historian's basic task, is the more important work. As a result of these publications, he was appointed associate professor in 1825 at the University of Berlin, where he taught as full professor from 1834 to 1871. Many of the students in his famous seminars were to become prominent historians, continuing his method of research and training in other universities. In his next book, Ranke, utilizing the extremely important reports of the Venetian ambassadors, dealt with the rivalry between the Ottoman Empire and Spain in the Mediterranean *(Fiirsten und Völker von Siid-Europa im sechzehnten und siebzehnten Jahrhundert*); from 1834 to 1836, he published *Die romischen Päpste, ihre Kirche und ihr Staat int sechzehnten und siebzehnten Jahrhundert* (changed to *Die romischen Papste in den letzen vier Jahrhunderten* in later editions)—a book that ranks even today as a masterpiece of narrative history. Rising above religious partisanship, Ranke in this work depicts the papacy not just as an ecclesiastical institution but above all as a worldly power.

Excursion into politics

Before this work appeared, Ranke the historian had been drawn briefly into contemporary history and politics. A disillusioning experience, it produced, however, a few short writings in which he expressed his scholarly and political convictions more directly than in his major works. Disregarding his real talents and misjudging the contemporaneous political dissensions, which in 1830 were intensified by the liberal July revolution in France, he undertook to edit a periodical defending Prussian policy and its rejection of liberal and democratic thinking. Only two volumes of the *Historisch-politische Zeitschrift* were published from 1832 to 1836, most of the articles being written by Ranke himself. While he tried to explain the conflicts of the times from a historical—and for him that meant nonpartisan—viewpoint, in essence he sought to prove that the French revolutionary development could not and should not be repeated in Germany. Ranke believed that history evolves in the separate development of individual men, peoples, and states, which together constitute the process of culture. The history of Europe from the late 15th century onward—in which each people, though sharing one cultural tradition, was free to develop its own concept of the state—seemed to him to confirm his thesis. Ranke dismissed abstract, universally valid principles as requirements for the establishment of social and national order; he felt that social and political principles must vary according to the characteristics of different peoples. To him the individual entities of greatest historical importance were states, the "spiritual entities, original creations of the human mind—even 'thoughts of God.' " Their essential task was to evolve independently and, in the process, to create institutions and constitutions adapted to their times.

In this respect Ranke's thinking is related to the philosopher G.W.F. Hegel's theory that what is real is also rational; yet, in Ranke's view, it is not reason that justifies what is real but historical continuity. This continuity is the prerequisite for the development of a culture and also for understanding historical reality. Hence, it is the historian's duty to understand the essence of "historicism": that history determines each event but does not justify it. In practice, however, Ranke endorsed the social and political order of his time—the European system of states, the German Federation with its numerous monarchies, and Prussia before the 1848 Revolution, with its powerful monarchy and bureaucracy, its highly developed educational system, and its rejection of liberal and democratic trends—as resulting from the European cultural process, a process that, according to him, would be demolished by democratic revolution.

The searchfor **objectivity.** But Ranke pleased no one; too devoted to the state for the liberals, he was not sufficiently dogmatic for the conservatives. He therefore returned to his historiographical work in which he thought he could more successfully attain his ideal of objectivity. From 1839 to 1847 the *Deutsche Geschichte im Zeitalter der Reformation (History of the Reformation in Germany,* 1845–47) appeared, the first scholarly treatment of that age. In 1847–48 there followed *Neun Biicher preussischer Geschichte (Memoirs of the House of Brandenburg and History of Prussia, during the Seventeenth and Eighteenth Centuries,* 1849), later expanded to 12 volumes; in 1852–61 the *Franzosische Geschickte, vornehmlich im sechzehnten uizd siebzehnten Jahrhundert,* published in English as *Civil Wars and Monarchy in France, in the Sixteenth and Seventeeth Centuries: A History of France Principally During that Period,* 1852; and, in 1859–69, the *Englische Geschichte, vornehmlich im sechzehnten und siebzehnten Jahrhundert*, published in English as A *History of England Principally in the Seventeenth Century,* 1875—each consisting of several volumes that, although partly rendered obsolete by later research, are still worth reading today for their great narrative skill. In these works, too, Ranke deals with the leading European states at decisive stages of their development within the European system. Ranke typically restricts himself to the Latin and Germanic nations as the protagonists of cultural development, among whom—from the 16th century on—the Protestant states had increasingly assumed leadership; and just as typically, he focusses on political history; *i.e.*, the foreign relations of states and their systems of government and administration. Because economic and social factors were barely reflected in the sources he used, appearing only dimly in the background as "forces" and "tendencies," Ranke found it increasingly difficult to understand the modern age of incipient social change.

His books on the late 18th and early 19th centuries *(Die deutschen Miichte und der Fürstenbund,* 1871–72; *Ursprung und Beginn der Revolutionskriege 1791 und 1792,* 1875; *Hardenberg und die Geschichte des preussischen*

Ranke's theory of history

Staates von 1793 bis 1813, 1877) are subtle accounts of complex political events but address themselves only indirectly to the central problems of a changing age. Like the *Englische Geschichte,* these books exhibit a certain bias against political and social change, especially the appearance of radical movements. In his lectures Ranke often dealt with the history of his time; they did not, apparently, differ in concept or emphasis from his books. History is regarded as a complex process of "historical life," which assumes its most effective "real spiritual" form in the great states and their tensions. The historian, as objectively as possible, must describe "how it really was," keeping the whole picture in mind while extracting the essence. Ranke was thus not an analyst but a "visual" historiographer. Aware of the limitations imposed by time and place on every historian, he attempted to achieve maximum objectivity principally by identifying himself not with a "party" but with the state. Yet his work demonstrates that his intellectual credo influenced his political views.

Ranke reached the peak of his fame as the most important living historian in the second half of the century. In 1865 he was ennobled and in 1882 made a privy counsellor. When Frederick William IV became mentally ill in 1857, Ranke finally withdrew from political life and, after his wife's death (1871), from social life also. Rejecting liberal democratic nationalism and distrusting Chancellor Otto von Bismarck's policy because he believed that it jeopardized the continuity of German history and embraced cooperation with popular movements, Ranke nevertheless welcomed the foundation of the empire in 1871.

In the meantime, failing eyesight had turned him into a lonely scholar who depended on the help of assistants. Yet, despite this handicap, at the age of 82 he began what he claimed to be his greatest work, a "world history" (nine volumes, 1881–88) leading up to the 15th century. Ranke thus fulfilled the task he had set himself as a young man: to tell the "story of universal history." Not a work of critical research or of historical and philosophical speculation but a wide-ranging account of the evolution of culture from the Greeks to the Latin-Germanic nations, it is actually a history of Europe in which the non-European world appears at best only marginally. He wrote it in the conviction that the peaceful evolution of culture was definitively protected against the danger of revolution and that the conflict between popular sovereignty and the monarchy had been settled once and for all in favour of the latter. He died in Berlin on May 23, 1886.

Assessment. Ranke's concept and writing of history predominated in German historiography up to World War I and even after; it also influenced a great many distinguished foreign historians who studied in Germany. Unfortunately, many of Ranke's disciples simply continued, canonized, and debased Ranke's concepts, retaining all of their limitations without the universality of view that gave them meaning. Ranke's own achievements, however, remain unquestioned. He contributed greatly to the progress of historiography: it became more self-assured in its method and proved itself capable of transforming the widely felt need for a historical understanding of the world ("historicism") into an interpretation of the past based on scientific research.

MAJOR WORKS

Geschichte der romanischen und germanischen Volker von 1494 bis 1514, with an appendix *Zur Kritik neuerer Geschichtsschreiber* (1824; *History of the Latin and Teutonic Nations,* 1846); *Fiirsten und Volker von Siid-Europa im sechzehnten und siebzehnten Jahrhundert:* vol. 1, *Die Osmanen und die spanische Monarchie* (1827; *The Ottoman and the Spanish Empires in the Sixteenth and Seventeenth Centuries,* 1843); *Die serbische Revolution* (1829; *A History of Serbia . . . ,* 1847); *Die grossen Miichfe* (1834); *Die römischen Piipste, ihre Kirche und ihr Staat, im sechzehnten und siebzehnten Jahrhundert,* 3 vol. (1834–36; subsequently expanded to *Die römischen Piipste in den letzten vier Jahrhunderten,* 1878; 10th ed., 1900; rev. ed., 1953; *The Ecclesiastical and Political History of the Popes of Rome . . . ,* 3 vol., 1840; and *The History of the Popes during the Last Four Centuries,* 3 vol., 1908); *Das politische Gespräch* (1836); *Deutsche Geschichte im Zeitalter der Reformation,* 6 vol. (1839–47; modern ed., 1925–26; *History of the Reformation in Germany,* 3 vol., 184547); *Neun Biicher preussischer Geschichte,* 3 vol. (184748; *Memoirs of the House of Brandenburg . . .,* 3 vol., 1849), subsequently expanded to *Zwolf Biicher . . .* (1874–79; modern ed., 1931); *Franzosische Geschichte, vornehrnlich in sechzehnten und siebzehnten Jahrhundert,* 5 vol. (1852–61; 4th ed., 6 vol., 1876–77; modern ed., 2 vol., 1954; *Civil Wars and Monarchy in France in the Sixteenth and Seventeenth Centuries,* 2 vol., 1852); *Englische Geschichte, vornehrnlich in sechzehnten und siebzehnten Jahrhundert,* 7 vol. (1859–69; 4th ed., 9 vol., 1877–79; modern ed., 1955; *A History of England, Principally in the Seventeenth Century,* 6 vol., 1875); *Geschichte Wallensteins* (1869); *Der Ursprung des Siebenjährigen Krieges* (1871); *Die deutschen Miichte und der Fiirstenbund,* 2 vol. (1871–72); *Ursprung und Beginn der Revolutionskriege, 1791 und 1792* (1875; 2nd ed., 1879); *Weltgeschichte,* 9 vol. (1881–88, incomplete). The collected edition of Ranke's works comprises 54 vol. (1867–90). There are important collections of his illuminating letters (1949 and 1950).

BIBLIOGRAPHY. HEINRICH VON SRBIK, *Geist und Geschichte vom deutschen Humanismus bis zur Gegenwart,* vol. 1 (1950); and GEORGE P. GOOCH, *History and Historians in the Nineteenth Century,* 2nd ed. (1952), treat in detail the position of Ranke in the development of historical thought and the modern science of history. The philosophical foundations of his view of history are set forth by FRIEDRICH MEINECKE, *Die Entstehung des Historismus* (1946); and CARL HINRICHS, *Ranke und die Geschichtstheologie der Goethezeit* (1954). Of the older investigations in this direction, GERHARD MASUR, *Rankes Begriff der Weltgeschichte* (1926); and ERNST SIMON, *Ranke und Hegel* (1928), are still important. RUDOLF VIERHAUS, *Ranke und die soziale Welt* (1957), deals with the role of social factors in Rankes historiography. Recent critical analyses are THEODORE H. VON LAUE, *Leopold Ranke: The Formative Years* (1950); and HELMUT BERDING, *Leopold von Ranke* (1971), in German.

(Ru.V.)

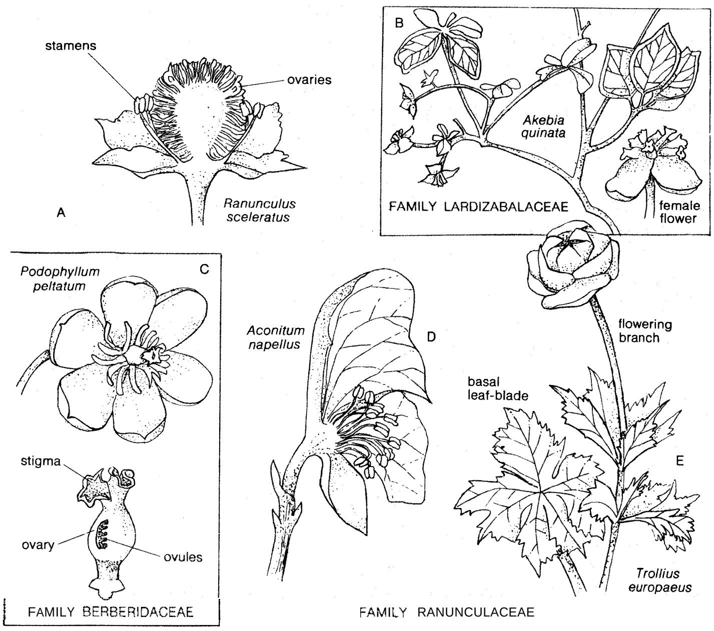
Ranunculales

The Ranunculales order is a large group of plants comprising 10 families, about 140 genera, and approximately 3,000 species. It is a rather diverse grouping, including mainly herbaceous plants and climbers, though some are woody shrubs. They are of considerable interest because they are abundant in most temperate areas, often forming a characteristic element of the floras. They also include many species of great ornamental value, which are grown in gardens in many parts of the world. Several species belonging to the order are common and noxious weeds, particularly in Europe and North America; indeed, many species contain compounds (mainly alkaloids) that are poisonous to man or to livestock. Some of these compounds are also important in folk medicine. In evolutionary terms the group is of especial importance, as it is considered by many botanists to be an early offshoot from the Magnoliales (*q.v.*)—the order generally considered to be the most primitive among the flowering plants.

GENERALFEATURES

Both vegetatively and florally the Ranunculales order is highly variable. There are, nevertheless, a number of features held in common by most members of the group. Among these are the herbaceous or softly wooded habit, and the usually alternately arranged leaves that lack the small leaflike appendages, called stipules, at the bases of their leaf stalks. In the flowers, the segments of the perianth, (sepals and petals) are usually spirally arranged, and the stamens are usually numerous and they, too, are often spirally arranged. There are usually several free carpels (female flower structures) that are spirally arranged or sometimes whorled, but they are rarely united or reduced in number to one in this order. Alkaloid chemical substances are generally present.

Distribution and abundance. The Ranunculales order is a cosmopolitan group and is found everywhere; its species are, however, more common in temperate and subtropical regions than in the tropics. Only one of the ten families of the order (Menispermaceae) is mainly tropical. Members of the order occupy a wide range of habitats, including freshwater, and the species are often very abundant. The barberry genus, *Berberis,* for exam-



Some vegetative and floral diversity in the buttercup order.

From *(Ranunculus sceleratus, Aconitum napellus)* Baillon, Eichler, Firbas, Rassner, Troll, Warming and *(Podophyllum peltatum)* Eichler, A. Gray, Hegi, Maout et Decaisne, Schumann, Warburg, Warming, Wettstein in A. Engler, *Syllabus der Pflanzenfamilien II* (1964), Gebruder Borntraeger, Berlin; (others) G.H. Lawrence, *Taxonomy of Vascular Plants* (1969), The Macmillan Company.

ple, is widespread in temperate South America and also in the Himalayas; the buttercup genus, Ranunculus, is common in Europe and North America, and the monkshood genus, *Aconitum*, is abundant in the Himalayas and Eastern Asia.

Importance. The importance of the group is related to its botanical and economic interest. When considered as an early offshoot from the most primitive flowering-plant stock, the group is of considerable evolutionary interest. Related to this is the fact that the group appears to show a series of structural features that can be construed as showing the origin of flower petals from sterile stamens. Many of the intermediate stages are present in existing species, and almost the whole series is shown by the largest family (Ranunculaceae), with various segments of the series being shown by the other families.

Interpretation of the possible origin of flower petals

The simplest stage in this series is shown by genera such as Anemone and *Pulsatilla,* in which the perianth consists of a single series (*i.e.*, sepals, or, as they are sometimes called, tepals) ; these are usually coloured and serve to attract insects. The stamens (male reproductive structures) are numerous, and all are fertile. Many species of *Clematis* are similar, but others have the outer stamens sterile and somewhat broader than the fertile inner ones. In the genus *Trollius* the outer stamens are more or less flattened and are sterile, the anther-bearing portion being represented by a nectar-secreting pit on the inner surface, near the base. These organs are coloured and very petal-like, though the main insect-attracting function is still carried out by the larger and brightly coloured sepals. In the genus Ranunculus the process is carried further, the sepals being small and usually greenish, whereas the nectar-secreting organs are larger and coloured and take on the main insect-attracting function. Like those of *Trollius*, these organs each bear a nectar-secreting pit on the inner surface, near the base, often covered by a fold or flap of tissue. These organs, which are clearly homologous with sterilized stamens, occur in a close spiral between the sepals and the fertile stamens. They are generally few in number (usually five), alternate in position with the sepals, and are conventionally referred to as petals by most taxonomists, although the Berlin school prefers to call them "honey leaves" (Honigbliitter).

In columbines of the genus Aquilegia the petals are more or less tubular structures, with an expanded and flattened forward portion and backwardly projecting spurs, at the base of which nectar is secreted. The petals are only slightly larger than the sepals, and both series are coloured and serve to attract insects. In *Helleborus, Nigella*, and related genera, the petals are again tubular but much smaller than the usually coloured sepals. They are often of very elaborate construction and serve only to secrete nectar. In those genera with highly zygomorphic (irregular or bilaterally symmetrical) flowers (*Aconitum*, *Consolida*, Delphinium), the petals are much-modified nectar-secreting structures, often enclosed in and hidden by the large coloured sepals.

Thus, there is a series of elaboration of nectar-producing structures and petal-like structures. It must be emphasized, however, that this series cannot be directly interpreted as showing an evolutionary lineage. What it does show, however, is various stages of a process of stamen sterilization and nectary and petal formation that may have taken place in the earliest angiosperms (flowering plants) or pro-angiosperms.

NATURAL HISTORY AND EVOLUTION

Life cycle. The life cycle of the Ranunculales order presents no great differences from that of other flowering plants. The group includes short-lived ephemeral annuais (*e.g.*, many species of the buttercup genus, Ranunculus), perennial herbs of longer or shorter duration, suffrutescent perennials (*i.e.*, those with a woody base that lives from year to year but with branches that die back each year), and woody shrubs and climbers that may persist for long periods. Reproduction is mainly by seed, though various forms of vegetative reproduction do occur (*e.g.*, rhizomes and stolons occur in many species, and bulbils are found in some forms of Ranunculus *ficar*ia). Formation of embryos without fertilization occurs, but it is apparently rare in the group.

Growth forms

Ecology. A very wide range of habitats is occupied by the Ranunculales order, including freshwater (both still and flowing), waste ground, field margins, grassland, pasture, forest, moorland, bogs and marshes, mountain pastures, and mountain slopes. The Menispermaceae family consists of mostly climbers and lianas in tropical forests, whereas the rest of the families in the order tend to avoid tropical areas or to grow only on high mountains (where there is a more temperate climate) within the tropic zone. In general, it may be said that most species of the Ranunculales order are either adapted to average moist habitats or to very wet ones, including even aquatic situations; true dry-habitat species are rare, and desert is essentially the only major plant habitat in which the Ranunculales order is not strongly represented.

The aquatic species of the genus Ranunculus, (Ranun*culus*, subgenus *Batrachium*) are an extremely variable group, difficult to classify, and some of them show the interesting phenomenon of heterophylly—the ability to produce leaves of different shapes depending, in this case, on the depth of the water in which they grow.

In Ranunculus *aquatilis* the plants may bear leaves of three different shapes, depending on the water levels; the submerged leaves are repeatedly divided on a pattern of thirds into threadlike segments; the leaves borne at the surface of the water are simple and palmately lobed, with more or less oblong, toothed segments; and the aerial leaves are divided two or three times by thirds into wedge-shaped segments. Thus, the precise type of leaves that are produced depends on the water levels: at times when the water is high, plants will mainly have leaves of the submerged type; when the water is low, most of the leaves will be of the aerial type. This variation causes difficulty in the classification of the aquatic buttercups.

Pollination. The Ranunculales order shows a very wide variation in methods of pollination, a variation that is reflected in the wide range of floral types found within the group. The mainly temperate families Ranunculaceae and Berberidaceae are the best known in this respect; the range of pollination types in the Ranunculaceae includes almost all of the types in the rest of the order.

In the Ranunculaceae family, as in the Ranunculales order as a whole, insect pollination is the general rule. It seems possible, however, that at least one or two species of *Thalictrum* may be wind pollinated, if not as a general and regular occurrence, then at least sometimes. Self-pollination may be the rule in many of the short-lived annual species in the family.

Insect-pollinated flowers

The insect-pollinated types may be divided into two broad classes — pollenflowers and nectar flowers, depending on what the visiting insect obtains from its visits to the flowers. In the first category are such genera as Anemone, *Pulsatilla*, and Clematis, in which no nectar is produced but from which the insects collect pollen. In the second category are such genera as Aquilegia (columbines), *Ranunculus* (buttercups), Helleborus, and Delphinium (larkspurs), which have well-developed nectaries.

In most cases insects are attracted to the flowers by the coloured perianths. The coloured parts may be sepals, as in Anemone and Clematis; petals (honey leaves), as in Ranunculus; a combination of the two, as in Aquilegia; or the very complex arrangements of sepals and petals containing nectaries in the strongly zygomorphic flowers of Aconitum (monkshood), Delphinium, and Consolida. In some species of *Thalictrum* (meadow rue)—e.g., T. kiusianum — the filaments of the stamens are swollen and coloured, and appear to act as insect attractors.

In all these types the flowers may be obligate outcrossers because either the female ovaries mature before the male anthers of the same flower (protogyny) or because of the reverse situation, in which the anthers mature first (protandry), though these characteristics are usually only weakly developed. Both the male stamens and the female carpels are spirally arranged, and they mature from the margin toward the centre of the flower, so that both sets of organs may contain mature members at the same time. The stamens and stigmas (pollen-receiving surfaces of the female flower structure) are usually situated in such positions that a visiting insect has to touch them in order to obtain pollen or nectar, thus effecting pollination.

In the Berberidaceae family (at least in the two shrubby genera Berberis and *Mahonia*) a different mechanism is employed, involving sensitive stamens. Nectar is secreted in pits on the inner petals, and insects are attracted by the brightly coloured sepals and petals, which, in sunny conditions, are widely spreading. Each of the six inner petals has two nectary pits near the base. The stamens are opposite these petals, and, in sunny conditions, lie along them, the base of the filament thus lying above the line of junction of the two nectary pits. An insect visiting the flower to obtain nectar will necessarily touch the base of the filament. This area is sensitive to touch, and the stimulus causes the filament to spring rapidly into an erect position, thus depositing pollen on the body of the insect. The pollen may then be transferred to another flower by the insect, thus effecting cross-pollination.

Relatively little is known of the pollination mechanisms in the other families of the order, but certain structural modifications suggest that highly adapted mechanisms may be in operation. In the genus *Cissampelos* and other related members of the Menispermaceae family, for instance, the anthers of the six stamens are united into a ring. Cross-pollination is made more certain in this family by the occurrence of separate male and female flowers.

Seed dispersal. Seeds are dispersed in various ways in the Ranunculales order, depending on the nature of the fruit. In those genera with fleshy fruits (*e.g.*, the genus Actaea, and many members of the family Berberidaceae) the seeds are mainly animal-dispersed, whereas in the majority of the order, in which dry fruits prevail, seeds are dispersed by more passive mechanisms. The seeds of many species of Ranunculus have a hard beak, formed by the persistent style, and often, also, a spiny or warty surface; these are dispersed by adhering to the fur of animals. The seeds of genera with single-seeded fruits show adaptation to wind dispersal — those of the genus Thalictrum often being winged and flattened, and those of *Pulsatilla* and Clematis bearing a long, plumose appendage. Genera with many-seeded podlike fruits usually have a spilling mechanism, by which the seeds are gradually shaken out of the open pods when these are agitated by the wind. In Helleborus foetidus and perhaps some other species of Helleborus, the pods, when ripe, are pendent, hanging near the ground surface. All the seeds from each pod are attached to a persistent, fleshy part of the placenta, which becomes detached upon ripening, causing the seeds to fall out, though still attached to the ribbon of placenta; this fleshy tissue is attractive to ants; the ants drag about the strips of seeds, ensuring dispersal.

Evolution and palaeontology. Very few fossil Ranunculales are known, and only a few forms from the Tertiary (about 50,000,000 years ago) have been certainly identified as belonging to the group. These reveal little of the evolution of the order but do indicate that the Menispermaceae family was formerly much more widespread than today, various leaf impressions having been found in Europe, North America, and Greenland.

Possible ancestors of the buttercup order

Most phylogenists believe that the Ranunculales order is an early offshoot from a broadly conceived Magnoliales stock. Some authorities stress the close relationship between the Illiciales order and the Lardizabalaceae, Sargentodoxaceae, and Menispermaceae group of families of the Ranunculales order, and they also point out the close structural relationships between the Berberidaceae, Nandinaceae, Podophyllaceae, Ranunculaceae, and Menispermaceae families, suggesting a common origin for all of them from a possible ancestor of form similar to the order Illiciales. Other workers, however, believe that the Ranunculaceae family within the Ranunculales order could be ancestral to all the other families of the order and that all of the features in which the other families of the order differfrom the Ranunculaceae family appear to represent evolutionary derivations from that family. It is to be doubted, however, whether many phylogenists would accept this latter view, because the Ranunculaceae family itself includes a wide range of types, some generally considered to be primitive, others showing characteristics usually considered to be advanced. In the present state of knowledge, the evolutionary pathways and relationships within the order remain obscure.

CLASSIFICATION

Distinguishing taxonomic **features.** The Ranunculales order is distinguished from its allies by a combination of features, among which the most important is perhaps the widespread occurrence of nectar-producing petals. It differs from the Magnoliales and Illiciales orders by the herbaceous or softly woody habit, the absence of aromatic compounds, the usual absence of stipules, and various floral characters. It differs from the Nymphaeales and Nelumbonales orders by its predominantly terrestrial habit, leaf type, and various floral characters; and from the more advanced Papaverales order by the absence of a latex-conducting vessel system, the free carpels, and other floral characters. It differs from the Paeoniales order, which was included in the Ranunculales order in many older taxonomic systems, by the presence of nectar-producing petals, and by the sequence of stamen maturation.

Annotated classification.

ORDER RANUNCULALES

Herbs, woody climbers, or shrubs. Shoots dimorphic (of 2 kinds) in Berberis, the leaves on the long shoots being replaced by spines. Leaves usually alternate, rarely opposite; exstipulate (without stipules)except in a few genera; variously shaped and divided; deciduous or evergreen; spiny-margined in some genera. Flowers solitary or in inflorescencesof varying degrees of complexity, usually with bracts, usually containing both sexes with flower parts arising at the base of the ovary (hypogynous). Perianth of 1, 2, or 3 (rarely more) series, usually spirally arranged. When the perianth is more than 1-seriate it is usually differentiatedinto an outer series of sepals and 1 or more inner series of petals (honey leaves). The series of the perianth may be in patterns of 3, 4, 5, or larger indefinite numbers. Pcrianth usually regular (radially symmetrical or actinomorphic) but is irregular (zygomorphic) in some genera of the family Ranunculaceae. Nectar absent, or secreted by nectaries of various shapes, often petaloid, or rarely by the ovary (*e.g.*, in the genus *Caltha*). Stamens usually numerous, spirally arranged, and maturing from the margin inward. They are fewer than or equal in number to the petals and opposite them in many Berberidaceae. Stamens are usually free (anthers united in some Menispermaceae). Anthers open by lengthwise slits or by flaplike valves in some. Pollen granular, variable in appearance. Ovary of several to many usually free (united in some Ranunculaceae) carpels, rarely reduced to a solitary carpel.

Ovules several with marginal or sub-basal placentation, or reduced to one. Fruit a berry or a group of follicles or achenes. Seeds usually with endosperm. Alkaloids usually present.

Family Lardizabalaceae

The plants of this family are climbers or rarely shrubs (*Decaisnea*), with alternate, palmate or pinnate leaves, trimerous flowers having 3 or 6 sepals, 6 petals (usually small, rarely absent), 6 stamens, and 3 (rarely more) carpels. The flowers are often functionally unisexual, with stamens represented by sterile stamen structures in the female flowers. The fruit is a group of berries or fleshy follicles, and the seeds have copious, fleshy endosperm. Species of Akebia and Decaisnea are grown as ornamentals; the blue, fleshy follicles of Decaisnea *fargesii* are very noticeable in autumn and never fail to attract attention. Eight genera and about 30 species with a disjunct distribution, most of the genera occurring in the Himalayas and eastern Asia; 1 (*Lardizabala*) is found in Chile.

Family Sargentodoxaceae

A family consisting of a single genus and species occurring in China and Southeast Asia. The only known species is a climber with alternate, trifoliolate leaves and with all floral parts spirally arranged. Sepals 6, petals 6, very small, stamens 6, opposite the petals, carpels numerous. The fruits are berry-like, and each one is stalked, the whole group being borne on the elongated receptacle, which is markedly thickened. The one species, Sargentodoxa cuneata, is sometimes cultivated as a cool-greenhouse climber in temperate areas.

Family Menispermaceae

A large family of woody climbers (rarely erect shrubs), many showing anomalous forms of secondary thickening. Leaves alternate, entire or lobed; flowers unisexual, usually small, sometimes strongly zygomorphic. Sepals 3 to 12 usually, but occasionally as few as 1 or much more numerous than 12; petals various, usually small, often absent; stamens 2 to indefinitely large numbers; when few, they are opposite the petals (if petals are present), anthers often united; carpels usually 3 to 6 but as few as 1, or very numerous. Seeds characteristically half-moon or horseshoe shaped. A few species of Stephania, Cocculus, and Cissampelos are grown as foliage ornamentals. Found more or less throughout the tropics with, however, some genera extending into temperate areas (*e.g.*, Japan, United States, Canada). About 65 genera and 430 species.

Family Ranunculaceae

The largest family in the order, consisting of herbs usually, (Xanthorrhiza, however, is a shrub with soft yellow wood), or rarely climbers (Clematis). Leaves usually alternate (opposite in Clematis), without stipules (with stipules in Thalictrum; Actaea and some others, however), simple to much divided, often with sheathing bases. Flowers actinomorphic or zygomorphic. Sepals 2 to 5 usually, but often numerous and often petaloid; petals zero to 5 usually, but often numerous, nectariferous, sometimes reduced to nectar-producing tubes or sacs. Stamens usually very numerous. Carpels numerous, but occasionally as few as 1, several- to 1-seeded. Fruit a group of achenes, follicles, or a berry. Widely distributed in temperate and subtropical regions of both hemispheres, mainly on mountains in the tropics. About 50 genera with some 2,000 species.

Family Glaucidiaceae

A family of 1 genus and 1 or 2 species from western China and Japan, often included in the Ranunculaceae family. The species are perennial herbs growing from rhizomes with few, palmately lobed, alternate leaves. Sepals 4, petaloid, usually blue; petals absent; stamens numerous; carpels 2. Fruit a group of follicles. Superficially the plants resemble the blue Himalayan poppies (Meconopsis in the family Papaveraceae, order Papaverales), and may indicate a relationship between the two orders. *Glaucidium palmatum* is frequently cultivated as a garden ornamental.

Family Hydrastidaceae

A family of 1 genus with 2 species, from North America. They are rhizomatous herbs with few, simple, cordate leaves. The flowers are small, with 3 sepals, and without petals. The fruit is a raspberry-like group of berries. Hydrastis canadensis, which was once important in folk medicine, has been exterminated by collectors in some parts of the U.S. The genus is usually included in the Ranunculaceae family.

Family Circaeastraceae

Another monotypic family, the one species occurring in the northwest Himalayas. It is an annual herb with spiny-toothed, rather barberry-like leaves, and very reduced flowers with 2 sepals, 2 (sometimes one) stamens, and 1 carpel. Its position is controversial, many authors including it in the Ranunculaceae family.

Family Podophyllaceae

A small family, whose circumscription varies greatly from one authority to another; many treat the group as part of the Berberidaceae family. The family includes 2 genera, each with few species, from eastern Asia and Atlantic North America. The plants are herbaceous with lobed leaves; the flowers have usually 6 sepals, 6 petals, and a single-chambered ovary, with ovules borne in 2 or more series on the adaxial suture. *Podophyllum* is sometimes grown as an ornamental; the fruits of P. peltatum (mayapple) are edible.

Family Nandinaceae

A monotypic family with its one species found in China and Japan. It is a small shrub with alternate, 2 to 3 pinnate leaves and small white flowers in large panicles. Sepals numerous, spirally arranged, petals 6, stamens 6, opposite the petals, carpel 1 with 1 pendent ovule. The one species, Nandina domestics, is frequently grown as an ornamental shrub, mainly for its attractive foliage, which is reddish when young, and which colours well in autumn.

Family Berberidaceae

A large family of 12 genera and about 650 species, most of these (about 500) belonging to the genus Berberis. The circumscription of the family varies greatly with authority, depending on how each individual author treats the Podophyllaceae family. As presented here, the family includes both herbaceous and shrubby genera from most temperate parts of the world. Leaves simple or variously compound, alternate (leaves on the long shoots in the genus Berberis represented by spines), spiny-margined in the shrubby genera. Flowers are variable. Sepals few to numerous, often in 1 to 3, trimerous whorls; petals present or absent, when present nectar-producing, spurred in some genera. Stamens 4–6, opposite to the inner petals, sensitive to the touch in shrubby genera, anthers often opening by flaplike valves. Ovary of 1 carpel; ovules numerous, placentation lateral or basal. Fruit usually fleshy.

Species of Berberis, Epimedium, *Jeffersonia, Mahonia, Leontice*, Ranzania and Vancouveria are cultivated as garden ornamentals.

Disagreements on relationships in the order

Critical appraisal. The contents of this order are not generally agreed upon among taxonomists, and various alternative classifications have been proposed. Specifically, there is no general agreement as to precisely what groups should be included in the Ranunculales order and how the groups should be related to each other; and there is little agreement as to which groups of genera should be recognized at family level.

The following points of agreement should be noted, however. The core of the order consists of the families Lardizabalaceae, Sargentodoxaceae, Menispermaceae, Ranunculaceae, Glaucidiaceae, Hydrastidaceae, Circaeastraceae, Podophyllaceae, Nandinaceae, and Berberidaceae; that is, those families comprising the order as presented here. The other groups in the order in other classifications—viz., the families Paeoniaceae, Ceratophyllaceae, Cabombaceae, Nymphaeaceae, Corynoçarpaceae, Coriariaceae and Sabiaceae—are more often placed in other orders, although the exact relationships of many of them are still controversial.

Within the order, as presented here, however, there is great variability, and it may be doubted whether or not it should be maintained as a whole. The division of it into two orders, Ranales and Berberidales, has been proposed, but this has not been widely accepted, and it may be an improvement to recognize its division into even three groups.

Regarding the placement of the Ranunculales among other orders, the close relationship of Magnoliales, Illiciales, and Ranunculales is generally agreed upon among phylogenists, as is a close relationship of the Ranunculales order with the Nymphaeales and Nelumbonales orders. A relationship with the primitive members of the flowering plant class Monocotyledoneae is postulated by some authors, whereas others suggest that the similarities between Ranunculales and Nymphaeales orders of the dicots and the Alismales order of the monocots are the result of convergent evolution. A fairly close relationship between members of the Ranunculales order (particularly among the Berberidaceae, Glaucidiaceae, and Podophyllaceae families) and the Papaverales order is widely accepted; this is based mainly on structural features, strengthened, however, by the occurrence of the same, or closely related, alkaloids in both groups. A close relationship between the Ranunculales and the Paeoniales orders is not now generally accepted.

BIBLIOGRAPHY. L.W.A. AHRENDT, "A Taxonomic Revision of *Berberis* and *Mahonia," J. Linn. Soc. Botany,* 57:1–410 (1961), the most complete and up-to-date account of these two woody genera; A. CRONQUIST, *The Evolution and Classification of Flowering Plants* (1968), a recent classification of all the orders of flowering plants for the advanced botanist; L. DIELS, "Menispermaceae," in A. ENGLER (ed.), *Das Pflanzenreich IV,* 46 (1910), the classical treatment of this family; A.S FOSTER, "Floral Morphology and Relationships of *Kingdonia," J. Arnold Arbor.,* 42:397–410 (1961), and "Floral Morphology of *Circeaster," ibid.,* 44:299–321 (1963), detailed studies on two of the smaller groups within the order; W.C. GREGORY, "Phylogenetic and Cytological Studies in the Ranunculaceae," *Trans. Am. Phil. Soc., n.s.,* 31:443–521 (1941), a classical account of the cytology, taxonomy, and evolution of the family; G.H.M. LAWRENCE, *Taxonomy of Vascular Plants* (1951), an authoritative descriptive treatment, with illustrations of higher plant families.

(J. Cul.)

Raphael

Working in the spirited atmosphere of the early 16th century, in the culture that came to be called the High Renaissance, Raphael was the most celebrated of those who brought Italian art to its highest level. He seemed to incarnate the spirit of Humanism—with its reverence for classical art and thought, its individualistic and inquiring spirit—and to give it incomparable expression.

At the end of the 15th century the states of Italy were in a politically critical condition, and Italian territory was about to become the battleground of opposing foreign armies. Early forewarnings of the turmoil of the Reformation were also disturbing its religious consciousness. Indeed, Raphael's work has been considered as the symptom of a moral and historical situation, the ultimate refinement of a world that had reached its climax and would before long find itself unable to survive: Raphael is thus the painter and architect of a moment of crisis whose works speak of his absolute assurance of artistic means and of the universality of his conceptual expression.

Early years at Urbino. Raphael was born at Urbino in the Marche, or marches, on April 6, 1483, the son of Giovanni Santi and Magia di Battista Ciarla. His mother died a few years later, in 1491. His father was, according to the 16th-century artist and biographer Giorgio Vasari, a painter "of no great merit." He was, however, a man of culture who was in constant contact with the advanced artistic ideas current at the court of Urbino. He gave his son his first instruction in painting, and before his death in 1494, when Raphael was 11, he had introduced the boy to Humanistic philosophy at the court, where the cultural achievements of the past and new ideas looking forward to the 16th century could be learned.

Early influences

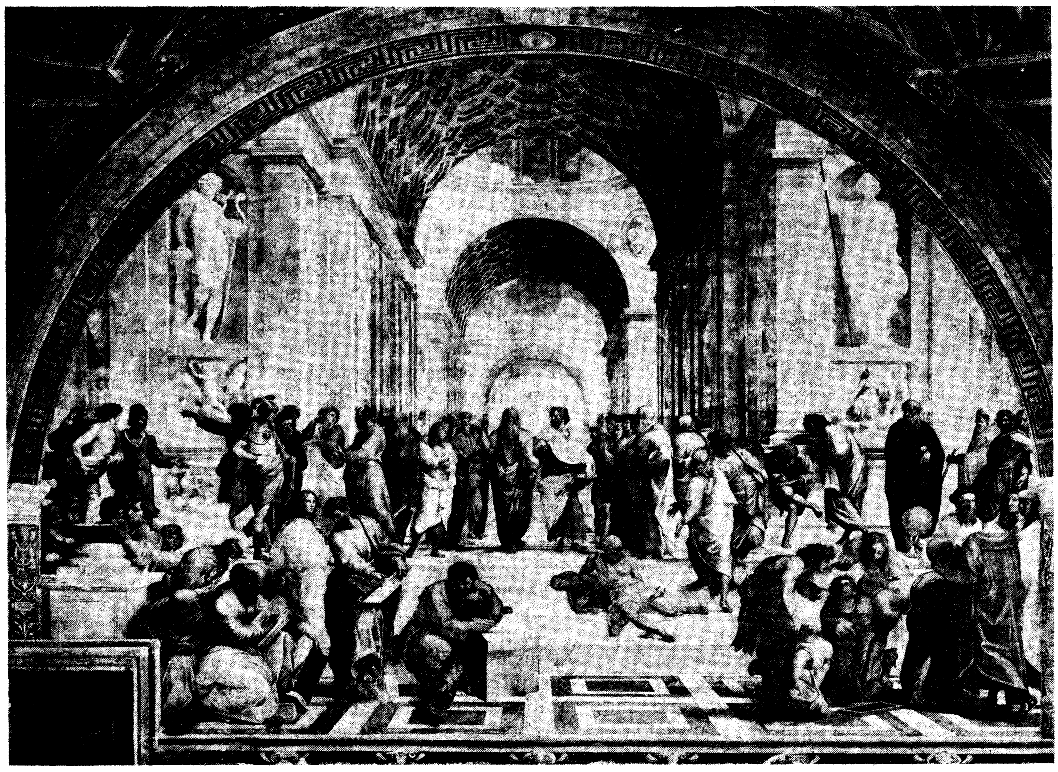
Urbino had become a centre of culture during the rule of Duke Federico da Montefeltro, who died seven months before Raphael was born. Duke Federico, the embodiment of the Renaissance ideal of the cultured prince, had greatly encouraged all the arts and had attracted the visits of men of outstanding talent to the court he kept in the beautiful ducal palace he had had reconstructed; such men included two Urbinates, the architect Donato Bramante, who was to befriend Raphael in Rome later, and the writer Baldassare Castiglione, whose *Il Cortegiano (The Courtier)* celebrated the reign of Guidobaldo and was universally acclaimed as the model for perfect conduct. Neither Bramante nor Castiglione were members of Federico's court; nor were Piero della Francesca or Leon Battista Alberti, two of the dominant spirits in the growth of the visual arts in Urbino during this period. The style of Piero, a painter with a strong mathematical bent, was diametrically opposed to that of Raphael, as it created spatial compositions of extreme austerity that emphasized perspective and the use of human forms as geometrical elements. Alberti was one of the major theorists in the development of perspective in painting and was a strong advocate of the idealization of nature and of architectural proportions based on musical harmonics. The foundation of Raphael's style was the system of artistic principles established by these artists and others such as Justus of Ghent, Ucello, Francesco di Giorgio, and the influence of Flemish painting. Later, in Florence with Leonardo da Vinci, he would synthesize from it a new system of painting as organic and ordered as that of the 15th-century artists, though less crystalline; the principles of order and proportionality he learned at Urbino would be developed by him in a personal and original manner. Although in Perugia and in Florence Raphael would be influenced by other major artists, Urbino constituted the basis for all his subsequent learning. Furthermore, the cultural vitality of the city probably accounted for the exceptional precociousness of the young artist, who, even at the beginning of the 16th century, when he was scarcely 17 years old, already displayed an extraordinary talent.

Apprenticeship at Perugia with Perugino. Vasari's statement that Giovanni Santi accompanied his son to Perugia and apprenticed him there to the great Umbrian painter Pietro Perugino cannot be accepted, for Raphael was hardly 11 at the time of his father's death. Nevertheless, the date of Raphael's arrival in Perugia cannot be long postponed, and several scholars place it definitely in 1495. What is certain, as attested by a notary's deed, is that he was no longer in Urbino on May 13, 1500. Furthermore, another document of December 10, 1500, declares that the young painter, by then called a "master," was commissioned to paint, together with a pupil of his father, an altarpiece to be completed by September 13, 1502. It is clear from this that Raphael's apprenticeship to Perugino was already under way and that he had given immediate proof of his mastery, so much so that between 1501 and 1503 he received a rather important commission—to paint the "Coronation of the Virgin" for the Cappella Oddi in S. Francesco, Perugia, where it remained until 1797, when it was stolen by Napoleonic troops. It was returned to the Vatican in 1815. Perugino was executing the frescoes in the Collegio del Cambio at Perugia between 1498 and 1500, enabling Raphael, as a member of his workshop, to acquire extensive professional knowledge. The art of Perugino, Raphael's first effective teacher, was in harmony with the Humanistic philosophy of the court of Urbino, especially in its iconography.

First major work

In addition to this practical instruction, Perugino's calmly exquisite style also influenced Raphael. The "Giving of the Keys to St. Peter," painted in 1481–82 by Perugino for the Sistine Chapel of the Vatican Palace in Rome, inspired Raphael's first major work, "The Marriage of the Virgin" (1504). Perugino's influence is seen in the emphasis on perspectives, in the graded relationships between the figures and the architecture, and in the lyrical sweetness of the figures. Nevertheless, even in this early painting, it is clear that Raphael's sensibility was different from his teacher's. The disposition of the figures is less rigidly related to the architecture, and the disposition of each figure in relation to the others is more informal and animated. The sweetness of the figures surpasses even that characteristic of Perugino, and the gentle relation between them does not exist in any comparable degree in Perugino's paintings. Three small paintings done by Raphael shortly after "The Marriage of the Virgin"—"Vision of a Knight," "Three Graces," and "St. Michael"—are masterful examples of narrative painting, showing, as well as youthful freshness, a maturing ability to control the elements of his own style. Although he had learned much from Perugino, Raphael by late 1504 needed other models to work from; it is clear that his desire for knowledge was driving him to look beyond Perugia.

Move to Florence. Vasari vaguely recounts that Raphael followed the Perugian painter Bernardino Pinturicchio to Siena and then went on to Florence, drawn there by accounts of the work that Leonardo and Michelangelo were undertaking in that city (their respective designs of the battles of Anghiari and Cascina for paintings in the Palazza della Signoria). By the autumn of 1504 Raphael had certainly arrived in Florence. It is not known if this was his first visit to Florence, but, as his works attest, it was about 1504 that he first came into substantial contact with this artistic civilization, which reinforced all the ideas he had already acquired and also opened to him new and broader horizons. He took in all that he saw, not passively but, as usual, to reaffirm his understanding of the



"School of Athens," fresco by Raphael. In the Vatican.
Anderson—Alinari

canons of Renaissance culture. With his incredible capacity to synthesize, Raphael learned from the works of a range of artists—from the old masters of the 15th century to his two great elder contemporaries, Leonardo da Vinci and Michelangelo. Giorgio Vasari records that he studied not only the works of Leonardo, Michelangelo, and Fra Bartolomeo della Porta, a Florentine proponent of the generalized idealism of the High Renaissance, but also "the old things of Masaccio," a pioneer of the naturalism that marked the departure of the early Renaissance from the Gothic. Some of his designs show that he also looked back to the sculptors Donatello and Verrocchio and the engraver Antonio Pollaiuolo, all of whom were avid students of ways of rendering movement in the human form. Thus, by constantly re-examining and re-evaluating the works of his contemporaries and predecessors, he was able to determine which of the artistic principles of the last 100 years were truly classic and which should be disregarded.

Still, his principal teachers in Florence were Leonardo and Michelangelo. Many of the works that Raphael executed in the years between 1505 and 1507, most notably the Madonnas, such as "The Madonna of the Goldfinch," the "Madonna del Prato," the "Esterházy Madonna," and "La Belle Jardinière," are marked by the influence of Leonardo, who since 1480 had been making great innovations in painting. These paintings were influenced by Leonardo's compositions, figure placement, and gestures, all of which are marked by an intimacy and simplicity of setting uncommon in 15th-century art. Raphael also owed much to Leonardo's lighting techniques; he made moderate use of Leonardo's chiaroscuro (*i.e.*, strong contrast between light and dark, seemingly resulting from a natural or fixed light source), and he was especially influenced by his sfumato (*i.e.*, use of extemely fine, soft shading instead of line to delineate forms and features). Raphael also became a master at this technique. Raphael went beyond Leonardo in creating new figure types whose round, gentle faces reveal uncomplicated and typically human sentiments but raised to a sublime perfection and serenity. At the same time, the extraordinary capacity to synthesize that characterizes Raphael enabled him to resolve a variety of treatment—from Masaccio to Michelangelo—of perspectives of movement, dramatic foreshortenings, and the plastic potentialities of the human figure.

Leonardo's influence on Raphael

Between 1504 and 1508 Raphael did not reside continuously in Florence but moved about, even returning to Urbino and, naturally, to Perugia. And it was a Perugian noblewoman who commissioned him, in 1507, to paint the "Deposition" that is now in the Borghese Gallery in Rome, in which the obvious influences of Michelangelo and also of Leonardo are transformed and worked up into an original synthesis.

In general, Raphael differed from Leonardo and Michelangelo, both painters of dark intensity and excitement, in that he wished to develop a totally fresh and completely calm, extrovert communication; through it he wished to achieve a more popular and universal form of expression. Meanwhile, although scarcely 25 years old, he was already in full command of his artistic means and was absolutely certain of his aesthetic principles. He was now ready to impose his style on Rome---on a society as astute and as critically apprised as any in Italy.

Last years in Rome. Raphael was called to Rome toward the end of 1508 by Pope Julius II at the suggestion of Bramante, who wanted to bring his fellow townsmen into the papal court. At this time Raphael was little known in Rome, but the young man soon made a deep impression on the volatile Julius and the papal court, and his authority as a master grew day by day. Identifying himself with Rome and the spirit of its aristocracy, he became so popular that he was called "the prince of painters." The Humanist Celio Calcagnini later said of him in a Latin epigram:

> It took many ancient heroes and a long age to build Rome, and many enemies and centuries to destroy it. Now Raphael has sought and discovered Rome in Rome: it takes a great man to seek, but discovery comes of God Himself.

The epigram is testimony to the success of the style by which the artist gained the respect of the whole city. The Roman aristocracy considered him one of its own; Cardinal Bibbiena in 1514 wished to give his niece in marriage to him; according to Vasari, Pope Leo X, the successor of Julius II and the cultured son of the Medici Lorenzo the Magnificent, had had the intention of making him a cardinal.

Raphael spent the last 12 years of his short life in Rome.

They were years of feverish activity and successive masterpieces. No test frightened him; he faced every task with modesty and deep studiousness, but he aimed at new heights each time and triumphed. Begun at the end of 1508 and completed in less than three years, the decoration of the Stanza della Segnatura in the Vatican Palace was perhaps his greatest work, reflecting the ideology and culture of Julius' pontificate and glorifying the Roman Church in history. The Stanza della Segnatura is one of the rooms in the Vatican papal apartments in which Julius II, who commissioned the decoration, was to live and work. Julius II was a highly cultured man who surrounded himself with the most illustrious personalities of the Renaissance. He entrusted Bramante with the construction of a new basilica of St. Peter to replace the original 4th-century church; he called upon Michelangelo to execute his tomb and compelled him against his will to decorate the ceiling of the Sistine Chapel; and, sensing the genius of Raphael, he committed into his hands the interpretation of the philosophical scheme of the frescoes in the Stanza della Segnatura, the theme of which he had discussed with the Humanists of his court. This theme was the historical justification of the power of the Roman Church through Neoplatonic philosophy. The two most important of these frescoes are called the "Disputa" and the "School of Athens." The "Disputa," showing a celestial vision of God and his prophets and apostles above a gathering of representatives, past and present, of the Roman Church, equates through its iconography the triumph of the church and the triumph of truth. The "School of Athens" shows Plato and Aristotle surrounded by philosophers, past and present, in a modern architectural setting; it illustrates the historical continuity of Platonic thought. In these works, the order conveyed by strong perspective and gradated figure placement and grace in gesture and expression are conjoined with an erudite iconography to make powerful philosophical statements. In executing this cycle, Raphael accepted and interpreted the spirit of Julius' ideology and the will to triumph that animated it and made it his own will and the foundation of the ordered, rational perfection of his artistic language.

Frescoes in the Stanza della Segnatura

The decoration of the apartments continued after the death of Julius in 1513 and into the succeeding pontificate of Leo X until 1517. In spite of this great undertaking, the last sections of which were left almost completely to his pupils, Raphael undertook at the same time a variety of other ambitious tasks—sacred and secular decorations in other buildings, portraits, altarpieces (in which his gentle saints and madonnas introduce a new typology), cartoons for tapestries, designs for dishes, and the painting of stage scenery. While he was at work in the Stanza della Segnatura, he also did his first architectural work, designing the church of S. Eligio degli Orefici. In 1513 the banker Agostino Chigi, whose Villa Farnesina Raphael had already decorated, commissioned him to do his funerary chapel in the church of Sta. Maria del Popolo. There Raphael showed himself to be a complete artist: he designed the chapel and a sculpture that was to adorn it and executed the cartoons for the mosaics in the cupola, creating a complex harmony of the different arts. In 1514 Leo X chose him to work on the basilica of St. Peter's alongside Bramante; and when Bramante died later that year, Raphael assumed the direction of the work, transforming the plans of the church from a Greek, or radial, to a Latin, or longitudinal, design. He also succeeded Brarnante in the decoration of the Vatican loggias, or galleries, in which he engaged himself and his whole workshop. The sweet simplicity of this lyrical decoration seemed to counterbalance the terrifying grandeur of the Sistine Chapel of Michelangelo.

Architectural work

As a result of the philosophical depth of many of Raphael's Roman works, his reputation as a Humanist and Neoplatonist spread through Rome. He counted among his friends there several men of letters—among them, Castiglione, Cardinal Bembo, the satirist Pietro Aretino, and Bibbiena—as well as many artists. In 1519 he designed the sets for the comedy I Suppositi, by Italy's epic poet Ludovico Ariosto. Raphael was a competent scholar, particularly interested in classical antiquity. In August 1515 Pope Leo put him in charge of the supervision of the preservation of marbles with valuable Latin inscriptions; two years later he was appointed commissioner of antiquities for the city, and he drew up an archaeological map of Rome. In this connection, the existence of a letter addressed to the Pope, the content of which, though indited by Castiglione, is said to be Raphael's (though a few scholars would like to attribute it to Bramante), is extremely significant in its preoccupation with the examination, measurement, and classification of ancient monuments according to their artistic and constructional qualities. His admiration of antiquity did not impede his projecting artistic schemes for future execution. At the end of his career, he was creating designs that looked beyond the High Renaissance to a new mode of expression. His plans for the Villa Madama in Rome, the construction of which was begun after 1516, exemplified a new architectural conception of enormous importance to the subsequent development of Italian architecture. Similarly prophetic in his painting, he revealed in his last work, the "Transfiguration" (commissioned in 1517), a new sensibility that is like the prevision of a new world, turbulent and agitated; in its composition it tends toward an expression that is already Baroque.

Raphael died on his 37th birthday, April 6, 1520. A mythical halo had surrounded him during the last years of his life, and the whole papal court mourned him. His funeral mass was celebrated at the Vatican, his "Transfiguration" was placed at the head of the bier, and his body was buried in the Pantheon in Rome.

MAJOR WORKS

PAINTINGS: "Altarpiece: the Crucified Christ with the Virgin Mary, Saints and Angels" ("Mond Crucifixion," c. 1501; National Gallery, London); "An Allegory" ("Vision of a Knight," c. 1501; National Gallery, London); "Three Graces" (c. 1501; Musée Condé, Chantilly, France); "St. Michael" (*c.* 1501;' Louvre); "Coronation of the Virgin" (*c.* 1501–03; Vatican Museum, Rome); "The Marriage of the Virgin" (1504; Brera, Milan); "Portrait of Agnolo Doni" (c. 1505; Pitti Palace, Florence); "Portrait of Maddalena Doni" (*c.* 1505; Pitti Palace); "The Madonna of the Goldfinch" (*c.* 1505; Uffizi, Florence); "The Madonna and Child with St. John" ("Madonna del Prato," c. 1505; Kunsthistorisches Museum, Vienna); "Esterházy Madonna" (c. 1505–07; Museum of Fine Arts, Budapest); "The Ansidei Madonna" (c. 1506; National Gallery, London); "The Deposition of Christ" (1507; Borghese Gallery, Rome); "La Belle Jardinière" (1507; Louvre); "The Niccolini-Cowper Madonna" (1508; National Gallery of Art, Washington, D.C.); "Stanza della Segnatura" (1508–11; Vatican, Rome); "Stanza d'Eliodoro" (1512–14; Vatican, Rome); "Triumph of Galatea" (1511–13; Villa Farnesina, Rome); "Sistine Madonna" (1513?; Gemäldegalerie, Dresden); "Stanza dell'Incendio" (1514–17; Vatican, Rome); Tapestry cartoons for Leo X (c. 1516; Victoria and Albert Museum, London); "Portrait of Balthasar Castiglione" (1516; Louvre); "Transfiguration" (1517–c. 1520; Vatican Museum, Rome); "The Holy Family of Francis I" (1518; Louvre); "St. Michael Vanquishing Satan" (1518; Louvre).

ARCHITECTURE: Chigi Chapel (c. 1513–14; Sta. Maria del Popolo, Rome); S. Eligio degli Orefici (c. 1516; Rome); Palazzo Pandolfini (c. 1516; Florence).

BIBLIOGRAPHY. V. GOLZIO (ed.), Raffaello nei *documenti*, nelle testirnonianze dei contemporanei e *nella* letteratura del suo secolo, (1936), the most important collection of sources and documents, including the biographies by Giorgio Vasari and Paolo Giovio and the letters and sonnets of the artist; OSKAR FISCHEL, "Santi Raffaello," in THIEME-BECKER, *Allgemeines* Lexikon der *bildenden Künstler*, vol. 29 (1935), a study that includes a complete bibliography on Raphael up to 1935; and Raphaels Zeichnungen, 2nd ed. (1898; Eng. trans., Raphael, 2 vol., 1948; reprinted in 1 vol., 1964); A.M. BRIZIO, "Raphael," in the Encyclopedra of World Art, vol. 11, col. 839–869 (1966), a study that includes a complete bibliography after 1935; RICHARD COCKE and PIERLUIGI DE VECCHI, The Complete Paintings of Raphael (1970, a translation of *L'opera completa* di Raffaeilo,presented by MICHELE PRISCO), critical apparatus and philology by PIERLUIGI DE VECCHI (1966), contains a general catalog of the artist's works; SIDNEY J. FREEDBERG, Painting of the High Renaissance in Rome and Florence (1961); MARIO SALMI (ed.), The *Complete* Workr of Raphael (1969); LUITPOLD DUSSLER, Raphael: A Critical Catalogue of his Pictures, Wall-Paintings and Tapestries (1971, translation of 1966 German publication).

Rare-Earth Elements and Their Compounds

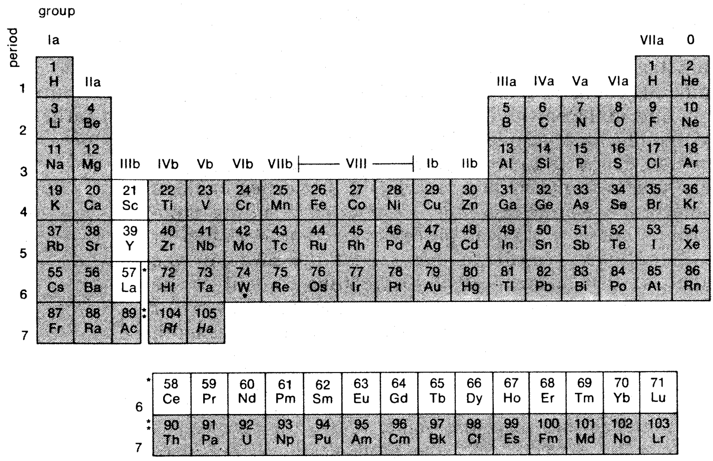
The rare-earth elements form a series of 17 chemically similar metals, all but one of which occur in nature. Often they are called simply rare earths, but this is a misnomer because the term earth properly is applied to the oxide of a metal rather than to the element itself. The rare-earth elements are not even particularly rare, though for a long time they were thought to be.

The 17 rare-earth elements (with their chemical symbols) are: scandium (Sc), yttrium (Y), lanthanum (La), cerium (Ce), praseodymium (Pr), neodymium (Nd), promethium (Pm), samarium (Sm), europium (Eu), gadolinium (Gd), terbium (Tb), dysprosium (Dy), holmium (Ho), erbium (Er), thulium (Tm), ytterbium (Yb), and lutetium (Lu).

Until the mid-20th century, there was not much use for pure rare-earth elements or compounds except cerium and lanthanum; mixtures of the rare earths, however, had found metallurgical and other uses. By the 1970s two of these elements, yttrium and europium, were being used in the red phosphors for colour-television tubes.

Positionin the periodic table

In the periodic table of the elements (see Figure). the rare-earth elements comprise three members of Group IIIb and all 14 members of one of two series of elements generally written apart from the main table. This long series is known collectively as the lanthanide series because it directly follows lanthanum in a different form of the table. The periodic table is based on the electronic structure of the atoms of the various elements, and this structure also determines the chemical behaviour of the elements themselves. As shown below (see *Electronic structure),* the rare-earth elements all have certain common features in the structure of their atoms, the fundamental reason for their chemical similarity.



period	Ia	IIa	IIIb	IVb	Vb	VIb	VIIb	VIII	VIII	VIII	Ib	IIb	IIIa	IVa	Va	VIa	VIIa	0
1	1 H																1 H	2 He
2	3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba	57 La *	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra	89 Ac :	104 *Rf*	105 *Ha*													

6 *	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu
7 :	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr

The positions of the rare-earth elements in the periodic table

The aqueous chemistry of all the rare earths is very similar and changes only slightly in progressing along the lanthanide series. Because of this similarity, it is difficult to separate individual rare earths. In the few cases in which the rare-earth ion can be oxidized or reduced to another valency, however, chemical separations can be carried out readily. Also, artificial mixtures of elements far apart in the series can be separated easily.

All of these elements form trivalent compounds, and in the crystal lattices (the regular arrangement of atoms in the solid forms) of such compounds, one rare-earth ion readily replaces another. The rare-earth metals when heated react strongly with nonmetallic elements to form very stable compounds. They are never found as the free metals in the earth's crust. Pure minerals of individual rare earths do not exist in nature; all their minerals contain mixtures of the rare-earth elements.

Promethium is never found in the earth's crust since it has no stable isotopes and is produced only by nuclear reactions; it can, however, be obtained in quantity from the fission products formed in nuclear reactors.

The chemical properties of scandium differ sufficiently from those of other rare-earth elements for it to have become segregated from them by the action of geological processes. Scandium seldom is associated with the rare earths in minerals.

History. The early Greeks believed that all matter was made up of four elements: air, earth, fire, and water. Earths were defined as materials that could not be changed further by the sources of heat then available. Until late in the 18th century, this Greek conception remained strong in chemistry, and oxides of metals such as calcium, aluminum, and magnesium were known as earths and were thought to be elements.

Discovery of the first rare earth

In 1794, Johann Gadolin, a Finnish chemist, while investigating a rare Swedish mineral, discovered a new earth in impure form, which he believed to be a new element and to which he gave the name ytterbia, from Ytterby, the village where the ore was found. The name, however, was soon shortened to yttria. In 1803, from the same mineral, later named gadolinite in Gadolin's honour, another new earth was reported in the literature independently by several chemists. The new earth became known as ceria, from the asteroid Ceres, which had only just been discovered (1801). Since yttria and ceria had been discovered in a rare mineral, and they closely resembled other known earths, they were referred to as the rare earths. Not until 1808 did the English chemist Sir Humphry Davy demonstrate that the earths as a class were not elements themselves but were compounds of oxygen and metallic elements. Later, a number of chemists verified the existence of ceria and yttria in gadolinite and found that these oxides were also present in a wide variety of other rare minerals. The elements of which yttria and ceria were the oxides were then given the names yttrium and cerium, respectively.

In the period from 1839 to 1843, Carl Gustaf Mosander, a Swedish chemist (and student of Berzelius), found that yttria and ceria were not even the oxides of single elements but were, in fact, mixtures. He reported that if the oxides were dissolved in strong acid and the resulting solution subjected to a long series of fractional precipitations as various salts (oxalates, hydroxides, and nitrates), two new elemental substances could be split off from the main component of each oxide. The two new oxides found in ceria he called lanthana and didymia, and the elements contained in them were named lanthanum and didymium. The new elements found (as their oxides) in yttria he called erbium and terbium, and the oxides were referred to as erbia and terbia. Mosander also was the first to obtain the rare-earth metals themselves from their oxides, although only in impure form. Mosander's researches puzzled the scientists of his time. He seemed to be finding a new group of elements of an entirely different type from any known previously. All formed the same classes of compounds with almost the same properties, and the elements could be distinguished from one another—at that time—only by slight differences in the solubilities and molecular weights of the various compounds.

In the next few years the literature on the rare earths became confused. There was, for example, considerable controversy for a number of years over the existence of didymium. The situation was considerably clarified in 1859 when an instrument called the spectroscope was introduced into the study of the rare earths. This instrument indicated the aatterns of light emission or absorption characteristic of the elements, and, with it, didymium was shown to have a characteristic absorption spectrum. From then on determination of spectra of various types became one of the most important tools in following the progress of the fractionation of rare earths. Somehow during this period the names used for the various fractions differed from laboratory to laboratory. Around 1860, by general agreement, it was decided to interchange the names of Mosander's earths, erbia and terbia.

Scandium and the periodic table

In 1869, when the Russian chemist Dmitry Ivanovich Mendeleyev first proposed the periodic table, he found it necessary to leave a blank at the position now occupied by scandium. He predicted in 1871, however, that a new element would be found to fit that blank in the table, and he also predicted certain properties of the element. The discovery of scandium a few years later (1879) and the agreement of its properties with those predicted by Mendeleyev helped to bring about general scientific acceptance of Mendeleyev's periodic table. Interestingly enough, one of the greatest weaknesses in the table was that it provided no logical place for the lanthanides, a difficulty that was not resolved for some years.

From 1843 to 1939 chemical fractionation of the mixed rare-earth salts obtained from many minerals was intensively investigated in both Europe and North America. Mosander's didymia was resolved into several oxides—samaria (samarium; 1879), praseodymia (praseodymium; 1885), neodymia (neodymium; 1885), and europia (europium; 1901). His terbia and erbia were resolved into holmia (holmium; 1878), thulia (thulium; 1879), dysprosia (dysprosium; 1886), ytterbia (ytterbium; 1876), and lutetia (lutetium; 1907).

During this period many of these elements were discovered independently by more than one investigator, but the credit for the discovery was usually given to the man who first separated sufficient quantities of the oxide to determine some of its properties and who published his results first.

As the scientists carried out their fractionations, they frequently observed changes in colour, apparent molecular weight, and spectra of the substances. Such changes were mainly responsible for the more than 70 claims for the discovery of new rare-earth elements during this period. Many of the observed changes were brought about by the concentration of different impurities, particularly the transition elements, in various fractions of the series. It is now known that such trace impurities in the rare-earth oxides can give rise to such colour changes and that such oxides can be made to fluoresce strongly and exhibit unique spectra.

Contributions of Welsbach

Shortly after Auer von Welsbach isolated praseodymia and neodymia in 1885 he invented an illuminating device that bears his name (Welsbach gas mantle), and a little later he produced a practical lighter flint. Both devices depended upon rare-earth elements. Although minerals rich in rare earths had up to that time been thought to be very rare, the demand for rare earths that developed as a result of Auer von Welsbach's inventions resulted in a worldwide search for rare-earth minerals, and it was found that one of them, monazite, existed in extensive deposits. Monazite, a phosphate of several rare-earth elements, was ideal for Auer von Welsbach's purposes, because it contained a high percentage of the element thorium, which was also used for the mantles. These were prepared by impregnating a cloth fabric with a solution of about 90 percent thorium nitrate, 10 percent cerium trinitrate, and minor amounts of other salts. When heated by a gas flame, these salts were converted to their oxides, which, when heated by the flame, gave off an intense white light. Cerium and iron form an alloy that emits sparks when struck. The discovery of this alloy by Auer von Welsbach started the flint industry. In 1913 about 3,300 tons of monazite were refined to produce the thorium and cerium used in gas mantles and the mixed rare-earth metals for flints and related products.

Atomic numbers of the rare-earth elements

The British physicist H.G.J. Moseley, while studying the X-ray emission spectra of the elements in 1913–14, found a direct relationship between the X-ray frequencies and the atomic numbers of the elements. This relationship made it possible to assign unambiguous atomic numbers to the elements and to verify their locations in the periodic table. In this way, Moseley was able to show clearly that there could be only 14 lanthanides, starting with cerium and ending at lutetium, and, at that time, all of the rare-earth elements had been discovered except for element 61. Because no stable isotopes (forms of the element with differing mass) of this substance exist in nature, it was not isolated until 1945, when one of its radioactive isotopes was separated from atomic fission products produced in a nuclear reactor. The element was named promethium after the Greek Titan who stole fire from the gods and gave it to mankind.

As in most fields of science, the present state of knowledge concerning the rare earths is the result of hundreds of scientists publishing thousands of papers and of the individual scientist making his advances based on the work that had been previously published. There were, of course, a number of men whose outstanding contributions changed the direction of the researches, but space does not permit referring to them by name.

Occurrence and abundance. The rare-earth elements are not rare in nature. They are found in low concentrations widely distributed throughout the earth's crust and in high concentrations in a considerable number of minerals. In addition, they are also found in many meteorites, on the moon, and in the sun. The spectra of many types of stars indicate that the rare-earth elements are much more abundant in these systems than they are in our solar system. Even promethium-147, which has a half-life (time required for one-half the material to undergo radioactive decay) of only a few years, has been observed in certain stars.

Cerium is reported to be more abundant in the earth's crust than tin, and yttrium and neodymium more abundant than lead. Even the relatively scarce lutetium is said to be more abundant than mercury or iodine.

Widespread distribution on earth

The rare-earth elements are found as mixtures in almost all massive rock formations, in concentrations of from ten to a few hundred parts per million (ppm) by weight. The fact that these elements have not been separated into minerals containing individual members of the family at any time in the earth's history--even after eons of repeated melting and resolidifying, mountain formation and erosion, exposure to hot vapour, and immersion in seawater—attests to the great similarity in properties of these elements. Nevertheless, rock formations resulting from some of these geological processes become enriched or depleted in rare earths at one end of the series or the other, so that an analysis of the relative content of the rare-earth elements is never exactly the same, even for similar rocks taken from different locations. In general, it has been found that the more basic (or alkaline) rocks contain smaller amounts of rare earths than do the more acid rocks, and it is believed that as these molten basic rocks intrude into the more acidic rocks, the rare earths are partially extracted into the more acidic rocks. Also, as this extraction takes place, the rare-earth elements of lower molecular weight (lanthanum, cerium, praseodymium, and neodymium) are taken up to a greater extent than the heavier elements.

New analytical methods involving activation' analysis (production of artificial radioactivity) and mass spectroscopy (separation of atoms on the basis of mass) have made it possible to make accurate measurements of the relative abundances of these elements, even when they are present in extremely small amounts. Such measurements are of great interest to geophysicists because they supply valuable information about the development of geological formations. The cooling of molten rocks and superheated water solutions that have percolated through rock under great pressure frequently produces minerals containing up to 50 percent rare earths. (For uniformity, these percentages are calculated as if the entire rare-earth content of the mineral were present in the form of oxides.) From the presence and composition of such minerals, geochemists can learn a great deal about the conditions, such as temperature and pressure, to which the rock mass was subjected. Similarly, the relative abundance of rare earths in the rocks on the moon is of great interest because of what it is expected to reveal about how the moon was formed and whether all or part of the moon was molten at any time.

The average content of rare-earth elements found in certain meteorites (chondrites) and in three types of common rocks is listed in Table 1. Included also is an estimate of the relative abundance of the elements in terms of the overall rank of all known elements and of their concentration in the earth's crust. It is now generally accepted that the relative values of the rare-earth elements in chondritic (granular) meteorites represent their overall relative abundance in the cosmos. The elements

Table 1: Abundance of the Naturally Occurring Rare-Earth Elements (parts per million)

element	earth's crust		average of 20 chondritic meteorites	composite of 40 North American shales	western North American Pre-cambrian granites	Kilauea basalt
	rank*	abundance				
Sc	46	5.0	—	—	—	—
Y	31	28.0	1.800	35.00	31.00	—
La	35	18.0	0.300	39.00	49.00	10.50
Ce	29	46.0	0.840	76.00	97.00	35.00
Pr	45	5.5	0.120	10.30	11.00	3.90
Nd	32	24.0	0.580	37.00	42.00	17.80
Sm	42	6.5	0.210	7.00	7.20	4.20
Eu	57	1.1	0.074	2.00	1.25	1.31
Gd	43	6.4	0.320	6.10	5.80	4.70
Tb	59	0.9	0.049	1.30	0.94	0.66
Dy	50	4.5	0.310	—	—	3.00
Ho	56	1.2	0.073	1.40	1.22	0.64
Er	54	2.5	0.210	4.00	3.20	1.69
Tm	65	0.2	0.023	0.58	0.53	0.21
Yb	53	2.7	0.170	3.40	3.50	1.11
Lu	60	0.8	0.031	0.60	0.52	0.20

*Expressed in a range of 1 to 105.

with even atomic numbers are much more abundant than the odd-numbered elements. Such information, together with the relative abundance of their isotopes, is of critical importance to astrophysicists because it bears on theories of the origin of the universe and the genesis of the chemical elements. (See CHEMICAL ELEMENTS, ORIGIN OF.)

This article is divided into the following sections:

I. Comparative chemistry of rare-earth elements

ELECTRONIC STRUCTURE

The existence of the rare-earth elements—a family of elements more alike in their properties than any other group of elements—was one of the puzzles the solution of which led to present-day understanding of atomic structure. It is now well-known that atoms are made up of positively charged protons, neutral particles called neutrons, and negatively charged electrons. The protons and neutrons are located in a minute nucleus at the centre of an atom, and the number of protons in the nucleus is equal to the atomic number of the element. In its neutral state, an atom has as many electrons surrounding the nucleus as there are protons in it. These electrons occupy a series of concentric shells: each successive shell contains a certain number of electrons and is designated by a quantum number, n, which can be any number from 1 to 7 for stable atoms. Electrons in shells with smaller n's usually are closer to the nucleus and are bound much more tightly to it than those in shells with larger n's. For this reason, the shells with low *n*'s fill first as one progresses from elements at the beginning of the periodic table to those with higher atomic number. When a new outer shell starts to fill with electrons, a new horizontal row, or period, of the periodic table begins and, because the outermost electrons in the lanthanide elements have an $n = 6$, these elements would appear in the sixth row of the periodic table if they were not generally moved to a special location (see PERIODIC I AW).

Electron shells and subshells in atoms

Within each shell, electrons occur in certain subshells, which are known (for reasons that are no longer important) as the *s*, p, *d*, and *f* subshells. Each subshell has its own quantum number *l*, which represents the angular momentum possessed by the electrons of that subshell. As extra protons and electrons are added to the atoms one at a time (in proceeding through the periodic table), the electrons fill the various shells and subshells in order of their stability. Because of the shapes of the subshells, it happens that *s* and *p* subshells of higher n are more stable than d and *f* subshells of lower n. Therefore, as electrons are added to atomic structure, they often enter shells of higher *n* before the lower shells are completely filled.

As the fifth period of the periodic table ends (with the element xenon), the 4*s*, 4*p*, and 4*d* subshells are filled, as are the 5*s* and 5*p* subshells. Then, with the beginning of the sixth period, two 6*s* electrons are added for the first two elements of the period, and a third electron goes into a 5*d* orbital as the element lanthanum is reached. Following lanthanum, however, as the next electron is added, it goes into the 4*f* subshell, which up to this point has remained vacant. This shell is capable of holding 14 electrons, and the addition of the next 13 electrons corresponds to the filling of this inner shell. The elements in which this occurs are the 14 lanthanide elements—from cerium through lutetium. These elements are much alike because the differences in their electronic structures chiefly involve the inner 4f electrons, whereas it is the outer *s* and *p* (and sometimes *d*) electrons that are involved in chemical bonding with other atoms and thereby determine the chemical behaviour of the elements. Although lanthanum atoms contain no 4*f* electrons, they resemble the atoms of the lanthanide elements closely, and it is not surprising that lanthanum should behave much as the lanthanides do (the name lanthanide, in fact, merely means lanthanum-like). Scandium and yttrium are elements in the same vertical file in the periodic table as lanthanum, and their atoms, too, have somewhat the same electronic structure but fewer filled shells, the outermost electrons in scandium being two 4*s* electrons and one 3d. In the case of yttrium, however, the outermost electrons are 5*s* and 4*d* electrons, respectively.

Location of lanthanide elements in the periodic table

Because of their general similarity in atomic structure, scandium, yttrium, lanthanum, and the 14 lanthanides are very similar chemically. This similarity is the reason they are found together in nature and also the reason they are so frequently classed together as the rare-earth elements. Interestingly enough, the element directly below lanthanum in the periodic table is the radioactive element actinium, which has properties that much resemble those of lanthanum. Actinium is followed by a series of elements known as the actinides, which correspond to the filling of the 5*f* shell, just as the lanthanides correspond to the filling of the 4*f* subshell. There is much resemblance between the two groups: for further coverage of the actinides, see the articles ACTINIDE ELEMENTS AND THEIR COMPOUNDS and TRANSURANIUM ELEMENTS.

Table 2 lists the lowest energy electronic configurations of the rare-earth elements in gaseous and in condensed

Table 2: **Some** Properties **of** the Rare-Earth Elements (values recommended by Ames Laboratory)

	Scandium	Yttrium	Lanthanum	Cerium	Praseodymium	Neodymium	Promethium	Samarium
Atomic number	21	39	57	58	59	60	61	62
Atomic weight	44.956	88.905	138.91	140.12	140.907	144.24	(145)*	150.35
Colour of element	silvery	silvery	silvery	silvery	silvery	silvery	silvery	silvery
Melting point (°C)†	1,541	1,522	921	799	931	1,021	1,168	1,077
Boiling point (°C)¶	2,831	3,338	3,457	3,426	3,512	3,068	2,700	1,791
Density at 25" C (g/cm³)	2.9890	4.4689	6.1453	6.672	6.773	7.007	—	7.520
Conduction electrons in metal	3	3	3	3 (3.1)	3	3	3	3
Valence in aqueous solution	**3**	**3**	**3**	**3,4**	**3**	**3**	**3**	**3,2**
Colour in aqueous solution	colourless	colourless	colourless	colourless	green	rose	colourless	yellow
Colour of oxide	white	white	white	off-white (Ce_2O_3); cream (CeO_2)	black (Pr_6O_{11}); green (Pr_2O_3)	blue		cream
Electronic configuration	$[Ar]\ 3d^{1}4s^{2}$	$[Kr]\ 4d^{1}5s^{2}$	$[Xe]\ 5d^{1}6s^{2}$	$[Xe]\ 4f^{2}6s^{2}$	$[Xe]\ 4f^{3}6s^{2}$	$[Xe]\ 4f^{4}6s^{2}$	$[Xe]\ 4f^{5}6s^{2}$	$[Xe]\ 4f^{6}6s^{2}$
Heat of fusion (kcal/g-atm)	3,369	2,732	1,482	1,238	1,652	1,705	—	2,061
Heat of vaporization7 (kcal/g-atm)	89.9	101.3	103.1	101.1	85.3	78.5	—	49.2
Electrical resistivity at 25° (ohm-cm X 10^{-6})	52	59	61–80	70–80	68	65	—	91
Nbel point (°K)	none	none	none	12.5	—	20	—	15
Curie point (OK)	none	none	none	—	—	—	—	—
Compressibility at 25° C (cm^2/kg X 10^{-6})	2.26	2.68	4.04	4.10	3.21	3.00	(2.8)	3.34
Crystal structure:	hcp	hcp	d-hcp	fcc	d-hcp	d-hcp	—	rhomb
Radius, metallic (Å)	1.640	1.801	1.879	1.820	1.828	1.821	—	1.804
Radius, ionic (Å)	0.732 (M^{3+})	0.893 (M^{3+})	1.061 (M^{3+})	1.034 (M^{4+}, 0.92)	1.013 (M^{4+}, 0.90)	0.995 (M^{3+})	0.979 (M^{3+})	0.964 (M^{2+}, 1.11)

	Europium	Gadolinium	Terbium	Dysprosium	Holmium	Erbium	Thulium	Ytterbium	Lutetium
Atomic number	63	64	65	66	67	68	69	70	71
Atomic weight	151.96	157.25	158.9254	162.50	164.930	167.26	168.934	173.04	174.97
Colour of element	silvery	silvery	silvery	silvery	silvery	silvery	silvery	silvery	silvery
Melting point (°C)†	822	1,313	1,356	1,412	1,474	1,529	1,545	819	1,663
Boiling point (°C)¶	1,597	3,266	3,123	2,562	2,695	2,863	1,947	1,194	3.395
Density at 25" C (g/cm³)	5.2434	7.9004	8.2294	8.5500	8.7947	9.066	9.3208	6.9654	9.8404
Conduction electrons in metal	2	3	3	3	3	3	3	2	3
Valence in aqueous solution	**3,2**	3	3	3	3	3	3,2?	**3,2**	3
Colour in aqueous solution	colourless	colourless	colourless	yellow tint	yellow	pink	greenish tint	colourless	colourless
Colour of oxide	white, greenish tint	white	brown (Tb_4O_7); white (Tb_2O_3)	white	yellowish white	pink	white, greenish tint	white	white
Electronic configuration	$[Xe]\ 4f^{7}6s^{2}$	$[Xe]\ 4f^{7}5d^{1}6s^{2}$	$[Xe]\ 4f^{9}6s^{2}$	$[Xe]\ 4f^{10}6s^{2}$	$[Xe]\ 4f^{11}6s^{2}$	$[Xe]\ 4f^{12}6s^{2}$	$[Xe]\ 4f^{13}6s^{2}$	$[Xe]\ 4f^{14}6s^{2}$	$[Xe]\ 4f^{14}5d^{1}6s^{2}$
Heat of fusion (cal/g-atm)	2,204	2,403	2,583	2.577	4,033	4,757	4,025	1,830	(4,457)
Heat of vaporization7 (kcal/g-atm)	(41.9)	95.3	93.4	70.0	72.3	76.1	55.8	36.5	102.2
Electrical resistivity at 25° (ohm-cm X 10^{-6})	91	127	114	100	88	71	74	28	60
Nbel point (°K)	90	—	229	179	132	85	57	none	none
Curie point (°K)	—	293.2	221	87	20	20	. . .	none	none
Compressibility at 25° C (cm^2/kg X 10^{-6})	8.29	2.56	2.45	2.55	2.47	2.39	2.47	7.39	2.38
Crystal structure:.	bcc	hcp	hcp	hcp	hcp	hcp	hcp	fcc§	hcp
Radius, metallic (A)	1.984	1.801	1.783	1.774	1.766	1.757	1.746	1.939	1.735
Radius, ionic (Å)	0.950 (M^{2+}, 1.09)	0.938 (M^{3+})	0.923 (M^{4+}, 0.84)	0.908 (M^{3+})	0.894 (M^{3+})	0.881 (M^{3+})	0.870 (M^{3+})	0.858 (M^{2+}, 0.93)	0.850 (M^{3+})

'Radioactive isotope separated from fission residues. †Corrected for new temperature scale. ‡hcp = hexagonal close packed; fcc = face-centred cubic; rhomb = rhombohedral; bcc = body-centred cubic. §Structure normally found at 25° C; other forms known. ¶R. Hullgren, R. Orr, and K. Kelley.

states, *i.e.*, as the free metal or in a crystalline or solution form as a compound. In a compound the $5d6s^2$ electrons (the superscript indicating the number of electrons in the subshell) are the valence, or bonding, electrons. These electrons are involved in the chemical bond and are usually paired with the electrons of the anion (the negative ion included with the rare-earth ion in the compound), with the result that they are no longer closely associated with the rare-earth atom. In the case of the metal, these electrons are free to wander throughout the crystal, being able to carry an electrical current and known, therefore, as conducting electrons. In the case of lanthanides in the gaseous state, the valence electrons are localized at the individual atoms, but in many cases the 5d electron can switch to the 4f subshell, giving rise to a state of different energy and different electronic configuration.

Lanthanide contraction

As the charge on the nucleus increases across the lanthanide series, it pulls the various subshells, especially the 5s and 5*p* subshells, closer to the nucleus. As a result, the radii of the lanthanide ions decrease as the atomic number increases. This effect is known as the lanthanide contraction.

PROPERTIES OF THE RARE-EARTH ELEMENTS AS A GROUP

Chemical **properties.** ***Rare-earth*** *componds.* The chlorides, bromides, iodides, nitrates, acetates, chlorates, and perchlorates of rare-earth elements are extremely soluble in water. The sulfates, $M_2(SO_4)_3 \cdot 8H_2O$, are sparingly soluble and have the property of being more soluble at 0° *C* (32" F) than at room temperature. The hydroxides, fluorides, carbonates, phosphates, and oxalates are insoluble in neutral or alkaline solutions.

Rare-earth ions form hundreds of different complexes with organic molecules. The compounds of these complexes can be soluble, sparingly soluble, or insoluble in aqueous solutions, but most of them are soluble in various organic solvents. More than 75 such families of lanthanide complexes have been studied.

The anhydrous salts (those free of water) are extremely stable, and, although they dissolve readily in water, they are rarely precipitated from solution; they must, therefore, be prepared by other methods, of which there are many. An anhydrous rare-earth chloride, MCl_3, for example, can be prepared by passing dry hydrogen chloride gas with slow heating over the hexahydrate ($MCl_3 \cdot 6H_2O$) or by adding chlorine gas (Cl_2) directly to the metal. Many anhydrous rare-earth compounds are dissolved by certain mixtures of molten salts; many of the rare-earth ions can exist in such solvents in more than one oxidation state (the number of electrons gained or lost by the atom). Compounds such as neodymium(II) chloride (in which the roman numeral indicates the oxidation state), $NdCl_2$, can be formed from neodymium metal and molten neodymium(III) chloride, $NdCl_3$.

Binary compounds with nonmetals

Rare-earth metals combine directly with nonmetallic elements to form very stable borides, carbides, nitrides, oxides, chlorides, bromides, phosphides, and sulfides. Most of these compounds can also be prepared by other methods, and many of them exist in several forms.

The common oxide is the trivalent form or sesquioxide (M_2O_3), and, in proceeding across the lanthanide series, the oxide of this formula is found in three different crystalline forms. These oxides all have high melting points, well above 2,000" C (3,600" F). They also form mixed crystals with other oxides such as alumina (Al_2O_3) and iron(III) oxide (Fe_2O_3). When the sesquioxides of cerium, praseodymium, and terbium are heated in air, they are converted to other oxides, with formulas CeO_2, Pr_6O_{11}, and Tb_4O_7, respectively. When heated in oxygen under pressure, the latter two are also converted to the dioxides PrO_2 and TbO_2. When the sesquioxide of any rare-earth element is heated with a small amount of the corresponding metal, it forms a black nonstoichiometric oxide formulated M_2O_{3-x} (x represents a small fraction) having a crystal structure like that of the sesquioxide but with vacancies at certain points in the crystal lattice. Rare-earth nitrides, MN, also have high melting points; nitrogen-deficient nitrides, MN, $_x$, often crystallize, so preparation of the pure compounds of ideal composition is difficult.

Monocarbides and dicarbides, MC and MC_2, exist for all of the rare-earth elements. A number of nonstoichiometric carbides exist for most of the rare earths. These compounds possess the same crystal structure over an appreciable composition range. In general, all the carbides melt in the range of 2,000" C or higher. Two types of borides, with formulas MB_4 and MB_6, are known for all the lanthanides. The hexaborides have properties that make them useful in electronic applications. Other borides also have been prepared for several of the rare-earth elements.

Solutions. Many rare-earth salts are soluble; the solutions are strong electrolytes (carriers of electric current), and when they are dilute, the salts are ionized completely into positively charged trivalent hydrated rare-earth ions and negatively charged anions. As the solutions become more concentrated, the rare-earth ions frequently form complexes (or associations) with the anion; these take the form either of a so-called outer complex, in which the rare-earth ion is surrounded by a layer of water molecules, with the anion outside these, or of an inner complex, in which the anion has moved next to the rare-earth ion, thereby displacing one or more water molecules.

Studies of rare-earth solutions are useful for understanding the behaviour of concentrated electrolytes, the theory of which is not so well understood as that of dilute electrolytes. The high charge on rare-earth ions enhances the effects of charged particles in solution, and the properties of the solutions change slightly and in a predictable manner for adjacent lanthanides. On the other hand, if the values of certain properties of the solutions are plotted against the atomic numbers of the lanthanides, simple curves are rarely obtained, because the size of the coordination sphere (the sphere encompassing the atoms or ions adjacent to the rare-earth ion) is changed by the lanthanide contraction in traversing the series. The coordination number can vary from 6 to 12, depending upon the size of the lanthanide ion, the nature of the anion or other complexed group (including water), and the solvent.

Formation of hydrates

The lanthanide salts that crystallize from saturated solution are usually hydrated (that is, contain molecules of water), and, in general, as one proceeds across the series, the number of water molecules in the crystal changes. Lanthanum, cerium, and praseodymium chloride, for example, crystallize at room temperature with seven molecules of water for each ion of the rare-earth metal, so that the formula of the hydrate is $MCl_3 \cdot 7H_2O$, in which M represents the ion of the rare-earth metal and Cl stands for a chloride ion. Other chlorides of the series crystallize as hexahydrates, $MCl_3 \cdot 6H_2O$. At different temperatures, pressures, and acidities, other hydrates can form; *e.g.*, praseodymium chloride hexahydrate ($PrCl_3$. $6H_2O$) forms above 60° C (140" F). If other metallic ions are present, substances of variable composition may result, in which either metallic ion can occupy a cation (positive ion) site in the crystal, or they can form orderly stoichiometric double salts; *e.g.*, a mixed nitrate of neodymium and magnesium [$2Nd(NO_3)_3 \cdot 3Mg(NO_3)_2 \cdot 24H_2O$].

With increasing alkalinity (basicity), rare-earth ions in solution hydrolyze (or react with water) as in the following equation:

$$M^{3+} + nH_2O \rightleftharpoons M(OH)_n^{(3-n)+} + nH^+.$$

From these solutions basic salts, such as basic nitrates or oxychlorides, can precipitate.

Physical **properties.** The pure rare-earth metals are bright and silvery. A bar of europium will tarnish almost immediately when exposed to air and will be entirely converted to the powdered oxide in a few days. Lanthanum, cerium, praseodymium, and neodymium also corrode readily in air; bars of these metals become encrusted with a thick layer of oxide in several weeks. Metallic yttrium, gadolinium, and lutetium, on the other hand, remain bright and shiny for years.

The properties of the rare-earth metals are frequently quite sensitive to the presence of impurities; for example, the light lanthanide metals will corrode much more rapidly if small amounts of calcium or magnesium or rare-earth oxides are present in the metal. The melting points and transition temperatures between different crystal forms (allotropic forms) can be changed drastically, frequently by several hundred degrees, when the metals are alloyed with other elements.

Small amounts of nonmetallic impurities also affect many of the properties of the rare-earth elements. Several thousand parts per million by weight of oxygen and even smaller amounts of nitrogen in the metals make them brittle. The effect of nonmetallic impurities on physical properties is determined by atomic percentages (that is, by the relative numbers of atoms present) and not by weight percentages; thus, in lutetium 300 parts per million by weight of hydrogen is about 5 percent on an atomic basis, whereas 1,000 parts per million of oxygen in lutetium by weight represents only 1 percent oxygen on the atomic scale.

In determining properties of the rare-earth metals it is obviously essential to work with well-characterized samples. The amount of each individual impurity present should be accurately known, as well as the previous history of the sample with regard to temperature and work. If the reported values of physical properties are to have much meaning, this information concerning the samples used must be given. Unfortunately, because of a lack of appreciation of the importance of impurities or a lack of proper equipment to adequately characterize the sample, there is a wide variation in the numbers reported in the literature for certain properties. In Table 2 are the values of the properties that—in the author's opinion—are the best values known. Some properties, such as elastic constants, resistivity, and effective magnetic moments, are very sensitive to temperature and show marked anomalies at and in the neighbourhood of crystal or magnetic transformations. Also, some properties depend on the angle at which they are measured with respect to the principal crystal axes in the metal.

In fact, the rare-earth metals do not resemble one another as closely as was generally believed in the early part of the 20th century. Physical properties differ as much across the lanthanide series as they do for most other series in the periodic table. The melting point of lutetium, for example, is almost twice that of lanthanum, and the vapour pressures of ytterbium and europium at 1,000° C (1,832°F) are millions of times greater than those of lanthanum and cerium. Lanthanum is a superconductor of electricity at 6" K (—267" C, —449" F), and gadolinium is a stronger ferromagnet at 0° than iron. The properties of adjacent pairs of lanthanide elements do, however, differ in a predictable and usually regular manner. This behaviour makes these metals uniquely valuable in studying theories of the metallic state, the formation of alloys, and the existence of intermetallic compounds. Each of the rare-earth metals readily combines with almost any other metallic element, and the resulting alloys exhibit a wide variety of properties: they can be hard or soft, brittle or ductile, and they can have high or low melting points. Some are extremely pyrophoric (ignite spontaneously), whereas others cause a coating to be formed on the surface of metals such as magnesium that protects the alloys from corrosion at elevated temperatures.

Rare-earth metals absorb hydrogen to form stable alloy-like hydrides in which percentages of the divalent compounds MH_2 range from zero to 100. These hydrides, brittle and metallic in appearance, have a bluish tinge. After absorption of hydrogen to yield the composition MH_2, further absorption occurs, finally yielding trivalent hydrides MH_3. During the change from MH_2 to MH_3, the properties become more saltlike. The amount of hydrogen per unit volume in yttrium hydride is considerably greater than that in water or liquid hydrogen, and this hydride does not develop a partial pressure of hydrogen gas equal to one atmosphere until the alloy has been heated to a white heat. Cerium metal, once the oxide surface film has been broken, absorbs hydrogen at room temperature and decomposes water vapour at higher temperatures, absorbing the hydrogen and reacting with the oxygen to form a layer of sesquioxide on the surface. The oxides, nitrides, and carbides of the rare-earth elements are soluble in the molten metals: as are the elements oxygen, nitrogen, and carbon. The exact form in which the dissolved substances are present is not known, but it is generally believed that the nonmetallic elements are present as interstitial atoms (atoms inserted in spaces left in the crystal structure). These dissolved nonmetallic elements remain in solid solution over a considerable composition range at temperatures near the melting point. As the metal is slowly cooled, however, the solubility decreases, and the dissolved elements precipitate as a second phase, probably as the sesquioxides, nonstoichiometric nitrides, and carbides. The diffusion rate (rate of movement) for nonmetallic elements in the metal is low below 800" C (1,472" F) and becomes progressively lower as the temperature is lowered. The properties of the metals containing these impurities, therefore, are dependent upon the heat treatment to which they have been subjected.

Heat treatment and crystal structure

Anisotropy. Rare-earth metals often exhibit anisotropy—differences in properties depending on which direction in the crystal they are measured—and the heat treatment of the sample is important in producing polycrystalline metal. It is possible by certain heat treatments to produce large grains oriented preferentially in a given direction. When properties that depend on crystal direction are measured on such polycrystalline samples, the results have little meaning unless the amount of preferred orientation is known.

Spectra of gaseous lanthanides

Energy states of lanthanides. The spark and arc spectra (patterns of emitted light) of the gaseous lanthanides are extremely complicated. There are literally tens of thousands of frequencies of light emitted by each of the lanthanides, and it requires very powerful instruments (spectrographs) to resolve them. This complexity arises from the fact that the lanthanides have an incomplete inner subshell, and the angular moments (spins and orbital motions) of the electrons in this subshell can combine in many ways to give many different energy states. In the most complicated case, that of gadolinium, there are 3,432 different states. Any of these states can combine with the many states arising from the three valence electrons, and this condition results in an incredible number of excited energy levels. The emitted frequencies represent transitions between any of these states. The situation is further complicated by the fact that the ionization potentials of these elements are extremely low, with the result that in the arc- and spark-light sources there are a great many ionized atoms, and their complicated spectra fall on top of those of the neutral atoms. Many thousands of these lines remain to be identified. A start has been made on this task, however, and a few levels—including the basic ones—have been identified. A thorough understanding of the energy levels of the rare-earth atoms will be of great value in arriving at a complete theory by which all the properties of an atom can be calculated from basic principles. Furthermore, the complete identification of the spectral lines of the rare-earth elements will be of great assistance to astronomers in identifying the many lines observed in stellar spectra that are believed to indicate the presence of rare-earth elements.

The first ionization potentials (the energy required to remove an electron from the neutral atom) of these elements are also difficult to determine accurately because of the complexity of the rare-earth spectra.

Rare-earth ions in solids and liquids. The sharp bands in spectra of solid rare-earth elements and compounds are much better understood. These bands arise from transitions between different energy states of the 4f subshell, and the position of the bands seems to be little affected by the atoms surrounding the lanthanide atoms. For this reason, scientists have been able to use these bands for more than a century to determine whether particular rare-earth ions are present in solids or liquids. The fine line structure of the bands, which can be resolved at low temperatures, is sensitive to the environment, and this effect makes such spectra a valuable tool for studying the forces that exist in solids and liquids.

Magnetic properties

The 4f electrons also are responsible for the strong magnetism exhibited by the metals and compounds of the lanthanides. In the incomplete 4f subshell the magnetic effects of the different electrons do not cancel out each other as they do in a completed subshell, and this factor gives rise to the interesting magnetic behaviour of these elements. At higher temperatures, all the lanthanides except lutetium are paramagnetic (weakly magnetic), and this paramagnetism frequently shows a strong anisotropy. As the temperature is lowered, many of the metals exhibit a point below which they become antiferromagnetic (*i.e.*, magnetic moments of the ions are aligned but some are opposed to others), and, as the temperatures are lowered still further, many of them go through a series of spin rearrangements, which may or may not be in conformity with the regular crystal lattice. Finally, at still lower temperatures, a number of these elements become ferromagnetic (*i.e.*, strongly magnetic, like iron). Some of the metals have saturation moments (magnetism observed when all the magnetic moments of the ions are aligned) greater than iron, cobalt, or nickel. They also show a strong anisotropy in their magnetic behaviour depending on the crystal direction. Study of the magnetism of rare-earth elements has had great influence on present-day theories of magnetism.

Electrical conductivity. The rare-earth metals, with the exceptions of cerium, ytterbium, and europium, have three electrons available for carrying electrical current. The space occupied by these electrons apparently represents more than 85 percent of the volume associated with the atom of each metal. Cerium is reported to have an average of 3.1 conducting electrons, presumably as the result of the existence of some of its atoms in a state in which four electrons are free to move through the metal. Pure cerium under high pressure or at low temperature assumes a high-density form in which the four-electron state assumes more importance. Europium and ytterbium are much less dense than the other lanthanides, and they have only two conducting electrons; the third valence electron has moved to an inside subshell (4*f*). In europium this electron half fills this subshell, and in ytterbium it completes it, the two configurations $4f^7$ and $4f^{14}$ being particularly stable. The electrical and chemical properties of these two metals therefore resemble those of magnesium, calcium, strontium, and barium (metals with two conducting electrons) more closely than those of the other lanthanides.

Nuclear properties. As a group the rare-earth elements are rich in total numbers of isotopes. averaging about 20 each. The elements with odd atomic numbers have only one, or at most two, stable isotopes, but those with even atomic numbers have from four to seven stable isotopes. Some of the unstable isotopes are feebly radioactive, having extremely long half-lives. The unstable radioactive isotopes can be produced in many ways; *e.g.*, by fission, neutron bombardment, radioactive decay of neighbouring elements, and bombardment of neighbouring elements with charged particles. The lanthanide isotopes are of particular interest to nuclear scientists because they offer a rich field for testing theories about the nucleus, especially because many of these nuclei are nonspherical, a property that has a decided influence on nuclear stability. When either the protons or neutrons complete a nuclear shell (that is, arrive at certain fixed values), the nucleus is exceptionally stable—the number of protons or neutrons required to complete a shell being called a magic number. One particular magic number—82 for neutrons—occurs in the lanthanide series.

II. Production and processing of rare-earth metals

SOURCES AND EXTRACTION

Though numerous minerals rich in rare earths are found in the earth's crust, many are extremely rare, and many more are found only in small pockets in more massive rocks. Although such minerals are of considerable research interest they are not used commercially. Monazite, a mixed phosphate of calcium, thorium, cerium, and various lanthanides, occurs in extensive deposits and is one of the main sources used commercially to obtain the light rare-earth elements. Monazite contains about 50 percent by weight rare-earth elements, in the approximate proportions 50 percent cerium, 20 percent lanthanum, 20 percent neodymium, 5 percent praseodymium, and lesser amounts of samarium, gadolinium, and yttrium. It also contains small amounts of the heavy rare-earth elements. The actual amounts of each element in the mineral vary considerably, depending on the point of origin of the monazite, because the various metallic elements can substitute for one another in the crystal lattice. The mineral probably formed as small crystals in rocks as they cooled, but as the mountains eroded away and were washed into the sea, the monazite, being denser than most other materials, settled first, while the lighter materials were carried farther out to sea. Apparently as a result of this action, sandbars containing monazite are found along the coasts of Brazil and southwest India. Concentrated deposits are also found on certain uplands, which are thought to have been the beaches of ancient seas or oceans and which were later uplifted. Such deposits in massive amounts are found in Australia, the U.S.S.R., and South Africa, and in the United States in South Carolina, Florida, and Idaho, and in many other locations. The mineral is dredged or scooped up, pulverized if necessary, and concentrated by flotation methods. Sometimes a magnetic-belt separator is used to pull the more magnetic monazite to one side in order to separate it from the nonmagnetic materials. The monazite is then shipped to rare-earth chemical plants. The mineral xenotime, a phosphate of yttrium and various lanthanides, is frequently found associated with monazite and may comprise from 1 to 10 percent of the mixed minerals. It is similar to monazite except that the metallic atoms are about 50 to 60 percent yttrium, and it contains more heavy lanthanides than light ones. Xenotime is one of the main sources of the heavy rare earths, and it can be separated from monazite by the magnetic-belt process because it is more magnetic than monazite.

Monazite

Another important source of light rare earths and europium is the mineral bastnaesite, a fluorocarbonate of lanthanum and cerium, with smaller amounts of neodymium and praseodymium. It is found in extensive deposits in eastern California. It contains almost no heavy rare earths, but there is enough europium (about 0.1 percent) to supply much of the world demand for this element. The mineral is also found in the U.S.S.R. and in parts of Africa. The rock is broken up by blasting and then is crushed and ground to a fine powder. The bastnaesite is separated from the other materials by the usual flotation methods and is then treated chemically so that it can be separated into europium, lanthanum, and cerium fractions by liquid–liquid extraction methods (see below *Liquid–liquid* extraction).

The niobium titanate minerals, such as fergusonite, euxenite, samarskite, and blomstrandine, are rich in the heavy rare-earth elements but are not used much commercially. The same is true of such silicates as gadolinite and allanite. Other commercial sources of rare-earth oxides are certain uranium- and apatite-mining operations in which the rare earths are obtained as a by-product even though the rare-earth content of the ores is low.

Very little scandium is found in rare-earth minerals. Most of the scandium produced commercially is a by-product from uranium processing—the scandium, which may be present in amounts up to five parts per million, being recovered from the uranium solution. There is, however, a rare mineral thortveitite—found in Norway—that contains up to 34 percent scandia, Sc_2O_3.

METHODS OF SEPARATION AND PURIFICATION

Generally, the rare-earth elements exist in dilute solution as trivalent ions. Quite early, however, it was found that a number of the elements could also exist in tetravalent or divalent form—including cerium(IV), samarium(II), europium(II), and ytterbium(II). If an element could be oxidized (to the tetravalent state) or reduced (to the divalent), then it could be removed readily from the other rare earths. Between the years 1930 and 1935, for example, about two kilograms (4.5 pounds) of extremely pure europium compounds were prepared by a separation process making use of the divalent form of europium. Although europium is one of the less abundant rare-earth elements, it was one of the first of the heavier rare earths to become generally available.

Purification of europium

Fractionation. Because the ions of the rare-earth elements are surrounded by tightly bound water molecules in aqueous solution, compounds of the rare earths formed from aqueous solutions have properties much alike and this similarity is particularly true for adjacent elements. The problem is still further complicated by the fact that one rare-earth ion can be substituted readily for another in crystal lattices, with the result that most precipitates consist of crystals of almost the same rare-earth mixture as the solution. Because of this behaviour, chemists of the 19th and early 20th centuries found it necessary to resort to laborious fractionation processes to isolate individual rare-earth elements. At the time, many different processes were used, such as fractional crystallization, fractional precipitation, fractional decomposition. and fractional extraction. All of these consisted of separating the mixed rare earths into two approximately equal fractions, one of which would be enriched in the lighter elements and the other in the heavier elements. Both fractions would then be put back into solution and the process repeated on each of them. Usually the adjacent inner fractions would be iecombined before proceeding to the next stage. Gradually, the lighter rare earths were collected in the beakers toward one end of the system, with the heavier elements concentrated at the other end.

As the quantity of material in the end beakers became small, it was usually customary to combine equivalent fractions from other similar runs. At this point the first series would be split into several independent groups, and a new fractionation process more suited to the elements in each fraction started. Needless to say, the quantity of a relatively pure rare-earth compound obtained from the end beakers was distinctly limited.

Fractional separation methods, particularly for adjacent heavy rare earths, are extremely slow and tedious. One investigator, for example, reported that he had recrystal lized the bromate salt of a thulium fraction 15,000 times

Fractional separation methods

and could see little difference between the first and last fractions. It is now known that even the purest fractions he obtained contained some ytterbium and erbium. If the elements are far apart in the lanthanide series, however, the task is simplified. It takes only a few partial precipitations, for example, to obtain a lanthanum–cerium–praseodymium fraction completely free of erbium, thulium, ytterbium, and lutetium. The most basic (nonacidic) of the rare-earth elements, lanthanum, is very well situated in the series in this respect because it occurs at one end, and a few fractionations suffice to separate a lanthanum–cerium fraction from the other rare earths. Since cerium in its tetravalent form has distinctly different chemical properties from a typical lanthanide in the trivalent state, it can be separated from lanthanum easily by ordinary chemical operations. Consequently, pure lanthanum and cerium compounds have been commercially available for many years, and even today several companies find the fractionation process the most economical method for producing compounds of these elements in ton quantities.

Ion exchange. Ion exchange is a method of separation based on differential absorption and elution (washing off) of substances from certain solid supporting materials, often powdered or finely divided materials held in glass tubes. The technique was first used in the rare-earth field during World War II to separate fission products obtained from nuclear reactors. In December 1943 a research group at the Oak Ridge (Tennessee) national laboratory announced that they had separated the mixed rare-earth elements from certain fission products by ion exchange on an organic resin into three fractions. The first fraction was shown to have radioactivity associated with yttrium, and the final peak to have cerium activity. The middle peak was thought to be a combination of the neodymium and element-61 activities. The group at Oak Ridge continued to develop the elution technique for separating fission products both with and without carriers (nonradioactive materials added to carry with them the radioactive isotopes). By the end of the war, they had succeeded in developing the processes so that they could separate the individual rare-earth elements of the cerium group (cerium through element 61) and yttrium. The carriers usually consisted of a few milligrams of each of the corresponding natural rare earths. In the meantime, a group at Iowa State University applied the ion-exchange process to the separation of gram quantities of adjacent rare earths and succeeded in separating the difficult pair praseodymium and neodymium in fairly high purities in gram quantities.

Ion exchange of fission products

The ion-exclzange elution method. Rare-earth-element separation by the ion-exchange elution process is carried out as follows: At the start of the process, the resin is saturated with monovalent cations, such as ammonium ion, NH_4^+, or hydrogen ion, H^+. Next, a solution of mixed rare-earth ions accompanied by strong acid anions is poured onto the top of the column. When the rare-earth ion encounters the cation-containing resin, it replaces three monovalent cations, and these—along with the strong acid anions—will flow through the column in solution and out the bottom. A band of resin saturated with rare-earth ions forms at the top of the ion exchanger and grows in length as more rare-earth solution is added. There is, however, little separation of individual rare-earth ions as this band forms. An eluant solution containing an anion that complexes with the rare-earth ion is then prepared, for example, an ammonium citrate solution of controlled acidity. This solution is then started flowing through the column to elute the rare-earth band down the column and out the bottom. When the main ions present in ammonium citrate in acid solution, $HCit^{2-}$ or H_2Cit^-, encounter rare-earth ions on the resin, complex ions form; these enter the solution phase, and three monovalent ions deposit on the resin in their place. When the rare-earth complexes reach the ammonium or acid resin, in front of the rare-earth band, the rare-earth ions are again deposited, and the band progresses down the column. The formation constants of the individual rare-earth complexes increase slightly with increasing atomic number. Because the various rare-earth ions on the resin are in equilibrium with the rare-earth complexes in solution as they pass over the band, there is a slight enrichment of heavy rare earths at the front of the band. As the band progresses down the column, this enrichment continues. At the same time, the band grows in length, since ammonium and hydrogen ions are also in equilibrium with the resin and their ions will deposit along with the rare-earth ions. After the band has travelled many band lengths, each rare earth exhibits a bell-shaped elution curve (concentration of rare-earth ions versus volume of eluant leaving the column) and these individual rare-earth bands travel down the column at different rates. The bands overlap badly at first, but after travelling many band lengths, they pull completely apart. The area under each curve is, of course, constant, because the amount of each rare earth on the column does not change, but the concentration of the rare earth in the resin gets less and less relative to the ammonium and hydrogen ions on the resin.

Limitations of the elution technique

With this type of separation, the original mixed rare-earth band must be quite narrow because the band has to travel many band lengths on a given column. The ions in the eluant are constantly overrunning the bands, with the result that large quantities of solution are needed; and the solution coming out the bottom of the column containing the successive pure rare earths is extremely dilute in rare earths. Such a process is obviously ideal for separating radioactive tracers, which one can count by means of radioactivity, and this process is frequently used in analytical chemistry, where only small amounts of the rare earths are separated. When it is necessary to obtain large amounts of rare earths in high purity, this process is not effective. It has the disadvantage of requiring far more chemicals than the displacement method developed later and described below. Furthermore, this process is not particularly adaptable to being scaled up to produce large quantities of ultrapure rare earths, nor is it well suited for recycling the water and chemicals. It does not give the purity of the individual rare earth that displacement methods can achieve. Finally, the elution process is slow compared with the displacement method.

The ion-exchange displacement method. The band displacement method of separating individual rare-earth elements was first published in 1952. This process is capable of being scaled up to handle any quantity of rare earths. The mixture can be resolved so that 98 or 99 percent of each individual rare earth can be recovered with less than 0.1 percent of other rare-earth impurities; and, if the rare earths are taken from the middle third of the bands, the sum total of other rare earths can be kept as low as 0.0001 percent. The same resins and type of equipment are used in this process as in the elution technique. Two strong chemical constraints, however, are imposed at the top and bottom edges of the rare-earth band. The eluant contains a strong complexing ion—such as a chelating agent, an organic molecule that wraps itself around the rare-earth ion, replacing all or most of the adjacent water molecules. The first constraint requires that the formation constants of the rare-earth complexes formed should be large enough so that, when the chelating agent encounters the top edge of the rare-earth band, it complexes in a short distance all of the rare-earth ions, moving them into solution and replacing them with the cation of the eluant. (The formation constant, however, should not be so large as to remove all the rare-earth ions from the solution phase.) The second chemical constraint occurs at the bottom edge of the rare-earth band: the original resin bed, called the retaining bed, down which the rare-earth band is moving, must have cations on its exchange sites that form a much tighter soluble complex with the chelating ion than do the rare-earth ions. Under this constraint the rare-earth complex promptly breaks up at the point where it encounters the retaining bed, and the rare-earth ions completely deposit in the bed, simultaneously removing an equivalent amount of the retaining-bed cations. With these constraints, the rare-earth band, after spreading out slightly to reach equilibrium, remains of constant length, with sharp top and bottom edges, no matter how far down the

The functions of chemical constraints

column it travels. The elution curve is flat-topped (rare-earth concentration remains constant over almost the entire band when plotted against volume of elute leaving the bottom of the column), and the percentage of rare-earth ions in the rare-earth band on the active sites of the resin is close to 100 percent. Here again, there must be a slight difference in the formation constants of the rare-earth chelates, so that the rare-earth ions are constantly interchanging as the eluant flows by the rare-earth band. As the band moves, the individual rare earths separate into individual flat-topped bands, which ride head to tail and never pull apart. By the time the band has travelled a tenth of its length, most of the heavy rare earths are already to be found in the front segments of the total rare-earth band, and, by the time it has travelled one length, all the individual rare earths are in separate bands, which overlap only slightly to give a narrow region consisting of a binary mixture of adjacent elements. These mixed regions, of course, must be recycled. By having a series of columns, however, the original rare earth band can be made very long, and, since the overlap regions are independent of band length, the bulk of each successive rare earth comes out the bottom of the column in high purity.

A number of companies have adopted the displacement process and, using it, have made available highly pure salts of the rare-earth elements of atomic number 59 and above (all the elements from praseodymium through lutetium) in any quantity at reasonable prices. This process has the distinct advantage of allowing the water and the eluting chemicals to be recycled and used over again. One long absorbed band can follow another down a series of columns if a short retaining-bed section is continually regenerated between the absorbed bands.

The first chelating agent used was ammonium citrate at such a low acidity that the citrate ion, Cit^{3-}, predominates in the solution. At this acidity the complex chelate ion, MCit_2^{3-}, forms. This process works well, but in 1954 it was improved by using a buffered ammonium solution of ethylenediaminetetraacetic acid (EDTA) with a cupric-ion retaining bed. A number of other chelating agents and types of retaining beds have also been investigated. Many of these work well, but none is markedly superior to EDTA. Some of these other systems, however, are used for certain regions of the rare-earth series because, for these regions, the processing is accomplished more quickly and cheaply.

Organic
and
aqueous
phases

Liquid–liquid extraction. Liquid–liquid extraction methods also find important applications in the rare-earth industry. The basic principles involved are similar to those operating in the ion-exchange processes. An organic solvent, such as tributyl phosphate, flows countercurrent to an aqueous stream containing the mixed rare-earth salts. Rare-earth complexes are formed with formation constants that vary somewhat across the series. The rare-earth ions can complex with their own anions to form neutral molecules that are soluble in the organic phase, or they can complex with molecules of the organic solvent and thereby join the organic stream. If desired, an organic chelating agent can be added to form complexes with the rare-earth ions. These complexes should be soluble in the organic liquid. As the aqueous phase flows past the immiscible organic stream, an equilibrium is set up between the rare-earth ions in the aqueous solution and the complex ions in the organic solvent. As the two streams flow past each other, the heavy rare-earth elements concentrate in one stream and the lighter ones in the other.

The equilibrium constant for the exchange of one rare-earth ion for another is usually small, with the result that the ions have to exchange with the complex many times before a clear-cut separation between two rare-earth ions is achieved. This process necessitates that the two liquids be in contact with each other through many stages. If the equilibrium constant is equal to 1, no separation will take place, and for adjacent rare earths it is difficult to find complexes—except in special cases—that differ much from that value.

The liquid–liquid extraction process suffers from the disadvantage that for a given system only one cut is made in the rare-earth series, and if 15 pure rare earths were desired 14 cuts would have to be made. It also suffers from the disadvantage that the distance the rare-earth ions must travel in order to complete one exchange is many times that required in the ion-exchange columns. It has the advantages, however, that more concentrated solutions can be used and that the process is more economical for handling large quantities of materials. So far, it has found application mainly in special cases. It is used in some industries to concentrate the total rare earths where their abundance in the original materials is low. It is also used for separating certain elements, such as lanthanum, cerium, europium, and yttrium, with which favourable equilibrium constants are found. This is the case for cerium and europium, because they can be extracted in their tetravalent and divalent forms, respectively. Yttrium is not a lanthanide, and its position in the rare-earth series can be changed by using different organic solvents or complexing agents. First, a complex is used that separates the yttrium and heavy lanthanides from the light lanthanides, and then a different system is used whereby the yttrium is shifted with respect to the lanthanide series, so that it can be separated from the heavy lanthanides.

The liquid–liquid extraction system has not been successful in separating the adjacent heavy rare earths in the quantities desired. If ultrahigh-purity rare earths are required, it is common practice--even in those cases where liquid–liquid extraction methods have been used—to place the somewhat impure rare earth on an ion-exchange column and to use the displacement method for further purification.

PREPARATION OF PURE METALS

Early metal-reduction methods. It is relatively easy to reduce anhydrous halides of the rare earths to metals. What is difficult, however, is to reduce them to high-purity metals in ingot form. The rare-earth metals have a great affinity for the nonmetallic elements—hydrogen, boron, carbon, nitrogen, oxygen, silicon, sulfur, phosphorus, chlorine, and bromine—and form very stable compounds with them. If a small amount of rare-earth metal ıs added to most other metals containing these elements present as impurities, it reacts with the impurities and removes them by gathering them together in nodules or transferring them to the slag phase. There has been a steady market for misch metal, a mixed rare-earth alloy, since Auer von Welsbach's time. A small addition of this alloy greatly improves the mechanical properties of many impure metals or alloys.

Also, hot rare-earth turnings (chips or curls from machining) can be used to produce extremely pure helium, neon, and argon by removing hydrogen, oxygen, nitrogen, carbon dioxide, and hydrocarbon vapours. As is often the case with the rare earths, however, other—and cheaper—materials perform this function equally well, and for this reason the rare-earth elements are seldom used for this purpose.

Finally, molten rare-earth metals dissolve almost all other metals and react with most compounds. They come close to being the hypothetical universal solvent of the ancients. The molten metal attacks any crucible in which it is melted, and the final product generally is a rare-earth-rich alloy of the crucible elements.

First
reduction
to a
metal

*The **metallothermic process.*** Mosander, in 1826, was the first to reduce a rare earth to a metal. He used a metallothermic reaction (heating with active metals) to reduce anhydrous chlorides made from his ceria with metallic sodium or potassium. His yields were low, 26 percent, and the metal existed as small nuggets in a solid slag, from which they could be separated only with difficulty. The metal was very impure; it contained considerable amounts of sodium or potassium and iron and other crucible materials. It also contained considerable amounts of hydrogen, oxygen, nitrogen, and carbon, as well as-a mixture of the ceria group of rare earths.

During the next hundred years, as the individual rare earths were discovered and separated, a number of scientists reduced many of the lighter rare earths to the metal-lit form using the metallothermic process—but sometimes varying it by substituting calcium, magnesium, and aluminum as the reductants and anhydrous fluoride salts as the reactants. Because of the scarcity of pure rare earths, however, as well as the difficulty in finding suitable crucible materials and the poor equipment for keeping out atmospheric gases, the metals were still so impure that no extensive studies could be made of their properties.

In 1935, samples of the purest rare-earth chlorides available were reduced to metals at relatively low temperatures in glass capsules with potassium vapour. This process gave free metals in the form of fine powder imbedded in potassium chloride; no attempt was made to separate the metal from the potassium chloride, because only such properties as crystal structure, density, and magnetic susceptibility were under investigation. Potassium chloride acted as an internal standard in the X-ray investigations, and magnetic susceptibilities could be corrected for the potassium chloride present. Although these metals were not really pure by modern standards—they contained appreciable amounts of potassium and rare-earth impurities—they yielded values for the lattice constants and densities of most of the rare-earth metals that lie within 1 percent of the best modern values.

Development of electrolytic techniques

Electrolytic processes. In 1875, the first successful preparation of rare-earth metals by an electrolytic process was reported. About five grams each of cerium, lanthanum, and didymium (neodymium and praseodymium) in compact form were prepared by electrolyzing the fused chlorides covered with layers of ammonium chloride. The electrolytic technique was later improved, and, in the period 1902–05, misch metal, cerium, lanthanum, praseodymium, neodymium, and samarium were prepared. In 1906, Auer von Welsbach started the commercial production of lighter flints, for which the misch metal was electrolytically reduced. In the years 1923–26, several improvements in the cell designs were made, and somewhat purer samples of lanthanum, cerium, and neodymium were prepared, along with some yttrium, although most of the latter metal deposited as powder.

The electrolytic process suffers from much the same difficulties as the metallothermic methods. It is difficult to find electrodes and cell materials that will stand up to molten rare earths and at the same time not introduce impurities into the ingot. It is also difficult to design cells that exclude all the atmospheric elements. The method works best for the low-melting rare earths, with which the cells can be kept sufficiently hot so that a molten pool of the metal forms in the bottom of the cell. With the higher melting rare earths, only powdered metal is formed, and it is difficult to separate it from the electrolyte in a pure form. By the early 1970s there had not been much success in finding electrolytes that can be heated above 1,000" C.

In 1931, a cell especially designed for producing much purer metals and also capable of reducing the halides of the heavier rare earths was employed to produce a quantity of cerium that contained only a small percentage of impurities and, somewhat later, the same apparatus was used to produce a number of other rare-earth metals, including europium, gadolinium, and yttrium.

By 1939 most of the rare-earth metals had been made in fair purity, and a number of their properties, such as magnetic behaviour, melting point, density, crystal structure, and chemical reactivity had been studied. All of these metals contained small amounts of metallic impurities and unknown amounts of nonmetallic impurities. Most of these impurities were not reported because analytical methods to determine them had not been developed at that time. Almost no work had been done on the properties of the rare-earth alloys except for those of cerium and lanthanum.

Importance of impurities

Modern techniques for producing ultrapure rare-earth metals. As purer rare-earth metals are produced, it is increasingly clear that many of their properties are extremely sensitive to small amounts of impurities. This phenomenon is particularly true with regard to magnetic and to nonmetallic impurities. For many industrial uses extreme purity is not required—nor even desired—since less pure metals can be produced much more cheaply. On the other hand, the presence of impurities can be critical in metal produced for research purposes, especially when experimental properties are being compared with predicted values, or in metal to be incorporated into solid-state devices. The processes described below are those used in making research-grade metal. If less pure metal is satisfactory, many of the steps described can be omitted, and the process can be terminated at the point where the desired purity is attained.

One especially favoured reduction process utilizes metallic calcium (Ca) and the rare-earth fluoride (MF_3). The reaction is as in the following equation:

$$3Ca + 2MF_3 \rightleftharpoons 2M + 3CaF_2.$$

Other metallothermic processes, however, can also be used, such as lithium (Li) metal and the rare-earth chloride (MCl_3):

$$3Li + MCl_3 \rightleftharpoons M + 3LiCl.$$

Variations of these methods, using lithium, sodium, potassium, or calcium as the reducing agent and any halide of the rare earth for the reactant, also are possible. For these alternative processes to succeed, however, the metal must be separable from the slag cleanly without introducing impurities, and all sources of contamination must have been eliminated.

The problem of obtaining sufficient quantities of highly pure individual rare-earth oxides has been solved by the development of the displacement-band method of separating rare earths on ion-exchange columns described above. If the oxides are obtained from the middle third of the pure rare-earth band, the total of other rare-earth impurities in the metals obtained from them does not exceed 10 parts per million. Fractions taken closer to the band edges contain somewhat larger proportions of such impurities. If the same equipment is used to prepare the different raw materials and to make the metals of a number of different rare earths, great care must be taken to prevent cross contamination of the rare earths.

Tungsten and tantalum crucibles

Limiting contamination. Contamination from the crucible cannot be eliminated entirely. Tungsten and tantalum make the best crucibles: they are little attacked by molten rare-earth metals at temperatures below 1,000" *C*, and the crucible material introduced into the rare-earth metals at higher temperatures, if harmful, can be removed by special techniques. Both tungsten and tantalum are available commercially in the form of both crucibles and thin sheets: to prepare them for use, these materials are thoroughly cleaned and baked in a high vacuum to remove impurities that may be adsorbed on their surfaces.

Introduction of impurities from the atmosphere can be largely eliminated by carrying out all operations in an environment of purified helium and by the use of modern high-vacuum ion pumps instead of oil pumps.

Finally, if ultrapure metals are to be obtained, the raw materials from which they are made must also be ultrapure or their impurities will end up in the rare-earth ingot. Commercial calcium is doubly distilled at low pressure in an atmosphere of pure argon or helium to remove iron and various nonmetals that it contains, and thereafter is rigorously protected from carbon dioxide and water vapour. Anhydrous rare-earth halides form oxyhalides upon contact with water vapour; therefore, the preparation of an anhydrous fluoride is carried out in two steps. The first is the passage of dry hydrogen fluoride over the powdered oxide, immediately sweeping away any water formed; and the second is the passage of the pure, dry hydrogen fluoride over the molten fluorides. If this is done, the oxygen content of the fluoride can be kept below ten parts per million.

Groups for standardized operation

Standardizing reduction methods. There are enough differences in the properties of the 17 rare-earth metals that the same reduction process does not work equally well for all of them. If metal containing less than 0.01 weight percent of impurities is desired, each element has to be treated somewhat differently. Many of the operations are the same for all reductions, and if the metals are divided into five groups, standardized operations can be applied for all metals in a group. The groups are as follows: Group I consists of those metals that have low melting points and high boiling points—lanthanum, cerium, praseodymium, and neodymium. Group II consists of those metals having high melting points and high boiling points—gadolinium, terbium, scandium, yttrium, and lutetium. Group III consists of those metals having high melting points, low boiling points, and, in addition. an appreciable vapour pressure at the melting point—dysprosium, holmium, erbium, and scandium. Group IV consists of those metals that have low boiling points—samarium, europium, ytterbium, and thulium. Group V, consisting only of promethium, would be included in Group II, except that serious difficulties result from the intense radioactivity of the metal. All operations with it must be carried out by remote controls, and this is usually done at special installations.

The calcium-reduction process works well for the metals of Groups I, II, III, and V, but not at all well for Group IV. Most of the trifluorides (SmF_3, EuF_3, YbF_3) of this group are reduced only to divalent fluorides, even when a large excess of calcium is used; the resulting material resembles, but is not, the metal. Thulium trifluoride is reduced to the metal, but the high vapour pressure of molten metallic thulium causes difficulty.

The standard procedure for producing the metal by calcium reduction is to load a tantalum container with enough rare-earth fluoride to yield a metal billet weighing about 300 grams (11 ounces). About 10 percent excess calcium is added to drive the reaction to completion. The crucible is then placed in a furnace with a helium atmosphere and heated above the melting point of the rare-earth metal or of the slag—whichever is greater—and held at that temperature until the reaction is complete and the metallic and slag layers have separated because of differences in their densities. After cooling to room temperature, the crucible is taken out of the furnace in the dry box, cut in two at the metal–slag interface, and all slag is knocked off the metal. Usually a bright metal surface can be obtained. The metal ingot, however, contains small amounts of calcium and rare-earth fluoride as impurities. Group I and Group II metals are then put in another tantalum crucible and replaced in the furnace for the boiling-off process. This time a high vacuum is used, and the metal is heated to about 1,400" to 1,500" C (2,552" to 2,732" F) and held there for some time, so that any volatile impurities, particularly calcium and rare-earth fluoride, evaporate. At these temperatures considerable amounts (1 to 3 percent) of tantalum dissolve in the molten metal. The furnace temperature is then slowly lowered until it is just above the melting point of the pure metal, at which temperature it is held for a few minutes to allow most of the tantalum to precipitate onto the walls or sink to the bottom of the crucible. (For Group I metals, the solubility of tantalum is about 50 parts per million or less at the melting point.) The ingot is then removed from the cooled furnace in the dry box, and the tantalum crucible and precipitates are machined off. The resulting ingot usually contains less than 0.01 percent total impurities.

Special problems with groups II and III

Removal of dissolved tantalum. The Group II metals, because of their higher melting points, still contain some tantalum as an impurity when they solidify; usually this tantalum appears as a second phase, showing up as black dots in the metallographic pictures of the metal. It is possible, however, to purify these metals further by distillation from a tantalum still. The still consists of a short tantalum crucible located in a high-vacuum furnace. Affixed to this crucible is an inverted crucible, which is out of the heating zone of the furnace, its upper part being 400" to 500" C (720" to 900" F) cooler because of radiation losses. The rare-earth metal can then be slowly sublimed (changed from a solid into a gas without passing through the liquid state) and resolidified in the inverted crucible. Because volatile impurities usually have different boiling points and heats of sublimation, it is possible (by choosing the right temperature) to sublime the metal in such a manner that the impurities can be separated from the metal. The nonvolatile tantalum remains behind in the still. Finally, the condenser, with its rather porous crystalline mass of rare-earth metal, is removed from the furnace, and the tantalum (or tungsten) is machined off in the dry box. The porous mass is then arc-melted into a billet on a water-cooled copper hearth under an inert atmosphere.

The Group III elements cannot be held at their melting points for long: because of their volatility a considerable quantity of the metal is lost. The boiling process therefore is omitted, but sublimation or volatilization works well, and—by the right choice of temperature in the still—both the volatile and inert impurities can be eliminated.

For the Group IV metals, only a distillation process is used. The pure oxide of the rare earth is dissolved in acid and reprecipitated using ultrapure chemicals to remove traces of calcium and magnesium often introduced from the water and chemicals used in the ion-exchange process. The precipitate is again converted to the oxide and placed in the still. Pure metallic lanthanum, cerium, or misch metal, which has been subjected to the boiling-off process to remove volatile impurities, is added in excess.

The reaction of europium oxide with lanthanum metal ($Eu_2O_3 + 2La \rightarrow La_2O_3 + 2Eu\uparrow$) takes place when the mixture is heated. Because the vapour pressure of europium is millions of times greater than that of lanthanum or of the oxide of either element, the metal distills away from the oxides, and the reaction goes practically to completion. If ultrahigh-purity metal is desired, a second distillation is performed.

Problems with thulium

Thulium poses special problems. Its melting point is so high that the molten metal acquires a considerable amount of impurity from the crucible. On the other hand, it has such a high vapour pressure at the melting point that it is practically impossible to melt it without losing much of the metal. Thulium is never melted, therefore, but is sublimed to the condenser, on which it forms solid crystals but not compact metal. If a solid bar is desired, the porous metal can be pressed into a tantalum tube and reduced to about half its diameter. The tantalum covering can then be machined off and a bar of compact metal obtained.

INDUSTRIAL USES

The properties of the 17 rare-earth elements in the form of their metals, alloys, or compounds—or some combination thereof—are so varied as to make them valuable for many industrial uses. Many other somewhat less costly materials, however, often will perform just as well; and when this is the case, the rare-earth elements are seldom used for these purposes. Only when their properties are unique is the extra cost justified industrially. Nevertheless, in 1969, the world shipments of light rare-earth oxides to be used in industry were reported to be 30,000,000 pounds (U.S.S.R. figures were not included but were substantial). About 20,000,000 pounds of oxides were obtained from bastnaesite, and the remaining 10,000,000 pounds were derived from monazite. About two-thirds of this production was consumed in the United States.

Use as catalysts

In 1969, about 12,000,000 pounds of rare earths were used in the United States to produce catalysts for the cracking of crude petroleum. The natural mixture of rare earths obtained from the minerals accounted for about 20 percent of that total, and the remaining 80 percent was made up of special mixtures of lanthanum, praseodymium, neodymium, and samarium. Rare-earth catalysts have been repeatedly recommended for use in numerous organic reactions, including the hydrogenation of ketones to form secondary alcohols, the hydrogenation of olefins to form alkanes, the dehydrogenation of alcohols and butanes, and the formation of polyesters. The extent to which these catalysts are used in industry seldom is made public, but there is no doubt that the rare earths show marked catalytic properties.

Another substantial use of rare-earth oxides is in the glass industry. Cerium oxide has been found to be a more rapid polishing agent for glass than rouge, and several million pounds a year are consumed in the polishing of lenses for cameras, binoculars, and eyeglasses, as well as in polishing mirrors and television faceplates. Glasses containing lanthanum oxide have very high refractive indexes and low dispersions. Such glasses are used in complex lenses for cameras, binoculars, and military instruments — for the purpose of correcting spherical and chromatic aberrations. Rare-earth oxides often are added to glass melts in order to produce special glasses. Neodymium is added to some glasses to counteract the yellowish tint caused by iron impurities. Very pure neodymium oxide, when added in sufficient quantities (1–5 percent), gives a beautiful purple glass. Praseodymium and neodymium are added to glass to make welders' and glassblowers' goggles, that absorb the bright-yellow light from the sodium flame. The same combination is sometimes added to the glass used in television faceplates to decrease the glare from outside light sources. A beautiful yellow ceramic stain results from the addition of about 3 percent praseodymium oxide to zirconium oxide. Cerium oxide increases the opacity of white porcelain enamels.

Use in metallurgy

The metallurgical industry is another heavy user of rare earths. Small amounts of misch metal and cerium have long been added to other metals or alloys to remove their nonmetallic impurities. Misch metal added to cast iron makes a more malleable nodular iron. Added to some steels, it makes them less brittle. The addition of misch metal to certain alloys has been reported to increase the tensile strength and improve the hot workability and the high-temperature oxidation resistance. The rare earths are particularly effective in iron–chromium and iron–chromium–nickel alloys to improve a number of their properties, especially their resistance to corrosion and oxidation. Yttrium metal is said to work even better than misch metal in removing impurities from certain materials. The flints of cigarette lighters are an alloy of misch metal and iron.

The addition of misch metal or pure rare-earth elements to magnesium increases its high-temperature strength and its creep resistance — that is, resistance to slow deformation under prolonged use. This alloy also makes better castings if small amounts of zirconium or other metals are added, and such alloys are used in jet-engine and precision castings. The addition of small amounts of rare-earth elements to aluminum has also been reported to give better castings.

Use in television screens

By far the heaviest user of ultrapure separated rare earths is the television industry. It has been found that if a small amount of europium oxide (Eu_2O_3) is added to yttrium oxide (Y_2O_3), it gives a brilliant-red phosphor. Colour television screens utilize red, green, and blue phosphors. In the past, a zinc–cadmium sulfide was used as the red phosphor, but it was not completely satisfactory because its fluorescent band was too wide, and it could not be made to fluoresce as intensely as the other phosphors. The Y_2O_3–Eu_2O_3 phosphor corrected these disadvantages and made possible much brighter and more natural coloured pictures. This use has been growing in many countries. Many of the early rare-earth screens used europium–yttrium orthovanadate phosphor, but the industry is shifting heavily toward the oxide phosphor. Some television companies have substituted gadolinium oxide for the yttrium oxide, and the television industry consumed in excess of 10,000 pounds of ultrapure gadolinium oxide in 1969. In the same year, Japan imported over 200 tons of xenotime in order to prepare rare earths for its television industry. The rare-earth phosphors are also finding use in mercury-arc lights, which are used for sporting events and special street lighting. Instead of the unhealthy-looking blue light of the mercury arc, the phosphors give an intense white radiation similar to daylight. Considerable amounts of mixed rare-earth fluorides are used to make cored carbon rods, which are used as arcs in searchlights and in some of the lights used by the motion-picture industry.

Yttrium-iron garnets are synthetic high-melting silicates that can be fabricated into special shapes for use as microwave filters in the communications industry. Yttrium-aluminum garnets also are being produced at an increasing rate for use both in electronics and as gemstones. Both of these synthetic minerals have much use in the jewelry business. These garnets have a high refractive index and a hardness approaching that of diamond. In a solid crystal form they are amazingly transparent, and they are being cut into imitation diamonds.

Another rapidly developing industrial application of rare earths is in the manufacture of strong permanent magnets. Alloys of cobalt with rare earths, such as cobalt–samarium, produce permanent magnets that are far superior to most of the varieties now on the market. Another recent development is the use of a barium phosphate-europium phosphor in a sensitive X-ray film that forms satisfactory images with only half the exposure.

Nuclear applications

Europium, gadolinium, and dysprosium have large capture cross sections for thermal neutrons — that is, they absorb large numbers of neutrons per unit of area exposed. These elements, therefore, are incorporated into control rods used to regulate the operation of nuclear reactors or to shut them down should they get out of control. In addition, rare-earth elements are used as burnable neutron absorbers to keep the reactivity of the reactor more nearly constant. As uranium undergoes fission, it produces some fission products that absorb neutrons and tend to slow down the nuclear reaction. If the right amounts of rare-earth elements are present, they bum out at about the same rate that other absorbers are formed.

Yttrium dihydride is used as a moderator in reactors to slow down neutrons. Certain rare earths are also used in shielding materials because of their high nuclear cross sections. Scandium metal is used as a neutron filter that allows neutrons only of a certain energy (two kiloelectron volts) to pass through.

Complexes of europium, praseodymium, or ytterbium with derivatives of camphor are useful reagents for analysis of optically-active organic compounds, which often are obtained as mixtures containing unknown proportions of two components that differ only in that their molecular structures are mirror images of each other (see ISOMERISM). Determination of these proportions can be very difficult, but the rare-earth complexes provide asymmetric environments in which each component absorbs electromagnetic radiation of a particular frequency in the presence of a strong magnetic field (a phenomenon covered in the article MAGNETIC RESONANCE). Composition of the mixtures then can be determined by measuring the intensities of the separate absorptions.

The rare earths have low toxicities and can be handled safely with ordinary care. Solutions injected into the peritoneum will cause hyperglycemia (excess of sugar in the blood), decreased blood pressure, spleen degeneration, and fatty liver. If solutions are injected into muscle about 75 percent of the rare-earth element remains at the site, the remainder going to the liver and skeleton. When taken orally, only a small percentage of a rare-earth element is absorbed into the body. Organically complexed ions are somewhat more toxic than solids or inorganic solutions. As is true for most chemicals, dust and vapours should not be inhaled, nor should they be ingested. Solutions splashed into the eyes should be washed out, and splinters of metal should be removed.

When handling rare-earth ores or minerals, dust should be avoided because many minerals contain other toxic elements, such as beryllium, thorium, and uranium. Finely divided rare-earth metals can ignite spontaneously, somewhat as magnesium does.

BIBLIOGRAPHY. J.W. MELLOR, *A Comprehensive Treatise on Inorganic and Theoretical Chemistry*, vol. 5 (1924), an early history of the rare earths, from 1790 to 1920; F. TROMBE *et al.*, *Éléments des terres rares*, vol. 1–2 (1960), reviews of rare-earth chemistry to 1958 (in French); F.H. SPEDDING and A.H. DAANE (eds.), *The Rare Earths* (1961), a broad review of rare-earth chemistry and metallurgy to 1960; K.A. GSCHNEIDER, JR., *Rare Earth Alloys* (1961), a review to 1960; C.M. LEDERER, J.M. HOLLANDER, and I. PERLMAN, *Table of Isotopes*, 6th ed. (1967), includes a review of radioactivity of the rare-earth elements; R.J. ELLIOTT (ed.), *Magnetic Properties of the Rare Earth Elements* (1971); E. WHEELWRIGHT (ed.), *Promethium* (in press), a volume devoted to this single man-made rare-earth element; L.H. AHRENS, *Origin and Distribution of the Elements,* vol. 30 of the International Series of Monographs in Earth Science (1968), a good reference source for the abundance of elements; L. ERYING (ed.), *Progress in the Science and Technology of the Rare Earths,* 3 vol. (1964–68), current papers and reviews given at annual rare-earth conferences.

(F.H.S.)

Rashi

Rashi's stature

Rashi (acronym for Rabbi Shlomo Yitzhaqi), medieval French commentator on the Bible and Talmud, the authoritative rabbinical compendium of law, lore, and commentary, holds a unique position in Jewish religious scholarship. In an exegesis-oriented literature, in which commentaries and supercommentaries constitute the most extensive as well as creative literary genre, Rashi is "the commentator" par excellence. The study of his classic commentaries by both schoolchildren and adults, the simple and the sophisticated, still serves among Jews as the most substantive introduction to biblical and post-biblical Judaism.

Shlomo (Solomon) Yitzhaqi (son of Isaac) was born at Troyes, France, in 1040. He studied in the schools of Worms and Mayence, the old Rhenish centres of Jewish learning, where he absorbed the methods, teachings, and traditions associated with Rabbi Gershom ben Judah (c. 960–1028/1040), called the "Light of the Exile" because of his pre-eminence as the first great scholar of north European Judaism. As Rashi transferred his scholarly legacy to the valley of the Seine (*c.*1065), his task was to consolidate and expand, preserve and innovate.

Personal and environmental factors combined to insure that Rashi would not be an ivory-tower scholar but very much involved in the relatively fast-moving social, economic, political, and theological developments of the 11th century. Troyes was a major commercial centre of the province of Champagne. Rashi, the de facto but unofficial head of the small Jewish community (about 100–200 people) in Troyes, was himself a winegrower. The major developments of the period are recorded or reflected in his commentaries, in the codes, liturgical works, and miscellaneous collections edited by his disciples, as well as in the hundreds of his responsa (Hebrew *teshuvot*)—official, authoritative answers to a wide range of questions addressed to him by various people on different occasions. Those texts reflect the intensified Jewish–Christian relations in business and commerce and the concomitant re-evaluation of the Jewish attitude toward Christianity; the church–state struggle known as the investiture controversy; emerging patterns of Jewish self-government (defined in special charters) in relation to both the existing feudal system and incipient urban civilization; the beginning of the Crusades, preached for the first time by Pope Urban II at Clermont (1096); the increased precariousness of the dispersed Jewish settlements; religious persecution and forced baptisms; and the hopes of messianism and the realities of martyrdom. Rashi also composed some penitential hymns (selihot), which revolve around twin themes: the harsh reality of exile and the comforting belief in redemption.

Rashi's method of Biblical interpretation

His Bible commentary illustrates vividly the coexistence and, to some extent, the successful reconciliation of the two basic methods of interpretation: the literal and the nonliteral. Rashi seeks the literal meaning, deftly using rules of grammar and syntax and carefully analyzing both text and context, but does not hesitate to mount Midrashic explanations, utilizing allegory, parable, and symbolism, upon the underlying literal interpretation. As a result, some of his successors are critical of his searching literalism and deviation from traditional Midrashic exegesis, while others find his excessive fondness for nonliteral homilies uncongenial. Yet it is precisely the versatility and mixture, the blend of creative eclecticism and originality, that account for the genius, the animation, and the unrivalled popularity of his commentary, which, symbolically, was the first book printed in Hebrew (1475). The commentary had a significant influence on Christian Bible study from the 12th-century Victorines to the Franciscan scholar Nicholas of Lyra (*c.*1270–1349), who, in turn, was a major source of Luther's Bible work, through the translators responsible for the King James Version of the Bible. Its influence continues in contemporary exegesis and revised translations. Rashi's custom to use a vernacular gloss to clarify the exact meaning of an obscure or technical term—there are over 3,000 of them in his works--also makes his commentary an important source for the study of Old French.

Rashi's commentary on the Talmud, based on the collective achievements of the previous generations of Franco-German scholars, reflects its genesis in the oral classroom instruction that Rashi gave in Troyes for several decades. The commentary, sometimes referred to as kuntros (literally, "notebook"), resembles a living tutor; it explains the text in its entirety, guides the student in methodological and substantive matters, resolves linguistic difficulties, and indicates the normative conclusions of the discussion. Unlike Maimonides' commentary on the Mishna (the authoritative compendium of Jewish Oral Law), which may be read independently of the underlying text whose relevant portions have been quoted or paraphrased in the commentary, Rashi's commentary is interwoven with the underlying text. Indeed, text and commentary form a unified mosaic. Rashi's work was literally epochal, and the agreement of subsequent scholars that the basic needs of text commentary had been fulfilled stimulated the rise of a new school of writers known as *tosafists*, who composed tosafot (glosses), refining, criticizing, expanding, or qualifying Rashi's interpretations and conclusions. The leaders of this great intellectual movement, which spread across all the cities of France and Germany, were Rashi's grandchildren—Samuel ben Meir (Rashbam), Jacob ben Meir Tam (Rabbenu Tam), and Isaac ben Meir (Ri ha-Zaken). Skillfully and honestly combining stricture and supplement, they were able to perpetuate and augment the achievement of the great Rashi, who died at Troyes on July 13, 1105.

BIBLIOGRAPHY. The best biography is the Hebrew study by A.M. LIFSCHITZ, *Rashi* (1966). The standard English biography remains M. LIBER, *Rashi* (1906). More specialized works include: I. AGUS, "Rashi and His School," in *The World History of the Jewish People,* vol. 2, pp. 210–248 (1966), the most recent survey of 11th-century Jewish history in the Franco-German sphere, with a good summary of intellectual and literary developments; HERMAN HAILPERIN, *Rashi and the Christian Scholars* (1963), a meritorious study, written for the non-specialist, of Rashi's influence on Nicholas of Lyra and Christian Bible study; and the *Rashi Anniversary Volume* (1941), a collection of scholarly essays on various aspects of Rashi's life and works.

(I.T.)

Rationalism

Rationalism is the philosophical view that regards reason as the chief source and test of knowledge. Holding that reality itself has an inherently logical structure, the Rationalist asserts that a class of truths exists that the intellect can grasp directly. There are, according to the Rationalists, certain rational principles—especially in logic and mathematics, and even in ethics and metaphysics—that are so fundamental that to deny them is to fall into contradiction. The Rationalist's confidence in reason and proof tends, therefore, to detract from other ways of knowing. Rationalism has long been the rival of Empiricism, the doctrine that all knowledge comes from, and must be tested by, sense experience. As against this doctrine, Rationalism holds reason to be a faculty that can lay hold of truths beyond the reach of sense perception, both in certainty and generality. In stressing the existence of a "natural light," Rationalism has also been the rival of systems claiming esoteric knowledge, whether from mystical experience, revelation, or intuition, and has been opposed to various irrationalisms that tend to stress the biological, the emotional or volitional, the unconscious, or the existential at the expense of the rational.

TYPES AND EXPRESSIONS OF RATIONALISM

Rationalism has somewhat different meanings in different fields, depending upon the kind of theory to which it is opposed.

Nativism in psychology and linguistics

In the psychology of perception, for example, Rationalism is in a sense opposed to the genetic psychology of the Swiss scholar Jean Piaget, who, exploring the development of thought and behaviour in the infant, argued that the categories of the mind develop only through the infant's experience in concourse with the world. Similarly, Rationalism is opposed to Transactionalism, a point of view in psychology according to which man's perceptual skills are achievements, accomplished through actions performed in response to an active environment. On this view, the experimental claim is made that perception is conditioned by probability judgments formed on the basis of earlier actions performed in similar situations. As a corrective to these sweeping claims, the Rationalist defends a nativism, which holds that certain perceptual and conceptual capacities are innate—as suggested in the case of depth perception by experiments with "the visual cliff," which, though platformed over with firm glass, the infant perceives as hazardous—though these native capacities may, at times, lie dormant until the appropriate conditions for their emergence arise.

In the comparative study of languages, a similar nativism was developed in the 1950s by the innovating syntactician, Noam Chomsky, who, acknowledging a debt to Descartes, exylicitly accepted the rationalistic doctrine of "innate ideas." Though the 4,000 languages spoken in the world differ greatly in sounds and symbols, they sufficiently resemble each other in syntax to suggest that there is "a schema of universal grammar" determined by "deep structures" or "innate presettings" in the human mind itself. These presettings, which have their basis in the brain, set the pattern for all experience, fix the rules for the formation of meaningful sentences, and explain why languages are readily translatable into one another. It should be added that what Rationalists have held about innate ideas is not that some ideas are full-fledged at birth but only that the grasp of certain connections and self-evident principles, when it comes, is due to inborn powers of .insight rather than to learning by experience.

Metaphysical and epistemological Rationalism

Common to all forms of speculative Rationalism is the belief that the world is a rationally ordered whole, the parts of which are linked by logical necessity and the structure of which is therefore intelligible. Thus in metaphysics it is opposed to the view that reality is a disjointed aggregate of incoherent bits and is thus opaque to reason. In particular, it is opposed to the logical atomisms of such thinkers as David Hume and Ludwig Wittgenstein, who held that facts are so disconnected that any fact might well have been different from what it is without entailing a change in any other fact. Rationalists have differed, however, with regard to the closeness and completeness with which the facts are bound together. At the lowest level, they have all believed that the law of contradiction "A and not-A cannot coexist" holds for the real world, which means that every truth is consistent with every other; at the highest level, they have held that all facts go beyond consistency to a positive coherence; *i.e.*, they are so bound up with each other that none could be different without all being different.

In the field where its claims are clearest—in epistemology, or theory of knowledge—Rationalism holds that some, at least, of man's knowledge is gained through a priori (prior to experience), or rational, insight as distinct from sense experience, which too often provides a confused and merely tentative approach. In the debate between Empiricism and Rationalism, Empiricists hold the simpler and more sweeping position, the Humean claim that all knowledge of fact stems from perception. Rationalists, on the contrary, urge that some, though not all, knowledge arises through direct apprehension by the intellect. What the intellectual faculty apprehends is objects that transcend sense experience—universals and their relations. A universal is an abstraction, a characteristic that may reappear in various instances: the number three, for example, or the triangularity that all triangles have in common. Though these cannot be seen, heard, or felt, Rationalists point out that man can plainly think about them and about their relations. This kind of knowledge, which includes the whole of logic and mathematics as well as fragmentary insights in many other fields, is, in the Rationalist view, the most important and certain knowledge that the mind can achieve. Such a priori knowledge is both necessary (*i.e.*, it cannot be conceived as otherwise) and universal, in the sense that it admits of no exceptions. In critical philosophy, epistemological Rationalism finds expression in the claim that the mind imposes its own inherent categories or forms upon incipient experience (see below *Kant and Hegel*).

Ethical and religious Rationalism

In ethics Rationalism holds the position that reason, rather than feeling, custom, or authority, is the ultimate court of appeal in judging good and bad, right and wrong. Among major thinkers, the most notable representative of rational ethics is Immanuel Kant, who held that the way to judge an act is to check its self-consistency as apprehended by the intellect: to note, first, what it is essentially, or in principle—-a lie, for example, or a theft—and then to ask if one can consistently will that the principle be made universal. Is theft, then, right? The answer must be "No," because, if theft were generally approved, no one's property would be his own as opposed to anyone else's and theft would then become meaningless; the notion, if universalized, would thus destroy itself, as reason, by itself, is sufficient to show.

In religion Rationalism commonly means that all of man's knowledge comes through the use of his natural faculties, without the aid of supernatural revelation. "Reason" is here used in a broader sense, referring to man's cognitive powers generally, as opposed to supernatural grace or faith—though it is also in sharp contrast to so-called existential approaches to truth. Reason, for the Rationalist, thus stands opposed to many of the religions of the world, including Christianity, which have held that the divine has revealed itself through inspired persons or writings and which have required, at times, that its claims be accepted as infallible, even when they do not accord with natural knowledge. Religious Rationalists hold, on the other hand, that if the clear insights of man's reason must be set aside in favour of alleged revelation, then human thought is everywhere rendered suspect—even in the reasonings of the theologians themselves. There cannot be two ultimately different ways of warranting truth, they assert; hence Rationalism urges that reason, with its standard of consistency, must be the final court of appeal. Religious Rationalism can reflect either a traditional piety, when endeavouring to display the alleged sweet reasonableness of religion, or an anti-authoritarian temper, when aiming to supplant religion with the "goddess of reason."

HISTORY OF RATIONALISM

Pre-Socratics

Epistemological Rationalism in ancient philosophies. The first Western philosopher to stress rationalist insight was Pythagoras, a shadowy figure of the 6th century BC. Noticing that, for a right triangle, a square built on its hypotenuse equals the sum of those on its sides and that the pitches of notes sounded on' a lute bear a mathematical relation to the lengths of the strings, Pythagoras held that these harmonies reflected the ultimate nature of reality. He summed up the implied metaphysical Rationalism in the words "All is number." It is probable that he had caught the Rationalist's vision, later seen by Galileo, of a world governed throughout by mathematically formulable laws.

The difficulty in this view, however, is that, working with universals and their relations, which, like the multiplication table, are timeless and changeless, it assumes a static world and ignores the particular, changing things of daily life. The difficulty was met boldly by the Rationalist Parmenides, who insisted that the world really is a static whole and that the realm of change and motion is an illusion, or even a self-contradiction. His disciple Zeno of Elea further argued that anything thought to be moving is confronted with a row of points infinite in number, all of which it must traverse; hence it can never reach its goal, nor indeed move at all. Of course, perception tells us that we do move; but Zeno, compelled to choose between perception and reason, clung to reason.

The exalting of rational insight above perception was

Plato and Aristotle

also prominent in Plato (*c.* 427–347 BC). In the *Meno,* Socrates dramatized the innateness of knowledge by calling upon an illiterate slave boy and, drawing a square in the sand, proceeding to elicit from him, step by step, the proof of a theorem in geometry of which the boy could never have heard (to double the size of a square, draw a square on the diagonal). Such knowledge, Rationalists insist, is certain, universal, and completely unlearned.

Plato so greatly admired the rigorous reasoning of geometry that he is alleged to have inscribed over the door of his Academy "Let no one unacquainted with geometry enter here." His famous "ideas" are accessible only to reason, not to sense. But how are they related to sensible things? His answers differed. Sometimes he viewed the ideas as distilling those common properties of a class in virtue of which one identifies anything as a member of it. Thus what makes anything a triangle is its having three straight sides; this is its essence. At othei times, Plato held that the idea is an ideal, a nonsensible goal to which the sensible thing approximates; the geometer's perfect triangle "never was on sea or land," though all actual triangles more or less embody it. He conceived the ideas as more real than the sensible things that are their shadows and saw that the philosopher must penetrate to these invisible essences and see with the eye of his mind how they are linked together. For Plato they formed an orderly system that was at once eternal, intelligible, and good.

Plato's successor Aristotle (384–322 BC) conceived of the work of reason in much the same way, though he did not view the ideas as independent. His chief contribution to Rationalism lay in his syllogistic logic, regarded as the chief instrument of rational explanation. Man explains particular facts by bringing them under general principles. Why does one think Socrates will die? Because he is a man, and man as such is mortal. Why should one accept the general principle itself that all men are mortal? In human experience such principles have so far held without exception. But the mind cannot finally rest in this sort of explanation. Man never wholly understands a fact or event until he can bring it under a principle that is self-evident and necessary; and he then has the clearest explanation possible. On this central thesis of Rationalism, the three great Greeks were in accord.

Nothing comparable in importance to their thought appeared in Rationalistic philosophy in the next 1800 years, though the work of Thomas Aquinas in the 13th century was an impressive attempt to blend Greek Rationalism and Christian revelation into a single harmonious system.

Rationalism in non-Western thought

Thoroughgoing Rationalism is to be found only in the philosophical tradition that has come down from Greece; the mysticism of India and the practicality of China have offered a less congenial soil for it. The nearest parallels to it in Eastern thought are found in the work of the Indian philosopher Sankara, who flourished about AD 800, and in that of the Chinese sage Chu-Hsi (1130–1200). Both were commentators on the ancient scriptures of their lands; and in ordering the scattered insights of these sources into intelligible systems, they did for their respective peoples something like what Aquinas did for the West in his harmonizing of Greek with Christian thought. Śaṅkara held, as Sir Sarvepalli Radhakrishnan has expressed it, that "the Absolute is the unattainable goal towards which the finite intellect strives." Perception is confined to what is transient and fragmentary; reason rises to truth that is timeless and universal; but even reason falls short of full understanding, which is achieved, if at all, only through mystical vision. Chu Hsi, who has influenced Chinese thought for the past six centuries, was a disciple of Confucius, though he had a stronger speculative interest than his master. He held that in all human minds a single reason was at work, which he called "the Way." All things were in some degree manifestations of it, and hence the understanding of the world lay in more complete identification with it.

Continental Rationalists

Epistemological Rationalism in modern philosophies. The first modern Rationalist was René Descartes (1596–1650), who was an original mathematician, whose ambition was to introduce into philosophy the rigour and clearness that delighted him in mathematics. He set out to doubt everything in the hope of arriving in the end at something indubitable. This he reached in his famous ***cogito ergo sum***, "1 think, therefore I am"; for to doubt one's own doubting would be absurd. Here then was a fact of absolute certainty, rendered such by the clearness and distinctness with which it presented itself to his reason. His task was to build on this as a foundation, to deduce from it a series of other propositions, each following with the same self-evidence. He hoped thus to produce a philosophical system on which men could agree as completely as they do on the geometry of Euclid. The main cause of error, he held, lay in the impulsive desire to believe before the mind is clear. The clearness and distinctness upon which he insisted was not that of perception but of conception, the clearness with which the intellect grasps an abstract idea, such as the number three, or its being greater than two.

His method was adopted in essentials by both Benedict Spinoza (1632–77) and G.W. Leibniz (1646–1716), who agreed that the framework of things could be known by a priori thinking. They differed from him, however, in their starting points. What was most undeniable to Spinoza was not the existence of his self but that of the universe, called by him Substance. From the idea of Substance, and with the aid of a few definitions and axioms, he derived his entire system, which he set forth in his *Ethics* in a formal fashion patterned after Euclid's geometry. Still, for both Spinoza and Leibniz much in nature remained stubbornly opaque. Leibniz distinguished necessary truths, those of which the opposite is impossible (as in mathematics), from contingent truths, the opposite of which is possible, such as "snow is white." But was this an ultimate distinction? At times Leibniz said boldly that if only man knew enough, he would see that every true proposition was necessarily true—that there are no contingent truths, that snow must be white.

Kant and Hegel

How, then, does reason operate and how is it possible to have knowledge that goes beyond experience? A new answer was given by Immanuel Kant (1724–1804) in his ***Critique of Pure Reason,*** which, as he said, involved a Copernican revolution in philosophy. The reason man can be certain that his logic and mathematics will remain valid for all experience is simply that their framework lies within his own mind; they are forms of arrangement imposed from within upon the raw materials of sensation. Man will always find things arranged in certain patterns because it is he who has unwittingly so arranged them. Kant held, however, that these certainties were bought at a heavy price. Just because a priori insights are the reflection of man's own mind, he cannot trust them as a reflection of the world outside himself. Whether the rational order in which man arranges his sensation—the order, for example, of time, space, and causality—represents an order holding among things-in-themselves (German ***Dinge-an-sich)*** he cannot hope to know. Kant's Rationalism was thus the counterpart of a profound Skepticism.

G.W.F. Hegel (1770–1831), the most thoroughgoing of Rationalist thinkers. attempted to break out of this Skepticism. He argued that to think of an unknowable is already to bring it within the sphere of what is known and that it is meaningless to talk of a region in which logic is invalid. Further, to raise the question "Why?" is to presume that there is an intelligible answer to it; indeed the faith of the philosopher must be that the real is the rational and the rational real, for this faith is implicit in the philosophic enterprise itself, As an attempt to understand and explain the world, philosophy is a process of placing something in a context that reveals it as necessary. But this necessity is not, as earlier Rationalists had supposed, an all-or-nothing affair issuing in a self-evident finality. Understanding is a matter of degree. What alone would wholly satisfy thought is a system that is at once all-inclusive and so ordered that its parts entail each other. Hegel believed that the universe constitutes such a whole and, as an idealist, held that it is a single, absolute mind. To the degree that the philosopher embodies and realizes this mind, his own mind will achieve both truth and reality. Indeed, the advance of civilization reflects the enlarging presence and control of such a system in the human spirit. Broadly similar Rationalistic systems were developed in England by F.H. Bradley (1846–1924) and Bernard Bosanquet (1848–1923) and in America by Josiah Royce (1855–1916).

Legalistic versus utilitarian ethics

Ethical Rationalism. The views of Kant were presented above as typical of this position (see above, *Types and expressions of Rationalism).* But few moralists have held to ethical Rationalism in this simple and sweeping form. Many have held, however, that the main rules of conduct are truths as self-evident as those of logic or mathematics. Lists of such rules were drawn up by Ralph Cudworth and Henry More among the Cambridge Platonists of the 17th century, who were noted for holding that moral principles were intrinsic to reality; and in the 18th century Samuel Clarke and Richard Price, defenders of "natural law" ethics, and the "common sense" moralist Thomas Reid also presented such lists. A 20th-century revision of this Rationalism has been offered by the Rational Intuitionists H.A. Prichard and Sir David Ross of Oxford under the name of deontology (Greek *deon*, "duty"), which respects duty more than consequences. Ross provides a list of propositions regarding fidelity to promises, reparation for injuries, and other duties, of which he says: "In our confidence that these propositions are true there is involved the same trust in our reason that is involved in our trust in mathematics." What is taken as self-evident, however, is not specific rules of conduct, but *prima facie* duties—the claims that some types of action have on men because of their nature. If a man is considering whether to repay a debt or to give the money to charity, each act has a self-evident claim on him; and their comparative strengths must be settled by a rational intuition.

The most influential variety of 20th-century ethical Rationalism has probably been the Ideal Utilitarianism of the British moralists Hastings Rashdall (1858–1924) and G.E. Moore (1873–1958). Both were teleologists (Greek *telos*, "end") inasmuch as they held that what makes an act objectively right is its results (or end) in intrinsic goods or evils. To determine what is right, reason is required in two senses: firstly, the inference to the consequences is an act of inductive reasoning; secondly, the judgment that one consequence is intrinsically better than another is a priori and self-evident. Moore thought that there is a single rule for all conduct—one should so act as to produce the greatest good—and that this is also a principle self-evident to reason.

Religious Rationalism. Stirrings of religious Rationalism were already felt in the Middle Ages regarding the Christian revelation. Thus the skeptical mind of Abelard (1079–1142) raised doubts by showing in his *Sic et Non* ("Yes and No") many contradictions among beliefs handed down as revealed truths by the church fathers. The greatest of the Medieval thinkers, Thomas Aquinas (1225–1274), was a Rationalist in the sense of believing that the larger part of revealed truth was intelligible to and demonstrable by reason, though he thought that a number of dogmas opaque to reason must be accepted on authority alone.

Expansion of religious Rationalism. Religious Rationalism did not come into its own, however, until the 16th and 17th centuries, when it took two chief forms: the scientific and the philosophic.

Galileo and Descartes

Galileo (1564–1642) was a pioneer in astronomy and the founder of modern dynamics. He conceived of nature as governed throughout by laws statable with mathematical precision; the book of nature, he said, is "written in mathematical form." This notion not only ruled out the occasional appeal to miracle; it also collided with dogmas regarding the permanent structure of the world—in particular with that which viewed the earth as the motionless centre of the universe. When Galileo's demonstration that the earth moves around the sun was confirmed by the work of Newton and others, a battle was won that marked a turning point in the history of Rationalism, since it provided a decisive victory in a crucial case of conflict between reason and apparently revealed truth.

The Rationalism of Descartes, as already shown, was the outcome of philosophic doubt rather than of scientific inquiry. The self-evidence of the *cogito,* seen by his "natural light," he made the ideal for all other knowledge. The uneasiness that the church soon felt in the face of such a test was not unfounded, for Descartes was in effect exalting the natural light into the supreme court even in the field of religion. He argued that man's guarantee against the possibility that even this natural light might deceive him lay in the goodness of the Creator. But then to prove this Creator, he had to assume the prior validity of the natural light itself. Logically, therefore, the last word lay with rational insight, not with any outside divine warrant. Descartes was inadvertently beginning a Copernican revolution in theology. Before his time, the truths regarded as most certain were those accepted from revelation; afterwards these truths were subject to the judgment of human reason, thus breaking the hold of authority on the European mind.

Deism

Four waves of religious Rationalism. The Rationalist attitude quickly spread, its advance forming several waves of general interest and influence. The first wave occurred in England in the form of Deism. Deists accepted the existence of God, but spurned supernatural revelation. The earliest member of this school, Lord Herbert of Cherbury (1583–1648), held that a just God would not reveal himself to a part of his creation only and that the true religion is thus a universal one, which achieves its knowledge of God through common reason. The Deistic philosopher John Toland (1670–1722), in his *Christianity Not Mysterious,* sought to show that "there is nothing in the Gospels contrary to reason, nor above it"; any doctrine that is really above reason would be meaningless to man. Attacking revelation, the freethinking polemicist Anthony Collins (1676–1729) maintained that the prophecies of the Old Testament failed of fulfillment; and the religious controversialist Thomas Woolston (1670–1733) urged that the New Testament miracles, as recorded, are incredible. Matthew Tindall (1657–1733), most learned of the English Deists, argued that the essential part of Christianity is its ethics, which, being clearly apparent to natural reason, leaves revelation superfluous. Thus the Deists, professing for the most part to be religious men themselves, did much to reconcile their public to the free play of ideas in religion.

Voltaire and Diderot

The second wave of religious Rationalism, less moderate in tone and consequences, was French. This wave, reflecting an engagement with the problem of natural evil, involved a decay in the natural theology of Deism such that it merged eventually with the stream that led to materialistic Atheism. Its moving spirit was Voltaire (1694–1778), who had been impressed by some of the Deists during a stay in England. Like them, he thought that a rational man would believe in God but not in supernatural inspiration. Hardly a profound philosopher, he was a brilliant journalist, clever and humorous in argument, devastating in satire, and warm in human sympathies. In his *Candide* and in many other writings, he poured irreverent ridicule on the Christian scheme of salvation as incoherent and on the church hierarchy as cruel and oppressive. In these attitudes he had the support of Diderot (1713–84), editor of the most widely-read encyclopaedia that had appeared in Europe. The Rationalism of these men and their followers, directed against both the religious and the political traditions of their time, did much to prepare the ground for the explosive French Revolution.

Hegelianism

The next wave of religious Rationalism occurred in Germany under the influence of Hegel, who held that a religious creed is a halfway house on the road to a mature philosophy, the product of a reason that is still under the sway of feeling and imagination. This idea was taken up and applied with learning and acuteness to the origins of Christianity by David Friedrich Strauss (1808–74), who published in 1835, at the age of 27, a remarkable and influential three-volume work, *Das Leben Jesu* (Eng. trans., *The Life of Jesus, Critically Examined*, 1846). Relying largely on internal inconsistencies in the Synoptic Gospels, Strauss undertook to prove these books to be unacceptable as revelation and unsatisfactory as history. He then sought to show how an imaginative people innocent of either history or science, convinced that a Messiah would appear, and deeply moved by a unique moral genius, inevitably wove myths about his birth and death, his miracles, and his divine communings.

Strauss's thought as it affected religion was continued by the philosophical historian Ernest Renan (1823–92) and as it affected philosophy by the humanist Ludwig Feuerbach (1804–72) of the Hegelian left. Renan's *Vie de Jésus* (1863; Eng. tranr., *Life* of *Jesus)* did for France what Strauss's book had done for Germany, though the two differed greatly in character. Whereas Strauss's work had been an intellectual exercise in destructive criticism, Renan's was an attempt to reconstruct the mind of Jesus as a wholly human person—a feat of imagination, performed with a disarming admiration and even reverence for its subject and with a felicity of style that gave it a large and lasting audience. Feuerbach's *Wesen des Christentums* (1841; Eng. trans, by George Eliot, *Essence of Christianity*, 1853) applied the myth theory even to belief in the existence of God, holding that "man makes God in his own image."

Evolution

The fourth wave occurred in Victorian England, following the publication in 1859 of Darwin's *Origin of Species*. This book was taken as a challenge to the authority of Scripture because there was a clear inconsistency between the Genesis account of creation and the biological account of man's slow emergence from lower forms of life. The battle raged with bitterness for several decades, but died away as the theory of evolution gained more general acceptance.

STATUS OF RATIONALISM

Religious. With increasing freedom of thought and wider acceptance of scientific views, Rationalism in religion has lost its novelty and much of its controversial excitement. To the contemporary mind, it is too obvious to warrant debate that reason and revelation cannot both qualify as sources of ultimate truth for, were they to conflict. truth itself would become self-contradictory. Hence theologians have sought accommodation through new interpretative principles that discern different grades of authenticity within the Scriptures and through new views of religious truth, existential rather than cognitive, that turn from propositional dogmas to the explication of lived human existence. Criticism of supernaturalism, however, is still carried on by such societies as the Rationalist Press Association, in Great Britain, and the Humanist Association, in the United States.

Ethical. Rationalism in ethics has suffered its share of criticism. Regarding its lists of rules—on the keeping of promises, the return of loaned goods, etc.—it has been argued, for example, that if they were specific enough to be useful (as in the rule against lying or stealing), they would tend to have exceptions—which no rule laid down by reason ought to have. On the other hand, if without exceptions, they would often prove to be tautologies: the rule of justice, for example, that we should give everyone his due would then mean only that we should give him what is justly his. After enduring a period of eclipse, however, during which noncognitive theories of ethics (emotive and existential) and relativism had preempted the field, rationalistic views, which agree in holding that moral standards do not depend upon the varying attitudes of persons or peoples, were receiving renewed attention in the mid-20th century. Prominent among these developments has been the "good-reasons" approach taken by the broadly-gauged scholar Stephen Toulmin, by Kurt Baier, and others, which examines the contexts of various moral situations and explores the kinds of justification appropriate for each.

Metaphysical. Typical of the ways of reasoning employed by Rationalists are two approaches taken to the metaphysical doctrine that all things are connected by internal relations: one a logical, the other a causal argument. An internal relation is one that could not be removed without affecting the terms themselves between which the relation holds. The argument runs: Everything is related to everything else at least by the relation "*A is different from* B." But difference is itself an internal relation, since the terms could not remain the same if it were removed. Hence everything is so connected with everything else that it could not be what it is unless they were what they are. The appeal to internal relations played an important part in the philosophies of Hegel, F.H. Bradley, and A.N. Whitehead (1861–1947).

The other line of argument is causal. Every event, it is maintained, is connected with every other, either directly or indirectly. Sir James Jeans argued that if the law of gravitation is valid, a man cannot crook his little finger without affecting the fixed stars. Here the causal relation is direct. It can also be shown that seemingly unrelated events are joined indirectly through their common connection with some remote historical event, by a chain of events leading back, for example, to Columbus' discovery of America. But if this had been different, all of its consequences would presumably have been different; thus an indirect and internal relation proves to have been present.

Many Rationalists have held with Spinoza that the causal relation is really a logical one—that a causal law, if precisely stated, would reveal a connection in which the character of the cause logically necessitates that of its effect; and if this is true, they maintain, the facts and events of the world must thus compose a single rational and intelligible order.

Rationalism and quantum physics

In the 20th century, such Rationalism met with a new and unexpected difficulty presented by quantum physics. According to the indeterminacy principle, formulated in 1927 by the German physicist Werner Heisenberg, it is impossible to discover with precision both the position and the velocity of a moving electron at the same time. This implies that definite causal laws for the behaviour of these particles can never be attained, but only statistical laws governing the behaviour of immense aggregates of them. Causality, and with it the possibility of rational understanding, seemed to be suspended in the subatomic world. Some Interpreters of the new physics, however, notably Max Planck, Albert Einstein, and Bertrand Russell, sustained the hopes of the Rationalists by insisting that what was excluded by the indeterminacy principle was not the fact of causality in this realm, but only the precise knowledge of it.

Indeed, some leaders of 20th-century science took the new developments in physics as on the whole supporting Rationalism. Protons and electrons, they contended, though beyond the reach of the senses, can still be known; and their behaviour, at least in groups, is increasingly found to conform to mathematical law. In 1932, Sir James Jeans, an astrophysicist and popularizer of science, said with a curious echo of Galileo, "the universe appears to have been designed by a pure mathematician."

Challenges to epistemological Rationalism. At first glance the claim of Empiricism that knowledge must come from sense experience seems obvious: how else could one hope to make contact with the world around him? Consequently, Rationalism has been sharply challenged—in the 19th century by the Empiricism of John Stuart Mill (1806–73) and in the 20th by that of the Logical Positivists. Mill argued that all a priori certainties are illusory: why does man believe, for example, that two straight lines cannot enclose a space? Is it because he sees it as logically necessary? No; it is because he has experienced so long and so unbroken a row of instances of it—a new one whenever he sees the corner of a table or the bordering rays of a light beam—that he has formed the habit of thinking in this way and is now unable to break it. A priori propositions, Mill claimed, are merely empirical statements of very high generality.

This theory has now been abandoned by most Empiricists themselves. Its implication that such statements as “2+2 = 4" are only probably true and may have exceptions has proved quite unconvincing. The Rationalist's rejoinder is that one cannot, no matter how hard he tries, conceive 2+2 as making 5; for its equalling 4 is necessary. But a priori knowledge is also universal. Neither of these two characteristics can be accounted for by sense experience. That a crow is black can be perceived, but not that it must be black or that crows will always be black; no run of perceptions, however long, could assure us of such truths. On the other hand, a priori truths can be seen with certainty—that if a figure, for instance, is a plane triangle within a Euclidean space, its angles must and always will equal two right angles.

Perhaps the most formidable challenge to Rationalism has come in the 20th century from such Logical Positivists as the Oxford Empiricist A.J. Ayer *(1910–*) and Rudolf Carnap (1891–1970), who had been a central figure in the Vienna Circle, where this movement first arose. Unlike Mill, they accepted a priori knowledge as certain; but they laid down a new challenge—the denial of its philosophic importance. A priori propositions, they said, are *(1)* linguistic, *(2)* conventional, and (3) analytic: *(1)* They are statements primarily of how one proposes to use words; if he says that "a straight line is the shortest line between two points," this merely reports his definition of "straight" and declares his purpose to use it only of the shortest. *(2)* Being a definition, such a statement expresses a convention to which there are alternatives; it may be defined in terms of the paths of light rays if one chooses. (3) The statement is analytic in that it merely repeats in its predicate a part or the whole of the subject term and hence tells nothing new; it is not a statement about nature but about meanings only. And since Rationalistic systems depend throughout upon statements of this kind, their importance is illusory.

To this clear challenge some leading Rationalists have replied as follows: (1) Positivists have confused real with verbal definition. A verbal definition does indeed state what a word means; but a real definition states what an object is, and the thought of a straight line is the thought of an object, not of words. (2) The Positivists have confused conventions in thought with conventions in language. One is free to vary the language in which a proposition is expressed, but not the proposition itself. Start with the concept of a straight line, and there is no alternative to accepting it as the shortest. (3) Some a priori statements are admittedly analytic, but many are not. In "whatever is coloured is extended," colour and extension are two different concepts of which the first entails the second, but is not identical with it in whole or part. Contemporary Rationalists therefore hold that the a priori has emerged victorious from the Empiricists' efforts to discredit such knowledge and the Positivists' attempts to trivialize it.

BIBLIOGRAPHY

Classics: Ancient Greek—PLATO, *Meno*; Modern—DESCARTES, *Meditationes de Prima Philosophia . . . (Meditations on First Philosophy)*: SPINOZA, *Ethics;* LEIBNIZ, *Monadologie (Monadology and Other Philosophical Writings)*; KANT, *Kritik der reinen Vernunft (Critique of Pure Reason);* 19th century—HEGEL, *Phänomenologie des Geistes (Phenomenology of Mind);* FRANCIS HERBERT BRADLEY, *Appearance and Reality.*

For Rationalism in the theory of knowledge, see BRAND BLANSHARD, *Reason and Analysis* (1962); GEORGE BOAS, *Rationalism in Greek Philosophy (1961)*; ERNST CASSIRER, *Die Philosophie der Aufklarung (1932;* Eng. trans., *Philosophy of the Enlightenment, 1951)*; M.R. COHEN, *Reason and Nature: An Essay on the Meaning of Scientific Method,* 2nd ed. *(1953)*: A.C. EWING, *Idealism: A Critical Survey* (1934); H.H. JOACHIM, *The Nature of Truth (1906);* A.E. MURPHY, *The Uses of Reason (1943);* H.J. PATON, *In Defence of Reason (1951);* BERTRAND RUSSELL, *Problems of Philosophy (1912);* W.H. WALSH, *Reason and Experience (1947).*

For Rationalism in metaphysics, see the classics listed above. For two outstanding examples from the present century, see J.M.E. MCTAGGART, *The Nature of Existence, 2* vol. (1921–27), together with the commentary of C.D. BROAD, *Examination of McTaggart's Philosophy,* 2 vol. (1933–38); and ALFRED NORTH WHITEHEAD, *Process and Reality (1929).* For Rationalism in ethics, see WILLIAM WOLLASTON, *The Religion of Nature Delineated (1722);* and KANT, *Die Metaphysik der Sitten (1785;* Eng. trans., *The Metaphysics of Morals, 1799).* For early forms of the appeal to self-evident rules, see RICHARD PRICE, *A Review of the Principal Questions in Morals (1758).* For later types of Rationalism, see G.E. MOORE, *Principia Ethica (1903);* W.D. ROSS, *The Right and the Good* (1930), and *Foundations of Ethics (1939).*

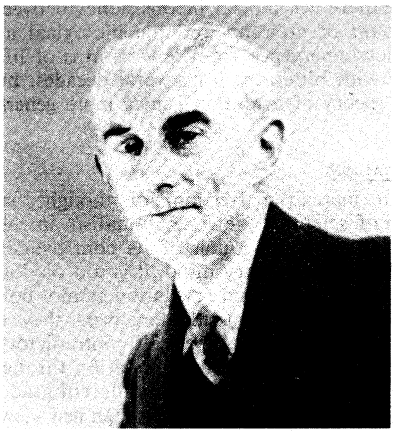
For Rationalism in religion, excellent standard works are: W.E.H. LECKY, *History of the Rise and Influence of the Spirit of Rationalism in Europe,* 2 vol. (1865); A.D. WHITE, *History of the Warfare of Science with Theology in Christendom, 2* vol. *(1910);* J.M. ROBERTSON, *A Short History of Freethought, Ancient and Modern, 2nd* ed., *2* vol. *(1906);* A.W. BENN, *History of English Rationalism in the Nirzeteenth Century, 2* vol. *(1906);* J.B. BURY, *A History of Freedom of Thought (1913).* SIGMUND *FREUD, Die Zukunft einer Illusion (1927;* Eng. trans., *The Future of an Illusion,* 1928), offers a psychoanalytic study of religious belief.

(B.Bl.)

Ravel, Maurice

One of the most original composers of his time in France. Maurice-Joseph Ravel was an exquisite craftsman, striving always for perfection of form and style, which he judged more valuable and important than an expression of personal feelings. For him, music was a kind of ritual, having its own laws, to be conducted behind high walls, sealed off from the outside world, and impenetrable to unauthorized intruders. When his Russian contemporary Igor Stravinsky compared Ravel to "the most perfect of Swiss watchmakers," he was in fact extolling those qualities of intricacy and precision to which he himself attached so much importance.

By courtesy of the French Embassy, Press and Information Division, New York



Ravel.

Ravel was born at Ciboure, near Saint-Jean-de-Luz, France, on March 7, 1875, of a Swiss father and a Basque mother. His family background was an artistic and cultivated one, and the young Maurice received every encouragement from his father when his talent for music became apparent at an early age. In *1889,* at 14, he entered the Paris Conservatoire, where he remained until *1905.* During this period he composed some of his best known works, including the *Pavune pour une infante défunte* (*Pavane for a Dead Princess),* the *Sonatine* for piano, and the *String Quartet.* All these works, especially the two latter, show the astonishing early perfection of style and craftsmanship that are the hallmarks of Ravel's entire oeuvre. He is one of the rare composers whose early works seem scarcely less mature than those of his maturity. Indeed, his failure at the Conservatoire, after three attempts, to win the coveted Prix de Rome for composition (the works he submitted were judged too "advanced" by ultraconservative members of the jury) caused something of a scandal. Indignant protests were published, and liberal-minded musicians and writers, including the musicologist and novelist Romain Rolland, supported Ravel. As a result, the director of the Conservatoire, Théodore Dubois, was forced to resign, and his place was taken by the composer Gabriel Fauré, with whom Ravel had studied composition.

Maturity of early works

Ravel was in no sense a revolutionary musician. He was for the most part content to work within the established formal and harmonic conventions of his day, still firmly rooted in tonality—*i.e.,* the organization of music around focal tones. Yet, so very personal and individual was his adaptation and manipulation of the traditional musical idiom that it would be true to say he forged for himself a language of his own that bears the stamp of his personality as unmistakably as any work of Bach or Chopin. While his melodies are almost always modal (*i.e.,* based not on the conventional Western diatonic scale but on the old Greek Phrygian and Dorian modes), his harmonies derive their often somewhat acid flavour from his fondness for "added" notes and unresolved appoggiaturas, or notes extraneous to the chord that are allowed to remain harmonically unresolved. He enriched the literature of the piano by a series of masterworks, ranging from the early *Jeux d'eau (Fountains,* completed 1901) and the *Miroirs* of 1905 to the formidable *Gaspard de la nuit* (*Gaspard of the Night,* 1908), *Le Tombcau de Couperin (Couperin: In Memoriam,* 1917), and the two memorable concerti. Both concerti date from 1931, and one was written for the left hand alone. Of his purely orchestral works, the *Rapsodie espagnole* (1907) and *Bolero* (1928) are the best known and reveal his consummate mastery of the art of instrumentation. But perhaps the highlights of his career were his collaboration with the Russian impresario Sergey Diaghilev, for whose Ballets Russes he composed the masterpiece *Daphnis et Chloe* (first performed 1912), and with the French writer Colette, who was the librettist of his best known opera, *L'Enfant et les sortilèges (The Child anti the Enchantments).* The latter work gave Ravel an opportunity of doing ingenious and amusing things with the animals and inanimate objects that come to life in this tale of bewitchment and magic in which a naughty child is involved. His only other operatic venture had been his brilliantly satirical *L'Heure espagnole* (first performed 1911). As a songwriter Ravel achieved great distinction with his imaginative *Histoires naturelles, Trois Poémes de Stéphane Mallarmé,* and *Chansons madécasses (Madagascan Songs).*

Collaborations with Diaghilev and Colette

Ravel's life was in the main uneventful. He never married, and, though he enjoyed the society of a few chosen friends, he lived the life of a semirecluse at his country retreat at Montfort-L'Amaury, in the forest of Rambouillet, near Paris. He served in World War I for a short time as a truck driver at the front, but the strain was too great for his fragile constitution, and he was discharged from the army in 1917.

In 1928 Ravel embarked on a four months' tour of Canada and the United States and in the same year visited England to receive an honorary degree of doctor of music from Oxford. That year also saw the creation of *Bolero* in its original form as a ballet, with Ida Rubinstein in the principal role.

The last five years of Ravel's life were clouded by aphasia, which not only prevented him from writing another note of music but also deprived him of the power of speech and made it impossible for him even to sign his name. Perhaps the real tragedy of his condition was that his musical imagination remained as active as ever. An operation to relieve the obstruction of a blood vessel that supplies the brain was unsuccessful, and the end came on December 28, 1937. Ravel was buried in the cemetery of Levallois, a Paris suburb in which he had lived, in the presence of Stravinsky and other distinguished musicians and composers.

MAJOR WORKS

Stage works

BALLET: *Daphnis et Chloé* (first performed 1912; also two suites for orchestra); *La Valse* (1920); *Bolero* (1928; also arranged for piano and piano four hands).

OPERA: *L'Heure espagnole* (1911); *L'Enfant et les sortilèges* (1925).

Orchestral works

ORCHESTRA ALONE: *Shdhe'razade,* overture (1898); *Pavane pour une infante défunte* (completed 1899); *Rapsodie espagnole* (1907).

SOLO INSTRUMENT AND ORCHESTRA: *Tzigane,* for violin and piano or orchestra (1924); *Piano Concerlo in G Major* (1931); *Piano Concerto in D Major for Left Hand* (1931).

Chamber music

String Quartet (1903); *Introduction et allegro,* for solo harp, string quartet, flute, and clarinet (1905); *Trio for Violin, Cello and Piano* (1914); *Sonata for Violin and Cello* (1922); *Sonata for Violin and Piano* (1927).

PIANO MUSIC: *Jeux d'eau,* for piano solo (1901); *Miroirs* (1905, basis for a later ballet); *Sonatine* (1905); Ma *Mère l'Oye,* piano, four hands (1908, later orchestrated and made into a ballet); *Gaspard de la nuit* (1908); *Le Tombeau de Couperin* (1917, later orchestrated); *Valses, nobles et sentimentales* (1911, later orchestrated and made into a ballet).

Vocal music

SONG CYCLES: *Shéhérazade,* for voice and piano or orchestra (1903); *Histoires naturelles,* for voice and piano (1906); *Trois Poèmes de Stéphane Mallarmé,* for voice, piano, string quartet, two flutes, and two clarinets (1913); *Chansons rnade'casses,* for voice, piano, cello, and flute (1926); *Don Quichotte à Dulcinée,* for voice and piano (1932).

Transcriptions of works by other composers

Mettuet pompeux (1918), from *Dix Pièces pittoresques* (1880) by Emmanuel Chabrier; *Pictures from an Exhibition* (1922), a piano suite with the same title by Modest Mussorgsky.

BIBLIOGRAPHY. An invaluable source of information is the *Catalogue de l'oeuvre de Maurice Ravel,* prepared by the FONDATION MAURICE RAVEL (1954), containing full details of every published and unpublished work and comprehensive bibliography. See also the biographical sketch (in French), dictated by Ravel to Roland-Manuel, and published in *La Revue Musicale* (December 1938), a special issue devoted to Ravel; and H.H. STUCKENSCHMIDT, *Maurice Ravel* (1966; Eng. trans., 1968). A fairiy full bibliography may be found in the full-length study of ROLLO HUGH MYERS, *Ravel: Life and Works* (1960).

(R.My.)

Ravenna

Ravenna is an industrial and agricultural city in northern Italy, on the Po River, about five miles from the Adriatic Sea. It arose on one or more islands near the mouth of the Po, just to the east of a wide band of marshy lowland. The city's location was the cause of its greatness and of the important role it played for 350 years as capital of the Western Roman Empire and later of Ostrogothic and Byzantine Italy. It was a safe base and refuge, immune to attack by land, with access to the sea, and it controlled passage to Rimini, one day's march to the south. During the course of centuries, silt from the river gradually filled in the lagoons to the west and between Ravenna and the sea.

History. The earliest inhabitants of Ravenna were probably Italic peoples who moved southward from Aquileia in about 1400 BC. According to tradition, it was occupied by the Etruscans and later by the Gauls, but information about the city before the time of Caesar is scanty.

Early settlement and growth

The Roman era (to 476 AD). Roman interest in the area centred on maintaining communications with Aquileia, Altino, and Spina. In 132 BC a major highway, the Via Popilia, was built through Ravenna connecting Rimini with these towns. In 89 BC Ravenna was given a special relation to Rome when it was designated a *civitas foederata* (allied community), and in the time of Augustus it became a self-governing Roman community. It was from Ravenna that Julius Caesar, on January 12, 49 BC, launched upon the decisive course that led eventually to his control of Rome.

During the war between Mark Antony and Octavian (later the emperor Augustus), Ravenna was occupied by Antony's supporters, who were thus able to communicate with him by sea. The city's special location caused Augustus to select it as home base of the fleet charged with guarding the eastern Mediterranean; for this use the port of Classis, which was able to accommodate 250 warships, was built about three miles from the city; a canal, the Fossa Augusta, connected it with the Po. All the commercial traffic coming from the East that was destined for northern Italy and the rest of southern Europe passed through this canal.

The growth of population and wealth made Ravenna one of the most important cities of Italy. The emperor Claudius built the first encircling defensive wall, and Trajan built an aqueduct. At the end of the 2nd century, Ravenna became the capital of Flaminia; at the end of the 3rd, of Emilia; and finally, of Flaminia and Picenum Annonarium. In the 5th century AD, as the Roman Empire decayed, Ravenna entered the most important period of its history.

Capital of the Western Roman Empire

In 402 the Western emperor Honorius, seeking a refuge from the invading Goths, took up residence in the city. From then until 476 it remained the imperial capital of the West and as such was embellished with magnificent monuments. With the fall of the Western Empire, it became the capital of Italy under the Ostrogothic kings Odoacer (reigned 476–493) and Theodoric (reigned 493–526). At the same time it became an episcopal see and acquired great prestige under Bishop Peter Chrysologus (died 450).

Under Byzantine rule (540–751). Ravenna played an important role in the Byzantine war to recapture Italy from the Goths; in 540 it was occupied by the Byzantine general Belisarius; when the war ended and the prefecture of Italy was established, Ravenna became capital.

The Byzantine emperor Justinian I (reigned 527–65) had Maximian of Pola named bishop of Ravenna (with the title of ***archiepiscopus),*** and the circumstances of the moment made him in effect the primate of Italy. Maximian dedicated some of the city's major basilicas. The fact that theirs was the church of the Italian capital led the archbishops of Ravenna, in the 7th century, to claim a pre-eminent status; for a time they even obtained autonomy from Rome.

The Byzantine struggle with the Lombards for control of Italy, beginning in 568, went on for practically two centuries; resistance was concentrated at Ravenna, which became the empire's citadel in Italy.

The exarchate of Ravenna. The exarchate (military governorship) of Ravenna was created before 584, and the exarch had the powers of *regnum et principatum Italiae* (kingship and prime authority in Italy); thus Ravenna in the 6th and 7th centuries was the centre of all administrative activity as well as the principal maritime trade centre of Italy. As such, it was a rich, cultivated, and populous city.

Rise of the archbishops of Ravenna

The slackening of control by Byzantium, which was involved in internal and external struggles in the 7th century, allowed the Lombards to gain strength, until the exarch finally had to confine his activity to Ravenna and its vicinity. Many public functions passed to the archbishops and to the popes, who provided for public works and even for the recruiting of armies. Loyalty to the Byzantine government declined when it showed itself incapable of resisting the Lombards. There were recurrent uprisings and, finally, open revolt, and an independent army of the exarchate was created.

Conquest by the Lombards (8th century). In 716 the Lombards under King Liutprand took Classis, and in 727 there was an insurrection against decrees of the Byzantine emperor Leo III banning the use of icons; the exarch was killed, and Liutprand held the city for three years. After the city had been liberated and retaken, Pope Zacharias, in 743, went in person and persuaded Liutprand to give up the city. But in 751 the Lombard king Aistulf captured it, thus ending Byzatine dominion in northern and central Italy. Pope Stephen II called on the aid of Pepin, king of the Franks, who expelled the Lombards in 754.

Domination by the archbishops. Though Ravenna's great historical importance had ended, it continued for some centuries in the role of capital of the former exarchate, now under the dominion of the archbishops. They were strengthened by their alliance with the local aristocracy, by quarrels between emperors and popes and the consequent enfeeblement of the latter, and by the political confusion of the times.

Concessions made by the Ottonian dynasty—which ruled the Holy Roman Empire from 962 to 1024—transformed Ravenna's archbishops into great imperial feudatories, completely independent of papal dominion. Ravenna also became the preferred seat of the later Saxon and Franconian emperors when they came to Italy. The archbishops were among the highest imperial dignitaries, and sided with the empire in its 11th-century struggle with the papacy. In 1080, Henry IV (emperor from 1084) had Archbishop Guiberto of Ravenna elected antipope under the name Clement III.

Decline of Ravenna

Under papal rule (1278–1859). Then began Ravenna's slow decline. The communes, which began to develop in Italy in the 10th or 11th century, dealt the final blow to the temporal power of the archbishops, whose domain finally was reduced to the city of Ravenna itself. After 1278, the city was under papal rule though power was held by the Polentani family. By the late Middle Ages, the city had lost its importance as a port, its trade having gone to the Venetians. In 1441 Venice established direct rule over Ravenna. But in 1509, as a result of the war of the League of Cambrai (against Venice), the city once again became part of the Papal States. In 1512 it suffered fire and devastation following the Battle of Ravenna, in which the city was seized by the French. It was soon recaptured.

Ravenna was doomed to decadence by the silting up of the port and by floods to which the rivers were subject. In the 18th century a new port, Corsini, was built, as was the Canale Candiano, which connected the port with the city. Ravenna's resources were thenceforth agricultural, providing wealth for the principal families, whose prosperity is attested by the handsome palaces of the 17th and 18th centuries.

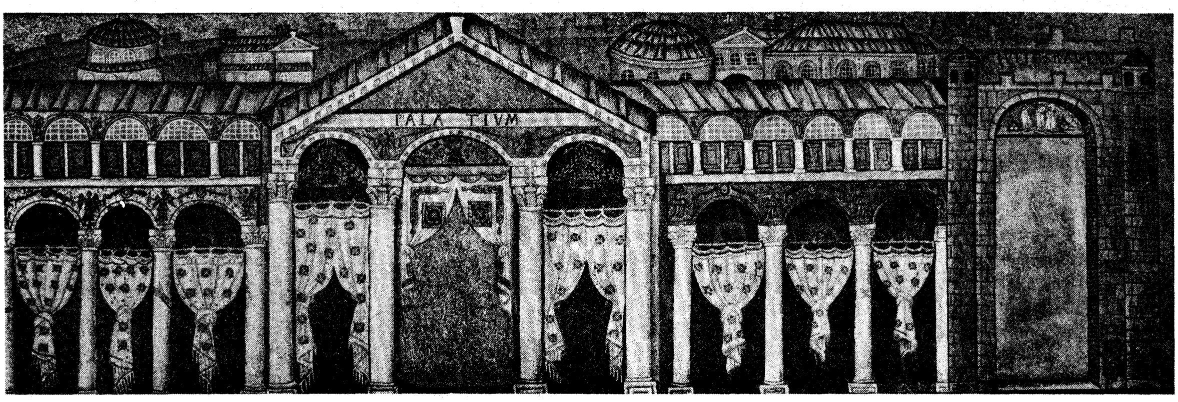
Napoleon occupied Ravenna in 1797. It was returned to the pope in 1815 and, with the exception of brief revolutionary episodes in 1831 and 1848–49, remained under this rule until 1859. In that year the city rebelled and proclaimed its union with the kingdom of Sardinia, which became the kingdom of Italy in 1861.

The modem city

Growth in the 20th century. Ravenna today, though still an agriculture centre, is also a growing industrial and port city. The administrative centre of the province of Ravenna, the city is surrounded by intensely cultivated and productive fields. Its population, 27,500 in 1931, rose in consequence of industrialization to about 60,000 in the 1960s; the size of the built-up area was tripled. In the early 1970s the population was about 133,000. Significant industrialization began about 1950. Principal enterprises include petroleum and natural-gas refining, the production of fertilizers and synthetic rubber, and the processing of oilseeds. Other industries and commercial activities have grown with these; by the 1970s the activity of the port had more than quadrupled. (During 1958–68, tonnage increased from 2,138,000 to 9,879,000.)

Links with the past. ***Modern excavations.*** Since about 1950 there has been extensive research in Ravenna (including the use of aerial photography), on the remains of pre-Roman and Roman times. North of the city the course of the Fossa Augusta was determined; to the southeast a settlement, perhaps the original one, was discovered. In the Classis area the floors and foundations of buildings and some seawalls, probably of the port, were found, but it is not yet possible to determine the plan of the whole.

Monuments. Many of Ravenna's monuments have been destroyed. Those that remain include the church of St. John Evangelist (S. Giovanni Evangelista), built in the 5th century by Galla Placidia, sister of the emperor Honorius, and sometimes mistakenly called her tomb. Older by a few years is the shrine of San Lorenzo, located at the right-hand extremity of the portico of Sta. Croce, attached to the imperial palace and entirely covered with splendid mosaics. Neon (bishop from 451 to 460) is credited with the mosaic decoration of the Catholic baptistery. Also erected in his time was the church of S. Pier Maggiore, now S. Francesco, which retains its original architecture. Another structure with complete mosaic decoration is the oratory of S. Andrea in Arcivescovado, built by Pietro II (archbishop from 494 to 520). The church of Sta. Agata was probably begun in the 5th century and completed some time later. Of Arian monuments there remain the cathedral (now Spirito Santo), almost entirely rebuilt in the 16th century, the adjoining baptistery with mosaic decoration, and S. Apollinare Nuovo, also with mosaic decoration, erected by Theodoric beside the royal palace. Its series of pictures showing the martyrs and the virgins was substituted by Archbishop Agnellus (6th century) for earlier mosaics. The mausoleum of Theodoric culminates in a marvelous cupola made of a single piece of stone from Istria, weighing almost 30 metric tons. The church of S. Vitale, the masterpiece of Byzantine art in



The palace of Theodoric and the rooftops of Ravenna, detail from a 6th-century mosaic in the nave of the church of S. Apollinare Nuovo, Ravenna.
Alinari

Ravenna, was begun by Bishop Ecclesius under the Ostrogothic queen Amalasuntha (died 535), and consecrated in 547; its presbytery is decorated with mosaics including two famous pictures representing Justinian and his wife Theodora. The basilica of S. Apollinare in Classe, begun in 535 and consecrated in 549, has a large mosaic in the vault of the presbytery. From the era of Venetian dominion there remain various palaces and a fortress, the Rocca Brancaleona.

BIBLIOGRAPHY. EDWARD HUTTON, *Ravenna: A Study* (1913), gives the history of the city up to the Renaissance with much information on the Byzantine churches. More recent is G. BOVINI, *The Ancient Monuments* of *Ravenna* (1955) and *Ravenna Mosaics* (1957), illustrated accounts; and A. TORRE, *Ravenna: Sforia di 3000 anni* (1967), a brief synthesis of the fundamental events of the city from its origin to the present day. AGNELLO (AGNELLUS), *Liber Pontificalis Ecclesiae Ravennatis,* ed. by O. HOLDER-EGGER (1878), a 9th-century writer, deals with church events up to 846, mixing history and fable.

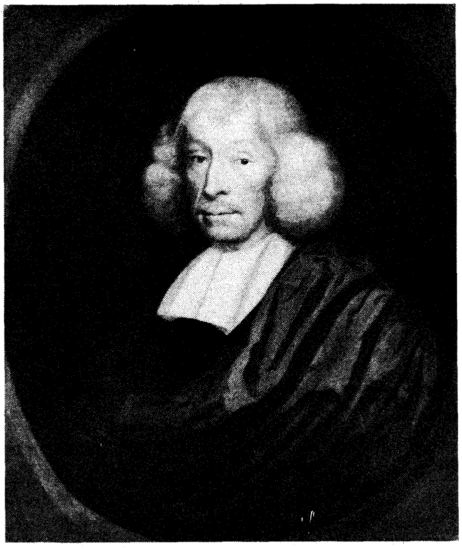
(A.To.)

Ray, John

John Ray was one of the leading naturalists during the age of scientific revolution in the 17th century. Primarily a botanist but learned throughout the entire range of natural history, he contributed significantly to the progress in taxonomy whereby the apparent chaos of organic nature was reduced to an ordered and intelligible classification. His natural theology, based on his work in natural history, influenced popular religious thought throughout the 18th century.

Ray was born on November 29, 1627, the son of the village blacksmith in Black Notley, a hamlet in Essex, England. He attended the grammar school in nearby Braintree. Ray was fortunate in growing up at a time and place in which a fund had been left in trust to support needy scholars at Cambridge. In 1644 he matriculated at one of the colleges there, Catherines Hall, and moved to Trinity College in 1646. Fortune again was with Ray. He had come to Cambridge at the right time for one with his talents, for he found a circle of friends with whom he pursued anatomical and chemical studies. He also progressed well in the prescribed curriculum, taking his bachelor's degree in 1648 and being elected to a fellowship at Trinity the following year; during the next 13 years he lived quietly in his collegiate cloister.

By courtesy of the National Portrait Gallery, London



John Ray, oil painting by an unknown artist. In the National Portrait Gallery, London.

Ray's string of fortunate circumstances ended with the Restoration. Although never an excited partisan, he was thoroughly Puritan in spirit and refused to take the oath prescribed by the Act of Uniformity. In 1662 he lost his fellowship. Prosperous friends, mostly students he had tutored at Trinity, supported him during the subsequent 43 years while he pursued his career as a naturalist.

That career had already begun with the publication of his first work in 1660, a catalog of plants growing around Cambridge; he had worked on the catalog for nine years, during which he trained himself, without benefit of preceptor, in natural history. After he had exhausted the Cambridge area as a subject for his studies, Ray began to explore the rest of Britain. An expedition in 1662 to Wales and Cornwall with Francis Willughby, when Ray's dismissal from Trinity was imminent, was a turning point in his life. Willughby, of a prosperous Warwickshire family, had been fired by Ray's enthusiasm for natural history while a student at Trinity. They now agreed to undertake a study of the complete natural history of living things with Ray responsible for the plant kingdom and Willughby the animal. Implicit was Willughby's material support of Ray.

Travels for collecting specimens

The first fruit of the agreement, a tour of the Continent lasting from 1663 to 1666, greatly extended Ray's firsthand knowledge of flora and fauna. Back in England, the two friends set to work on their appointed task, with Ray established more or less permanently in Willughby's home, Middleton. In 1670 Ray produced a *Catalogus Plantarum Angliae* ("Catalog of English Plants"). Then in 1672 Willughby suddenly died. Ray took up the completion of their project as a sacred duty owed to him; he was aided by an annuity of £60 left to him by Willughby and by his position as tutor to the Willughby children. In 1676 Ray published F. *Willughbeii . . . Ornithologia* (Eng. trans. *The Ornithology* of *F. Willughby. . . ,* 1678) under Willughby's name even though he had contributed at least as much as Willughby. Ray was already at work on F. *Willughbeii . . . de Historia Piscium* (1685; "History of Fish") when Willughby's mother, his remaining support in the household, died. Willughby's widow, who had never liked Ray or his work, immediately forced him from the house. After he completed the "History of Fish," also with Willughby's name, she refused to aid in its publication. Eventually the Royal Society, of which Ray was a fellow, financed the publication.

Meanwhile, Ray returned to his native village of Black Notley. In 1673 he had married a young governess in the Willughby household, Margaret Oakley, and after 11 childless years, they produced four daughters. The girls collected caterpillars and butterflies, which they brought to their father to identify and classify. Among the younger generation of scientists, Ray formed a new circle of friends. From the cottage in which he lived, perhaps not in poverty but certainly not in plenty, Ray, now more than 60 years old, far removed from an adequate library, and tormented by sores on his legs and recurrent diarrhea, poured out an incredible volume of scientific work.

Important publications

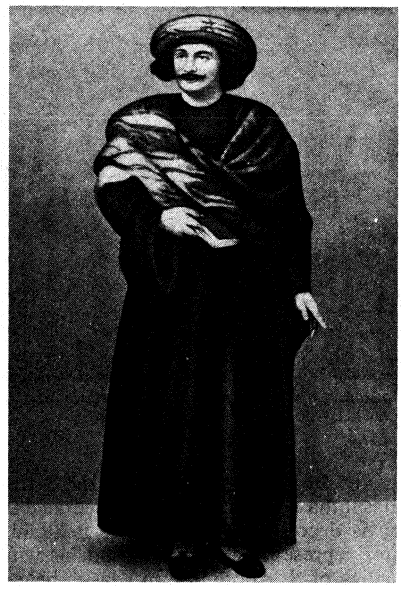
Ray had never interrupted his research in botany. In 1682 he had published a *Methodus Plantarum Nova* (revised in 1703 as the *Methodus Plantarum Emendata . . .*), his contribution to classification, which insisted on the taxonomic importance of the distinction between monocotyledons and dicotyledons, plants whose seeds germinate with one leaf and those with two, respectively. Ray's enduring legacy to botany was the establishment of species as the ultimate unit of taxonomy. On the basis of the *Methodus,* he constructed his master work, the *Historia Plantarum,* three huge volumes that appeared between 1686 and 1704. After the first two volumes, he was urged to compose a complete system of nature. To this end he compiled brief synopses of British and European plants, a *Synopsis Methodica Aviunt et Piscium* (published posthumously, 1713; "Synopsis of Birds and Fish"), and a *Synopsis rnethodica Animalium Quadrupedum et Serpentini Generis* (1693; "Synopsis of Quadrupeds"). Much of his final decade was spent on a pioneering investigation of insects, published posthumously as *Historia Insectorum.* In all this work, Ray contributed to the ordering of taxonomy. Instead of a single feature, he attempted to base his systems of classification on all the structural characteristics, Including internal anatomy. By insisting on the importance of lungs and cardiac structure, he effectively established the class of mammals, and he divided insects according to the presence or absence of metamorphoses. Although a truly natural system of taxonomy could not be realized before the age of Darwin, Ray's system approached that goal more than the frankly artificial systems of his contemporaries. He was one of the great predecessors who made possible Linnaeus' contributions in the following century. Nor was this the sum of his work. Ray had always asserted that "Divinity is my profession," and in the 1690s he also published three volumes on religion. *The Wisdom of God Manifested in the Works of the Creation* (1691), an essay in natural religion that called on the full range of his biological learning, was his most popular and influential book. It argued that the correlation of form and function in organic nature demonstrates the necessity of an omniscient creator. This argument from design, common to all the leading scientists of the 17th century, implied a static view of nature that impeded the development of evolutionary ideas even throughout the 19th century. Still working on his *Historia Insectorum*, John Ray died at Black Notley on January 17, 1705, at the age of 77.

BIBLIOGRAPHY. The definitive study is the biography by CHARLES E. RAVEN, *John Ray, Naturalist* (1942). See also the early sketches in *Further Correspondence of John Ray*, ed. by R.W.T. GUNTHER (1928); and *Memorials of John Ray*, ed. by EDWIN LANKESTER (1846). For an assessment of his work, see RICHARD PULTENEY, *Historical and Biographical Sketches of the Progress of Botany in England*, 2 vol. (1790); and JULIUS VON SACHS, *Geschichte der Botanik, vom 16. Jahrhundert bis 1860* (1864; Eng. trans., *History of Botany, 1530–1860,* 1890); and for the intellectual background to his thought, see R.S. WESTFALL, *Science and Religion in Seventeenth-Century England* (1958); and JOHN GREENE, *The Death of Adam* (1959).

(R.S.W.)

Ray, Rammohan

Rammohan Ray, who is sometimes called the father of modern India, founded a movement for a renaissance of Hindu culture in 19th-century Bengal. He challenged traditional culture by organizing religious dissenters and championing educational, social, and political reforms.



Rarnmohan Ray.

By Courtesy of the Nehru Memorial Museum and Library. New Delhi

He was born on May 22, 1772, in the village of Rādhānagar, Hooghly district, in British-ruled Bengal. His family, which was prosperous and of the Brahmin caste, was in the service of a prominent landholder. His father, Ramkanta Ray, claimed descent from Narottama Thakur, a follower of the 16th-century Bengali Vaiṣṇava (cult of Viṣṇu) reformer, Caitanya. His maternal forebears were the Bhattacharyas of Chatra, chief priests of the Śākta sect (mother-goddess cult), with whom Vaiṣṇavas, historically, had few dealings.

Little is known of his early life and education, but his home life was marked by active religious conflict. He seems to have developed unorthodox ideas at an early age; after the death of his father, his mother unsuccessfully attempted to disinherit him on grounds of apostasy.

Alienated from his family, Rammohan supported himself by moneylending, managing his small estates, and speculating in British East India Company bonds. In 1805 he was employed by John Digby, a lower company official. Through Digby he was introduced to Western culture and literature. For the next ten years Rammohan drifted in and out of British East India Company service as Digby's assistant.

Revival of Vedanta doctrines

Rammohan continued his religious studies throughout this period. In 1803 he had composed a tract denouncing religious divisions and superstition and advocating "natural religion" in which reason guides to ". . . the Absolute Originator who is the first principle of all religions." By 1815 his spiritual roots were more clearly discernible, when he composed a brief summary of the *Vedānta-sūtras* (an ancient Sanskrit religious treatise), in Bengali and Hindi, entitled *Vedāntagrantha.* In the same year he published vernacular and English translations of his abridgment of an unknown compendium of Vedanta doctrines, the *Vedāntasāra.* There followed Bengali and Hindi translations of the *Kena* and *Īśa Upaniṣads* (1816), the *Kaṭha* and *Māṇḍūkya Upaniṣads* (1817), and the *Mundaka Upaniṣad* (1819). The central theme of these texts, for Ray, was the worship of the Supreme God, beyond human knowledge, who supports the universe. These publications established Rammohan, in the eyes of contemporaries, both as a modern exponent of the Vedānta school of Hindu philosophy and as a scriptural nonconformist. He rejected the label of reformer. By translating the sacred Sanskrit *Upaniṣads* into modern Bengali, Rammohan violated a long-standing tradition. In appreciation of these translations, however, the French Société Asiatique, in 1824, elected him to an honorary membership.

In 1815 Rammohan founded the short-lived Atmiya Sabha (Friendly Society) to unite his growing following.

These activities attracted the attention of Baptist missionaries, with whom he began work on a new Bengali translation of the New Testament. This lasted long enough for a dispute to arise over the divinity of Christ. For Rammohan the issue was the same as with his Hindu critics: the unity of the godhead. Yet he published, in 1820, the ethical teachings of Christ, excerpted from the four Gospels, under the title *Precepts* of *Jesus, the Guide to Peace and Happiness.* A vigorous debate with missionaries followed, centring upon the authority of the Bible and the doctrine of the Trinity. Ray closely argued the Unitarian anti-trinitarian position, publicly challenging Christian spiritual supremacy.

Shift to social and political action

In 1823, when the British imposed censorship upon the Calcutta press, Rammohan, as founder and editor of two weekly newspapers, organized a protest, arguing in the spirit of the American and French revolutions, in favour of freedom of speech and religion as natural rights. This protest marked a turning point in Ray's life, away from preoccupation with religious polemic toward social and political action. In 1822 he founded the Anglo-Hindu School and foui years later the Vedanta College. When the Rengal government had proposed a more traditional Sanskrit college, in 1823, Rammohan protested that classical Indian literature would not prepare the youth of Bengal for the demands of modern life. He proposed, instead, a modern, Western curriculum of study. Rammohan, furthermore, led a protest against the outmoded British legal and revenue administration, advocating the separation of the legal and revenue functions, and the adoption of the *pañcāyat* system, the Indian jury-trial system.

Rammohan Ray's opposition to *sati* (ritual death of widows upon the funeral pyres of their husbands) placed him in the centre of the biggest public controversy of his generation. In 1818 he had issued his first pamphlet denouncing the rite from sacred literature. He followed in 1820 with another, arguing from Hindu law. Two years later he published *Brief Remarks Regarding Modern Encroachments on the Ancient Rights of Females According to the Hindu Law of Inheritance.* His newspaper *Sambad Kaurnudi* joined in the already growing public outcry against *sati.* Rammohan's actual influence on the passage of the 1829 act prohibiting *sati* is not clear, though there were many who afterward remarked upon his role in the debate. It has been widely accepted that he had the effect of emboldening the government to act decisively, though he himself favoured caution in governmental interference in public religious life.

In August 1828 he formed the Brahmo Samaj (Society of Brahmā). The deed of the first Sarnaj building declared its purpose to be

> . . . a place of public meeting of all sorts and descriptions of people, without distinction, as shall behave and conduct themselves in an orderly, sober, religious and devout manner: for the worship and adoration of the Eternal, Unsearchable and Immutable Being, who is the Author and Preserver of the Universe. . . .

The Brahmo Samaj was to play an important part, later in the century, as a Hindu movement of reform.

Mission to Europe

The following year, Ray journeyed to England, as the unofficial representative of the titular King of Delhi, to petition the East India Company for an increase in the royal pension. The King granted him the title "Raja," though it was unrecognized by the British. Ray's personal objectives for the visit were to lobby for reforms in Indian government and to support the abolition of *sati.* He was received with adulation, especially by English Unitarians and by King William IV. During the summer of 1833 he travelled to Paris, where he was received by King Louis-Philippe. But his health had declined in Europe, and, on September 27, 1833, he died in the care of Unitarian friends at Bristol, England.

Rammohan Ray's importance in modern Indian history rests upon the facts that he revived interest in the ethical principles of the Vedānta school as a counterpoise to the Western assault on Indian culture and contributed to the popularization of the Bengali language, while at the same time he was the first Indian to apply to the Indian environment the fundamental social and political ideas from the American and French revolutions.

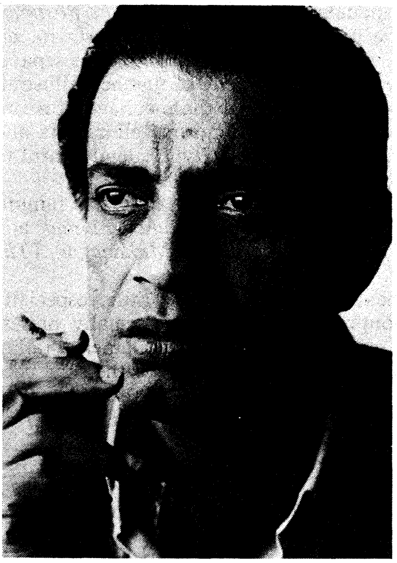
BIBLIOGRAPHY. SOPHIA DOBSON COLLET, *The Life and Letters of Raja Rammohun Roy,* ed. by D.K. BISWAS and P.C. GANGULI, 3rd ed. (1962), is the best known, most comprehensive study of the life of Rammohan Ray, although it contains some inaccuracies. KALIDAS NAG and DEBAJYOTI BURMAN (eds.), *The English Works of Raja Rammohun Roy* (1945), is the most up-to-date edition of Ray's English writings. RAMAPRASADA CHANDA and JATINDRA MAJUMDAR (eds.), *Selections from Official Letters and Documents Relating to the Life of Raja Rammohun Roy* (1938); and JATINDRA MAJUMDAR (ed.), *Raja Rammohun Roy and Progressive Movements in India* (1941), shed light on details of Ray's life from letters, legal documents, and contemporary newspapers. As a critical estimate of Ray's life and writings, SUSHIL KUMAR DE, *Bengali Literature in the Nineteenth Century, 1757–1857,* 2nd ed. rev., pt. 2, ch. 3 (1962), is abbreviated but free of the eulogistic tone that characterizes much of the literature on Ray.

(B.C.R.)

Ray, Satyajit

Of all the film directors of India—which is one of the leading nations of the world in motion-picture production—the most famous is the Bengali Satyajit Ray, who developed entirely apart from the industry as an independent. His first film, *Pather Panchali* (1955), won the attention of audiences throughout the world to his subtle and sensitive work, and his subsequent efforts have confirmed the initial evaluation of him as one of the most important contemporary directors. His inspiration has much in common with the thought and lyricism of the Bengali poet Rabindranath Tagore.

Camera Press—Publix



Satyajit Ray.

Tagore's influence on Ray

Ray was born May 2, 1922, at Calcutta, into a family of artists—his grandfather was a well-known painter, and his father, who died in 1923, was the author of works that are now Bengali classics. The future director completed his secondary schooling at Presidency College in Calcutta. He was then destined for scientific research, but through the influence of Tagore, a family friend, he turned to artistic studies. He took courses given by the poet at Visva-Bharati in Santiniketan.

After Tagore's death in 1941, Satyajit Ray returned to Calcutta. Hoping to make a career in illustrating, he joined a publicity firm and soon became artistic director. In 1945, while illustrating the works of the Bengali writer Bibhuti Bhnsan Bandyopadhyay, the idea of becoming a film maker occurred to him. His interest was heightened by his contact with the great French director Jean Renoir, who made his film *The River* (1951) in India. But Ray's plans to adapt Bandyopadhyay's work to the screen met with many obstacles. Producers hesitated to entrust so much responsibility to a beginner, especially as Ray wanted to employ nonprofessionals and to show the life and customs of the Bengali peasants without anecdotes of the sort that were common in Indian films. Seven years later, in 1952, thanks to the help of friends, he succeeded in raising the capital to start the film. Except for the camera operator, Subrata Mitra, none of the people who plunged into this adventure had ever worked in films. At the end of three months their small capital was exhausted. But after long talks, the West Bengal government agreed to provide the needed finances for the first film of the anticipated trilogy.

The film, *Pather Panchali,* was completed in 1955 and was a great success. Acclaiming its significance, the Asian public was enthusiastic, and the next year it won a prize at the highly prestigious film festival at Cannes, France. This success assured the completion of the other two films of the trilogy: *Aparajito,* which earned the "Golden Lion" award of the 1957 Venice Film Festival, and *Apu Sansar,* made in 1959. In these films, Ray, in the style of an Oriental Dickens, relates the story of a Bengali child, Apu, from youth to maturity. The slow rhythm of the narrative creates a timeless poetic climate, halfway between dream and reality.

Between *Aparajito* and *Apu Sansar,* Ray produced two films of an entirely different sort: *Paras Pathar,* a satirical film adapted from a novel by Parasuram, and *Jalsaghar* (1958; *The Music Room),* an amusing comedy of manners.

As soon as the Apu trilogy was completed he undertook a long documentary illustrating the life and work of Tagore, which was released in 1960. To celebrate the centennial of the poet's birth, in 1961, he brought to the screen three of Tagore's novels grouped under the title of *Teen Kanya* ("Three Girls"). Its Western version (called *Two Daughters)* includes only two stories: *The Postmaster,* which tells the adventures of a village postman, his love affair with a young orphan, and their final separation; and *The Conclusion,* a comedy of manners illustrating the disastrous consequences of the marriage of a student to an eccentric girl. The humorous characters and the portrayal of small details of daily life lend charm to these minor sketches.

Devi (The Goddess), his best film after the famous trilogy, was followed in 1962 by *Kanchenjunga* and *Abhijan (The Country).* He made *Mahanagar (The Big City)* in 1963.

Ray's style

Ray usually rehearses his actors very little, especially when they are nonprofessionals, and he usually prefers the first film of a scene, because of its spontaneity, rather than a "retake." In response to criticisms of the slow pace of his work, he has defended its rhythm as a reflection of the style of the life he is trying to interpret. He has cited as the greatest influences on his work the documentaries *Nanook of the North* (1920–21) and *Louisiana Story* (1948) by the American Robert Flaherty, and *Earth* (1930), set on a collective farm, by the Soviet director Alexander Dovzhenko.

MAJOR WORKS

Pather Panchali (1955; On the Road); Aparajito (1956; The Unvanquished); Jalsaghar (1958; The Music Room); Paras Pathar (1958; The Philosopher's Stone); Apu Sansar (1959; The World of Apu); Devi (1960; The Goddess); Rabindranath Tagore (1960); Teen Kanya (1961; Two Daughters); Kanchenjunga (1962); Abhijan (1962; The Country); Mahanagar (1963; The Big City); Charulata (1963); Kapurush-o-Mahapurush (1965; The Coward and the Saint); Nayak (1966; The Hero); Goopy Gyne, Bagha Byne (1969; The Adventures of Goopy and Bagha); Days and Nights in the Forest (1970).

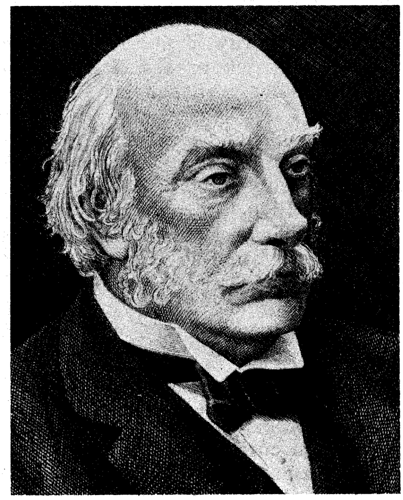
BIBLIOGRAPHY. MARIE SETON, *Portrait of a Director: Satyajit Ray* (1971), is a detailed biographical and critical study, profusely illustrated and including an appendix of Ray's own writings.

(Je.M.)

Rayleigh, John William Strutt, Lord

Lord Rayleigh, last of the great British classical physicists, made notable contributions to every branch of physical science known in his day. His profound insight and enormous capacity for detailed work enabled him to solve many problems previously considered intractable as well as to suggest new lines of research that materially furthered 20th-century science and technology. His sound judgment and diplomatic manner made him an unusually valuable coordinator in British academic and governmental scientific circles.

By courtesy of the International Telecommunication Union, Geneva



Rayleigh, engraving by R. Cottot (20th century).

Rayleigh was born John William Strutt at Terling Place, Witham, in the county of Essex, England, on November 12, 1842. Poor health throughout his childhood and youth frequently interrupted his education, and it was necessary for him to be withdrawn from both Eton and Harrow. In 1857 he began four years of private study under a tutor. In 1861 Strutt entered Trinity College, Cambridge, from which he was graduated with a B.A. as Senior Wrangler (highest honour) in the Mathematical Tripos in 1865, a distinction that foreshadowed his future success. He early developed an absorbing interest in both the experimental and mathematical sides of physical science, and in 1868 he purchased an outfit of scientific apparatus for independent research. In his first paper, published in 1869, he gave a lucid exposition of some aspects of the electromagnetic theory of James Clerk Maxwell, the physicist, in terms of analogies that the average man would understand.

The Theory of Sound

An attack of rheumatic fever shortly after his marriage in 1871 threatened his life for a time. A recuperative trip to Egypt was suggested, and Strutt took his bride, Evelyn Balfour, the sister of Arthur James Balfour, the well-known politician and statesman, on a houseboat journey up the Nile for an extended winter holiday. On this excursion he began work on his great book, *The Theory of Sound,* in which he examined questions of vibrations and the resonance of elastic solids and gases. The first volume appeared in 1877, followed by a second in 1878, concentrating on acoustical propagation in material media. After some revision during his lifetime and successive reprintings after his death, the work has remained the foremost monument of acoustical literature.

Shortly after returning to England he succeeded to the title of Baron Rayleigh in 1873, on the death of his father. Rayleigh then took up residence at Terling Place, where he built a laboratory adjacent to the manor house. His early papers deal with such subjects as electromagnetism, colour, acoustics, and diffraction gratings. Perhaps his most significant early work was his theory explaining the blue colour of the sky as the result of scattering of sunlight by small particles in the atmosphere. The Rayleigh scattering law, which evolved from this theory, has since become,classic in the study of all kinds of wave propagation.

Rayleigh's one excursion into academic life came in the period 1879–84, when he agreed to serve as the second Cavendish professor of experimental physics at Cambridge University, in succession to James Clerk Maxwell, the distinguished Scottish physicist. Rayleigh took his university duties very seriously both with respect to the instruction of students and to the carrying out of a vigorous research program on the precision determination of electiical standards. A classical series of papers, published by the Royal Society, resulted from this ambitious work. After a tenure of five years he returned to his laboratory at Terling Place, where he carried out practically all of his scientific investigations.

A few months after resigning from Cambridge, Rayleigh became secretary of the Royal Society, an administrative post that, during the next 11 years, allowed considerable freedom for research.

Discovery of argon

Rayleigh's greatest single contribution to science is generally considered to have been his discovery and isolation of argon, one of the rare gases of the atmosphere. Precision measurements of the density of gases conducted by him in the 1880s led to the interesting discovery that the density of nitrogen obtained from the atmosphere is greater by a small though definite amount than is the density of nitiogen obtained from one of its chemical compounds, such as ammonia. Excited by this anomaly and stimulated by some earlier observations of the ingenious but eccentric 18th-century scientist Henry Cavendish on the oxidation of atmospheric nitrogen, Rayleigh decided to explore the possibility that the discrepancy he had discovered resulted from the presence in the atmosphere of a hitherto undetected constituent. After a long and arduous experimental program, he finally succeeded in 1895 in isolating the gas, which was appropriately named argon, from the Greek word meaning "inactive." Rayleigh shared the priority of the discovery with the chemist William Ramsay, who also isolated the new gas, though he began his work after Rayleigh's publication of the original density discrepancy. In 1904 Rayleigh was awarded the Nobel Prize for Physics for his role in the discovery of argon; Ramsay received the award in chemistry for his work on argon and other inert elements. The next year Rayleigh was elected president of the Royal Society.

A decided catholicity of taste characterized Rayleigh's choices in research. It was not his habit to devote his exclusive attention to one problem over a long interval but to divert it frequently to a wide variety of topics, such as optics, acoustics, electricity, or fluid mechanics. He preferred to work on several research projects at once, but he was by no means erratic; his work was consistently of high quality throughout his career, both in the brilliance and thoroughness of experimental execution and in the lucidity and precision of his writings. He not only expanded and consolidated existing knowledge in important branches of the physical sciences, but he also opened up important new research frontiers, especially in acoustics and fluid dynamics.

In his later years, when he was the foremost leader in British physics, Rayleigh gave freely of his time and helpful counsel by serving in influential advisory capacities in education and government. In 1908 he accepted the post of chancellor of Cambridge University, retaining this position until his death. He was also associated with the National Physical Laboratory and government committees on aviation and the treasury.

Retaining his mental powers until the end, he worked on scientific papers until five days before his death, on June 30, 1919, at Terling Place.

BIBLIOGRAPHY. ROBERT JOHN STRUTT (4th Baron Rayleigh), ***Life of John William Strutt, Third Baron Rayleigh*** (1924), augmented ed. (1968); R. BRUCE LINDSAY, ***Lord Rayleigh: The Man and His Work*** (1970), contains a complete bibliography of Rayleigh's writings and includes excerpts from his 450 scientific articles.

(R.B.L.)

Realism

Understood in its broadest philosophical sense, Realism connotes any viewpoint that accords to the objects of man's knowledge an existence that is independent of whether he is perceiving or thinking about them. Though it may seem strange to the unphilosophical layman that the independent existence of objects "out there" should be questioned, the philosopher, faced with the many profound challenges that Idealists have posed against the independence of objects, knows that the problem is far from trivial.

The Idealist challenge

Clearly, Idealists have argued, musical tones such as middle C do not have existence as tones in the air; they appear, instead, to be qualities that the mind itself generates when the zppropriate hair cells in the organ of Corti are stimulated. Nor does the colour purple have existence as a quality in the world outside of the mind; there can be, in fact, no such thing as a beam of pure (monochromatic) purple light inasmuch as purple is a unique kind of colour that is perceived when vibrations at the opposite extremes of the visual spectrum (red and violet) are mixed together in the same beam. At least in this one case, the colour seems created by the mind. But if this is so of purple, the Idealists ask, is it not true, also, of all colours? Similarly, under certain circumstances heat is felt as cold and rotation as oscillation. It is not surprising, therefore, that philosphers have asked what, if any, residual properties an object might have in and of itself after due allowance has been made for those qualities that the mind and perspective of the observer have imposed upon it; nor is it surprising that they have asked what, if anything, it would mean to insist on the objective existence of an object of which all of the qualities were mental. Realists, on the other hand have held that, in spite of the foregoing considerations as proposed by the Idealists, there still remains a sense in which objects can have an existence that is independent of minds.

Realism exists, however, in several strikingly different versions: its objects may be, for example, either individual things (such as "the Moon"), or merely particular qualities of things (such as "roundness," "yellowness"), or species and genera of things (such as "moons," "planetary bodies"). In one way or another, however, whether from the viewpoint of the things themselves or from that of the human activities related to them, Realism tends to stress some definite function of the independent existence of objects.

NATURE AND SCOPE OF REALISM

Realism and the problem of knowledge. One of the major problems confronting Realism involves the distinction between private, public, and so-called ontological objects. A private object is a sheer datum (such as a perceived patch of yellow) taken purely as an uninterpreted item in the knower's own inner experience; a public object is one that the mind has projected into an objective conceptual frame of space and time shared in common with other minds, an object that the mind has constituted as a percept (such as the perceived Moon)—though it is still acknowledged to be in part mental (*e.g.*, its yellowness; or its visual size, which is larger when it is near the horizon); and an ontological object is the Kantian "thing-in-itself" (the Moon as it really is), which may as well consist of monads, of God's thoughts, of will, or of action, as of force and matter. Though Realists and Idealists both acknowledge that the knower transcends the private object, they make different assumptions about the relationship between the public and the ontological object: on the one hand, the Realist holds that the physical sphericity, yellowness, and hardness perceived (or perceivable) in the public object are in some degree actual properties of the ontological thing-in-itself; the Idealist, on the other hand, holds that the public object is merely a phenomenon, from which little can be inferred about the underlying *onta* (realities), least of all of their basic qualities, which are probably quite different from the roundness and hardness of the perceived object. The Idealist may, in fact, surmise that the nature of the *onta* is conveyed more faithfully in the fundamental mental tone of the public object—in the colours, feelings, and durations (which are of the nature of mind)—than in its specific material properties. (A third contender, the metaphysical solipsist, would hold that the ontological object does not exist at all.)

Private, public, and ontological objects

Similarly, if a particular thing regarded in its particularity (such as the Moon) is distinguished from a universal—

i.e., an entity comprising the essence of the thing (moonness)—that which it shares with all the other things of the same species or genus (as with the moons of Jupiter)—then a form of Realism can be defined as asserting the independent reality of universals, which it may even exalt above that of particulars.

Contrast with other views

Accordingly, Realism may be variously opposed to the tenets of other philosophical positions. As opposed to Nominalism, which denies that essences (or the specific and generic natures of things) have any reality at all (except as names), and conceptualism, which grants such universals reality only as concepts within the mind, Realism allows to the specific or generic nature of the thing a distinct existence in reality outside the mind. Against Idealism (see above), it asserts that the existence of sense objects (such as the perceived Moon) and that of their qualities is external to thought. In opposition to phenomenalism and sensationalism, which regard objects as comprising mere private volleys or families of disconnected sense fragments, Realism grounds objects in real unified and enduring substances. Unlike conventionalism, a philosophy of science that regards scientific laws and theories as freely chosen constructs devised by the scientist in order to describe reality, Realism holds that laws and theories have determined and real counterparts in things.

The term Realism first aaaeared earlv in the 19th century, though the adjective Realist dates from the late 16th century. These terms have been applied, however—often retroactively—to various systems that have arisen throughout history.

In its broadest scope, the term Realism has application in a number of distinct areas. In literature, art, and aesthetics (see ART, PHILOSOPHY OF; AESTHETICS), in law (see LAW [WESTERN], PHILOSOPHY OF), and in philosophy, it emphasizes real existence or relation to it. The present article is concerned solely with Realism in the philosophical area.

Philosophical senses of Realism. Even within philosophy Realism has a wide range of applications. Though a definitely modern term, Realism is freely used today for tenets of the Greek and medieval epochs, as well as for the modern period.

Basic kinds of Realism. Among philosophical Realisms, two fundamentally different kinds can be distinguished: the Realism of natures and the Realism of things. In the Realism of natures, that which is viewed as having an existence external to the mind is an entity that, in some sense, is set apart in the world of things—an entity that is variously understood as the Form or Idea in which a thing participates, such as "manness" or "bedness" (Platonic Realism), as the essence or *to ti ēn einai*, "the 'what it is' of a thing" (Aristotelian Realism), or as its nature, either absolute, specific, or generic (medieval Realism or the Realism of universals), or, finally, as laws or theoretical models abstracted from scientific observations (see SCIENCE, PHILOSOPHY OF: III. *Deeper issues and broader involvements of science: Status of scientific propositions and concepts or entities).* In the Realism of things, on the other hand, that which is viewed as having an existence external to the mind is the total, concrete, and individual object of experience, which the Realist regards as retaining its chief properties even when left unseen. This Realism, too, can be variously conceived: the externality of the world, for example, can be regarded as simply and obviously given (commonsense Realism); the object itself, though external, can be viewed as the sole entity standing before the mind and grasped by it (neo-Realism); or the object can be conceived as, in some sense, duplicated, so that the mind directly encounters only a counterpart of the external object and not the object itself (critical Realism), a counterpart which was sometimes regarded as a representation of it (representational Realism).

Realism of natures

Distinctions among the Realisms. As previously noted, the term Realism has been applied retroactively to the transcendence of the Platonic Forms or Ideas, to the extent that for Plato the natures of things have, in the ideas of them, an existence more real than that of sensible, individual things (see METAPHYSICS). Yet from its emphasis on ideal as opposed to concrete existence, this Platonic doctrine would be classed as an Idealism instead of a Realism. In the parallel issue in Aristotelianism, the stand that the universals, or specific and generic natures, exist only in the mind but are nonetheless grounded in the real forms of things has been called a moderate Realism. Aristotle himself, however, vigorously denied that the universals have any substantiality *(Metaphysics,* Z: 13–14; 1038^b8–1039^b19), which clearly suggests that, for him, the universals have no existence independently of cognition; this tends, in this first context, to invalidate the designation Realism for the Aristotelian doctrine.

Correspondingly, Realism is used to describe medieval views that allowed species and genera some kind of distinct existence outside the mind. There it meant that not only individual men and individual animals and so on exist outside cognition but also that the specific nature of man and the generic nature of animal and the like have an existence of their own in the outside world. For Realism, objects "fall into" such categories as humanness, mountainness, etc., *naturally.* For its opponents, however, this is not always the case: thus, in terms of a modern illustration, graniteness—that which all the granite rocks share in common—does not exist except as an artificial category set up by the mind (conceptualism) because it merges by imperceptible gradations into diorite or felsite as its mineral composition and texture gradually change.

Yet in actual fact the various medieval doctrines do not fit neatly under these divisions. In the philosophies of several medieval Scholastics, for instance, both the particular thing and the universal are distinguished in one way or another from a third entity, the specific or generic nature *taken absolutely* in itself. This so-called absolute nature is given a "being of its own" by the influential Persian Avicenna, an early-11th-century philosopher and physician; and—in the 13th century—by Henry of Ghent, an eclectic Christian Scholastic, and by the voluntarist Duns Scotus, the greatest medieval British Scholastic, who gave the absolute nature a reality that was distinct in form from the individual thing, but unitively contained in it. Thomas Aquinas gave it no being at all. Though these views reflect radically different metaphysical settings, they all variously bar the natures from real existence when separated in any way from the individual.

In its conventional applications to Greek and medieval thought, accordingly, Realism turns out to be an elusive and even confusing notion. It seems to be an inept way of emphasizing difficulties that are significantly present in the philosophies of these epochs, which require understanding and solution. But the granting of extramental existence to the generic and specific natures has raised more difficulties than it has solved.

Realism of things

All of these ancient and medieval doctrines—whether Realistic, conceptualistic, or nominalistic—accept the external existence of individual sensible things. From this viewpoint, they would all be Realisms in the second main sense of the term, that of the Realism of ordinary things, which is the sense in which Realism is predominantly employed in the modern era. Here it means the epistemological (or theory-of-knowledge) view that things taken as individual wholes exist outside cognition.

HISTORY OF WESTERN REALISM

In these and other related ways, modern writers have seen the philosophic attitude called Realism continually surfacing in the stream of Western thought, suggesting that it is a perennial feature (see PHILOSOPHY, HISTORY OF WESTERN).

Ancient Realism. In pre-Socratic thought, even in Parmenides (late 6th century), known for reducing reality to the One, the relevant reality of the objects of cognition was everywhere assumed. In Plato (5th and 4th centuries BC), the separate and more excellent existence of the natures, or Forms, was strongly asserted at times, though quite often the immanence of the form in individuals was just as surely implied without any satisfactory reconciliation of the contradiction. With Philo of Alexandria, a

Scholastics, Cartesians, and Platonists

Hellenistic Jewish philosopher in the 1st century AD, the existence of the Platonic Forms was located within the mind of God, a view also found in the early 5th century in Augustine of Hippo (*De diversis quaestionibus*, "On Diverse Questions"). In the medieval Augustinian tradition, for instance in Anselm of Canterbury, the influence of this interpretation persisted. In the early 6th century, on the other hand. Boethius, perhaps the intellectual founder of the Middle Ages, in transmitting Aristotelian logic to the West, presented the universal notions with a strong cast of Platonic Realism, while acknowledging that the Aristotelian view was different. Among medieval thinkers in the early 12th century, such as William of Champeaux, the Parisian logician and theologian, the Platonizing tradition of Boethius was dominant, though it was brought under fire by such men as Roscelin, founder of nominalism, who saw universals as mere words. With the stormy controversialist Peter Abelard, foremost dialectician of his time, the Boethian Realism was attacked. But more than mere words were required to justify universality; in his view, universals were concepts signifying real things and had their ultimate basis in the divine ideas, as in the Augustinian tradition.

Moderate Realism

Medieval Realism. The approach of the early medieval thinkers to the problem of universals was made from the side of logic. But it soon involved theological issues, which, when added to the much deeper study of its metaphysical backgrounds made by Boethius, led the scholars to place the relevant questions in a different setting. The natures or common essences of things came to be scrutinized from a threefold viewpoint: as existent in sensible things, as existent in the mind, and as absolutely existent in themselves. This subjected the problem to metaphysical investigation. In that setting, Aquinas allowed neither being nor unity to be attributed to the nature taken absolutely. Duns Scotus, however, accorded it a lesser unity than that of the individual and gave it a kind of being proportionate to this real specific unity, but which required unitive containment of the nature by the individual. In these different ways the nature, so taken, provided the ground for the universal that existed only in the mind. Views incorporating this feature have been called moderate Realisms, though the designation is open to the same objection as it is in its application to Aristotle (see above, *Distinctions among tlze Realisnzs).* In later Scholastic tradition the currents became badly confused, and unending controversy raged on the various kinds of universals and the respective status of each type.

Modern Realism. In the familiar formula *cogito ergo sum* ("I think; therefore, I am") proffered by the first notable modern philosopher, René Descartes, methodical thinking was rooted in thought itself, thus raising the problem of how any material world outside of thought could be reached philosophically. In Descartes and a half century later in the British Empiricist John Locke, an external origin for sensations was accepted, though without any thoroughly philosophical justification. Rather, the denial of an external world was regarded as too absurd to be countenanced. In this perspective Locke's philosophy displayed a commonsense Realism. According to one of Locke's contemporaries, the Cartesian Nicolas Malebranche (known for his claim that God's will is the true cause of motion), religious faith guaranteed the external world. The Cambridge Platonists, a sober group of 17th-century moral and religious Rationalists, in a similar atmosphere of faith and with a Cartesian understanding of sensation, acquiesced in the external existence of sensible things while, against a Neoplatonic background, they accorded a respectively greater reality to the objects of intellectual cognition. For Berkeley, an early 18th-century Empiricist and Idealist, the scriptural guarantee was lacking because matter was nowhere mentioned in the revealed descriptions of the sensible universe; accordingly, in his view, no sensible world outside cognition was left. But in David Hume, who marked the climax of the Empiricist movement, even the cognitive subject, or soul. vanished.

Views based on common sense

Facing the impossibility of a genuine philosophical justification for arguing to an external world from the starting point of mind or idea, Claude Buffier, an early-18th-century French Jesuit, and shortly later, the Scottish Realists, leaned explicitly on common sense as the motive for accepting the world's external existence. Most prominent in this school was Thomas Reid, an opponent of paradox and skepticism. And John Witherspoon, who was called from Scotland to the presidency of Princeton University, held that "the impression itself implies and supposes something external that communicates it, and cannot be separated from that supposition." Consequently, the attempt of Berkeleian immaterialism "to unsettle the principles of common sense by metaphysical seasoning" could, in his view, never produce conviction.

20th-century Realism. Around the turn of the 20th century, a strong revolt against Kantian subjectivism and the dominant Idealisms appeared in such thinkers as William James, a psychologist and Pragmatist; Bertrand Russell, perhaps the most inflnential logician and philosopher of his time; and G.E. Moore?a meticulous pioneering Analyst. Thus it was that very early in the century philosophers came to use Realism, as opposed to Idealism, for their own ways of thinking. In 1904, James signalled the resurrection of natural Realism. In 1910, W.P. Montague of Columbia University and Ralph Barton Perry of Harvard University and several others signed an article entitled "The Program and First Platform of Six Realists," and followed it with a cooperative volume *The New Realism* (1912). New Realism, or neo-Realism, in defending the independence of known things, explained that in cognition "the content of knowledge, that which lies in or before the mind when knowledge takes place, is numerically identical with the thing known." To other Realists this epistemological monism, as Perry called his theory of knowledge, failed to extricate itself from the egocentric predicament (*i.e.*, from the incapacity of the mind to transcend its private experience) that they all professed to see in the logic of Idealism. Nor could it give a satisfactory explanation of the mind's proneness to error, or even of cognition itself as being significantly different from the things known. Another type of Realism was advanced against neo-Realism in a similarly cooperative volume entitled *Essays in Critical Realism* (1920), by the naturalist George Santayana and several others. To the monism of the neo-Realists such writers opposed an epistemological dualism, in which the object in cognition and the object in reality are numerically two at the time of perception. They divided, however, into a majority group and a minority group on the status of the immediately given object. For the majority group this datum was not an existent but merely an essence; for the others it was an existent—a mental or psychic existent for some and a physical (brain) existent for others. Here agreement failed, and the cooperative effort of the critical Realists soon fell apart.

These and the ensuing discussions left a recognized distinction between representative and direct Realism: for representative Realism the immediate confrontation of cognition occurred over against a mental representation of the external object; for direct Realism the confrontation was immediately with the thing existent outside of cognition. The critical Realists themselves, in claiming that the datum was not an object as such but only the means of perceiving it, disavowed any representationalism; but in others, who proposed that the sense-datum was the image directly apprehended and was markedly different from the physical object, representative Realism was definitely present. Representationalism may also be seen in the Realism of the Belgian Neoscholastic Desire Mercier, who founded the school of Louvain, and in the physiological Idealisms of contemporary neurophysiologists. In essence, representationalism was the inferential procedure employed by Descartes and Locke for reaching the external world. Direct Realism, on the other hand—as defended by recent writers—acknowledged no intermediate object between cognition and the external thing perceived.

Naive and tacit Realisms

Within the ambit of contemporary discussions, naive Realism was the label for any unquestioning belief that things in reality correspond exactly to human cognition of them. Expressly meant as a prephilosophical attitude, naive Realism can hardly be included under philosophical procedures. Yet its appropriateness to the man in the street has also been widely challenged; for the ordinary man is keenly interested in distinguishing critically between reality and figments of cognition and is continually doing so in ordinary life. He does not proceed, however, as did the aforementioned Realisms: he does not first regard the object in terms of its status in cognition and then explore its relation to reality. But to come under the notion as introduced by the Realists, naive Realism must be explained in terms of the cognitional relation—*e.g.*, as one of the "three typical theories of the knowledge relation." It is, accordingly, a philosophical category, though historians and controversialists shun the listing of recognized philosophers under such a title.

Metaphysical Realisms

Further, philosophies that neither bore the name of Realism nor defined reality in terms of its relation to cognition are frequently called Realisms today. Aristotelianism, for example, explained the reality of things through their substantiality, Thomism through their existence in themselves, Scotism through the metaphysical priority of a nature possessed in common, and contemporary linguistic philosophy, as in John Austin, an important mid-20th-century Oxford Analyst, through a completely ostensive view of language; yet all have been seen as Realisms. The process philosophies of the Pragmatist John Dewey and of Alfred North Whitehead, an influential cosmologist and metaphysician, and—still more controversially—the philosophy of Charles Sanders Peirce, an individualistic American logician and Pragmatist, may also be taken as Realisms, even though they did not stress the basic relation to cognition; for these thinkers agreed that things as a fact do have, or may have, existence outside cognition, even though this existence was not reached from cognition nor defined through its relation to cognition. With them the cognitional relation was only an inessential afterthought. Serious interest in explaining as Realism the traditional tenets of pre-Cartesian philosophies may be seen in the writings of many contemporaries. Yet for Aristotle and Aquinas, what was meant by the reality of sensible things is already established metaphysically, through their substantiality or ontal (real) existence, before they are compared with cognition; hence to bring in the further notion of Realism for this purpose seems meaningless. The notion is therefore extraneous to the philosophical procedures of thinkers who locate the starting point of their philosophy in some actuality of the real thing itself; for relation to cognition does not play an operative role in their basic procedures. Only by means of entirely extrinsic bonds can they be grouped with the genuine Realisms.

MAJOR ISSUES AND EVALUATION OF REALISM

From the foregoing survey, the major issues upon which Realism focusses attention stand out clearly. For both speculative and practical reasons, men wish to distinguish sharply between what they call reality and what they recognize or suspect to be merely products of their own cognition. Accordingly, the ancient Platonic concentration on specific and generic natures and, in Aristotle, the essential role played by the universal in reasoning led to a close scrutiny of the way in which these natures exist. Undoubtedly, they exist in human thought. For the Realistically inclined thinker, however, their crucial role tends to demand counterparts if human thinking is to bear on what really exists. Still more drastic are the post-cartesian philosophies in which the existence of external things themselves does not enter human cognition in direct confrontation. Finally, the mathematical and scientific constructs, which have been so fruitful in man's struggle for mastery over nature, seem to require for the Realistically minded thinker some counterpart in the things themselves in order to provide an adequate philosophical explanation of their success.

Problems and current status

When the issues are faced in the foregoing manner, some lines of a procedure common to the various explicit Realisms emerge. Universals, sensations and perceptions, scientific formulas and laws are all found to be existent in cognition. From that sure starting point, attempts are made to show that objects either corresponding to them or identical with them exist outside the mind. That pattern seems to be the general procedure followed in any way of thinking that has spontaneously given rise to the notion of an epistemological Realism and that can, with historical and philosophical significance, be labelled such. As is likewise apparent from the foregoing survey, this way of thinking follows a dubious procedure: Realism is not primarily a doctrine of the existence of things but rather a doctrine of cognition. In Realism, cognition is regarded as the object most present to itself; *i.e.*, a man knows his own thought processes more intimately than anything else. But the genuine Realist seems unwittingly to take the material thing as his model in conceiving cognition. No external object can be more present to a material thing than that material thing itself. If cognition is conceived after this analogy, it will be what is most present to itself and will have to be the starting point from which the Realist reasons. This starting point seems to offer no exit. The objects reached from it can be only internal products or occurrences in the mind, for it offers nothing more basic from which to reason than the cognition itself. Any philosophically genuine Realism seems, in consequence, prone to failure in its basic objective.

Accordingly, Realism, in the senses responsible for the epistemological use of the term, has long since ceased to inspire vigorous debate. In a Platonic tradition that continues in modern thought, however, Realism in respect to the natures of things is by no means dead. In regard to the Cartesian problem of the world's external existence, representative Realism seems to have shared the fate of the sense-datum and to be quite inoperative outside of the wake of neurophysiological writings. But attempts at direct Realism are still made. In the scientific field the opposition to conventionalism and to retaining a merely instrumental status for laws or theories has remained a lively issue (see SCIENCE, PHILOSOPHY OF: III. *Deeper issues and broader involvements of science: Status of scientific propositions and concepts or entities).* Moreover, modern means, such as the electron microscope (which shows molecules in real existence) and the hope of being able to see atoms foretoken a greater correspondence of scientific constructs with the structure of reality than had previously been demonstrated. But this verification process consists in a comparison of reality with thought rather than in any attempt to reach reality from cognition alone.

BIBLIOGRAPHY. There is no single monograph that covers comprehensively the whole topic of Realism. The following works will be helpful for the study of its particular phases.

Medieval Realism: M.H. CARRÉ, *Realists and Nominalists* (1946, reissued 1967).

Neo-realism and critical Realism: E.B. HOLT *et al., The New Realism* (1912, reprinted 1970); DURANT DRAKE *et al., Essays in Critical Realism* (1920, reprinted 1968); RENÉ KREMER, *La Théorie de la connaissance chez les néo-réalistes anglais* (1928); T.E. HILL, "Realistic Theories," *Contemporary Theories of Knowledge,* pp. 77–205 (1961, reissued 1980), good general coverage of the American and English fields.

Neoscholastic Realisms: LÉON NOËL, *Le Rkalisme immkdiat* (1938); and works by ÉTIENNE GILSON: *Le Rkalisme méthodique* (1936), presents the author's own views; ch. 5, "Vade Mecum of a Young Realist," trans. by W.J. QUINN, in R. HOUDE and J. MULLALLY (eds.), *Philosophy of Knowledge* (1960); and *Réalisme Thomiste et critique de la connaissance* (1939), a critique of the leading Neoscholastic views.

Linguistic approach: MARTIN LEAN, *Sense-Perception and Matter* (1953, reprinted 1973); J.L. AUSTIN, *Sense and Sensibilia* (1962).

Universals in contemporary thought: I.M. BOCHENSKI, ALONZO CHURCH, and NELSON GOODMAN, *The Problem of Universals* (1956); FARHANG ZABEEH, *Universals* (1966); PANAYOT BUTCHVAROV, *Resemblance and Identity: An Examination of the Problem of Universals* (1966, reissued 1982); NICHOLAS WOLTERSTORFF, *On Universals: An Essay in Ontology* (1970).

Scientific Realism: BERTRAND RUSSELL, *Our Knowledge of the External World,* rev. ed. (1926, various printings); W.H. WERKMEISTER, *The Basis and Structure of Knowledge* (1948, reissued 1968), bibliography, pp. 420–438; MORITZ SCHLICK, "Are Natural Laws Conventions?" in H. FEIGL and MAY BRODBECK (eds.), *Readings in the Philosophy of Science,* pp.

181–188 (1953); ROMANO HARRÉ, *Theories and Things* (1961); GROVER MAXWELL, "The Ontological Status of Theoretical Entities," in *Minnesota Studies in the Philosophy of Science*, 3:3–27 (1962); J.J.C. SMART, *Philosophy and Scientific Realism* (1963).

Other 20th-century Realisms: G.D. HICKS, *Critical Realism* (1938); JAMES FEIBLEMAN, *The Revival of Realism* (1946, reissued 1972); WILFRID SELLARS, *Science, Perception and Reality* (1963); R.M. CHISHOLM (ed.), *Realism and the Background of Phenomenology* (1961), convenient bibliography, pp. 290–304; J.D. WILD (ed.), *The Return to Reason: Essays in Realistic Philosophy* (1953); E.B. MCGILVARY, *Toward a Perspective Realism* (1956); D.M. ARMSTRONG, *A Materialist Theory of the Mind* (1968); INGEBORG WIRTH, *Realismus und Apriorismus in Nicolai Hartmanns Erkenntnistheorie* (1965); GUSTAV BERGMANN, *Realism: A Critique of Brentano and Meinong* (1967).

(J.O./L.H.St.)

Recife

The capital of the state of Pernambuco, in the northeastern region of Brazil, Recife stands on one of the easternmost points of the Atlantic coast of South America, where two river mouths form a lagoon around a central island.

Recife is Brazil's sixth largest city, with a population in 1980 of almost 1,205,000. It combines historic reminders of Brazil's colonial past with the bold lines and skyscrapers of 20th-century town planning. The city has been called the Venice of Brazil because it is crossed by the Capibaribe and Beberibe rivers, its quarters linked by numerous bridges. It is also a centre of learning and of the arts. The city is the major metropolis of Northeast Brazil. Greater Recife includes not only the old city of Olinda, five miles (eight kilometres) to the north, but also a ring of industrial towns.

The name is derived from the Portuguese *recife*, or *arrecife* ("reef"), an allusion to the offshore chain of rocks that shelters the city's harbour and the neighbouring beaches from the full force of the sea. The reef, however, is breached by the discharge of the Rio Capibaribe into the Atlantic. The name Pernambuco, often applied to the port by itself, is taken from the Tupi Indian expression *para-nambuco* ("broken sea").

The Dutch occupation

History. In the second quarter of the 16th century Recife was merely an anchorage that handled the exports and imports of Olinda's wealthy Portuguese colonists. It was raided by French pirates in 1561 and by the English in 1595. In 1630 it was captured by the Dutch, who held it for 24 years. It prospered under the governorship of Count John Maurice of Nassau and is still sometimes called the Cidade Mauricia in his honour. The Dutch occupation was, however, constantly resisted by the Portuguese settlers, whose campaigns in the nearby hills mark the beginning of Brazilian nationalism.

Recife's libertarian spirit was shown on many occasions. In 1710 the inhabitants revolted against the magnates of Olinda in what is now called the War of the Mascates (*i.e.*, "peddlers") because the small tradesmen of Recife tried to organize a municipality of their own. During the revolt of 1817 a republic was formed that lasted for 90 days. In 1824 there was an attempt to establish the Confederation of the Equator as a republic independent of the Brazilian Empire. This spirit may also be seen in the socialistic revolution of the Praieiros of 1848 and in the series of civil movements for reform, best exemplified by the campaign that culminated in the decree of 1888, abolishing slavery. In 1823 Recife became the official capital of the province of Pernambuco.

Landscape. Recife stands at an altitude of 26 feet (eight metres) above sea level. The climate is humid and hot: annual rainfall varies between 60 and 80 inches (1,500 and 2,000 millimetres), and the temperature ranges from a maximum of about 90° F (32° C) to a minimum of 63° F (17° C). The trade winds, however, play a mitigating role during the hottest season, from October to March.

The *bairros*

The Rio Capibaribe winds about Recife from the west and is joined by the Beberibe River flowing from the northwest. It is to these two rivers, whose branches separate the various *bairros*, or quarters, of the city, that Recife owes its peculiar charm.

Recife, on the Rio Capibaribe.
Carl **Frank**

The *bairro* of Recife proper is that of the port. It can be considered either as an island or as the tip of a peninsula because the sand-built isthmus that links it with Olinda is very narrow. One of the oldest *bairros*, it contains some splendid houses of the colonial period, and banks and consulates line its broader avenues. The port itself has been modernized both for coastal and for oceangoing traffic. Its waterfront, backed by customs warehouses, is almost two miles long and can accommodate ships of 25-feet draft everywhere and ships of 31 feet at some berths. Nearby are the great storage tanks of the petroleum companies that supply the northeastern region of Brazil.

The *bairro* of Santo Antônio stands on an island, formerly named Antônio Vaz, in the lagoon, inland from Recife proper and from the isthmus. The commercial centre of the city, it contains the best cinemas, restaurants, and hotels, in addition to the law courts, the state treasury, the palace of government, the public library, the Santa Isabel Theatre, and the offices of Latin America's oldest newspaper still in circulation, the *Diário de Pernambuco.* Though such magnificent avenues as Guararapes, Dantas Barreto, and Nossa Senhora do Carmo have replaced many of the narrower streets of the colonial era, many churches of historical interest remain: Santo Antônio, São Francisco, Nossa Senhora do Carmo, Espirito Santo, São Pedro dos Clérigos, and others. The beautiful Nassau Bridge memorializes Count John Maurice, whose own palace was on this island.

Adjacent to Santo Antônio is the *bairro* of São José, where narrow streets and colonial houses still predominate, despite the pressures of urban development. The basilicas of the Penha and of São José de Ribamar are noteworthy among the churches.

The Boa Vista quarter

The *bairro* of Boa Vista, on the mainland west of the central island, is partly residential and partly commercial. It is also the centre of Recife's student life, though some schools of the universities have been moved to the Cidade Universitaria in the city's outskirts. The splendid premises of the Faculty of Law are located in Boa Vista.

The outer ring of industrial towns that are regarded as belonging to Greater Recife includes Paulista, Jaboatão, São Lourenço da Mata, and Cabo.

Railroads, highways, airlines, and shipping connect Recife with other parts of Brazil. Guararapes Airport serves both international and domestic flights. The city maintains a bus and trolley service, and there also are many private bus companies.

Population. In the 1980 census metropolitan Recife had almost 2,400,000 inhabitants. The majority are of European origin, but Africans and American Indians are also represented. There is a considerable population of mixed racial background. Most of the people are Roman Catholics, but other Christian persuasions and other religions are also found.

Economic life. Recife has shared in the prosperity of northeastern Brazil that resulted from development promoted by Sudene (Superintendência para o Desenvolvimento do Nordeste), a federal organization; signs of poverty continue, nevertheless, to be widespread. Industries of many types have been promoted in the peripheral areas of Greater Recife, and the city itself has expanded rapidly; as in other developing cities, however, there are beggars in the streets. New apartment buildings have continued to rise in the centre of the city and in the suburbs; many homeless people are, however, obliged to sleep outdoors. Branches of the major banks of Rio de Janeiro, São Paulo, and Minas Gerais do considerable business in Recife, and there also are U.S., French, and Italian concerns. The U.S. Agency for International Development (AID) has offices in the Recife financial district.

Administration, services, and institutions. Although the government of Pernambuco, with its seat in Recife, is invested with legislative, executive, and judiciary powers affecting the entire state, the city of Recife is governed by a mayor (*prefeito*) and the Chamber of Councillors (*vereadores*). Security is normally maintained by the civil police, but the state military police and some of Brazil's armed forces are also based in Recife.

Reservoirs outside the city supply Recife with water. Electricity is provided by CHESF (Companhia Hidrélétrica de São Fernando), which supplies hydroelectric power to the metropolitan area.

There are more than 50 hospitals and sanatoriums in Recife, together with organized emergency services.

Universities

Recife provides primary education, at public expense, to some 260,000 children each year, while public and private schools give secondary education to some 60,000. Institutions of higher learning include the Federal University of Pernambuco, the Federal Agricultural University of Pernambuco, the Catholic University of Pernambuco, and the numerous research institutes attached to them. The independent Joaquim Nabuco Institute of social researches, which is distinguished for its anthropological studies, is also located there.

The Pernambuco State Public Library in Recife has more than 80,000 books, and the Federal University's Faculty of Law has a library of 113,000 volumes. The Portuguese Reading Room houses about 40,000 books. Besides the State Museum, there are museums of sugarcane and of popular art. Recife has six daily newspapers, four television stations, and seven radio stations.

The cultural milieu. Recife has a symphony orchestra, a conservatory of music, and several theatrical companies, including the nationally famous Pernambuco Amateur Theatre and the Popular Theatre of the Northeast. Reflecting the area's distinctive cultural composition are the folklore festivals: the Xangô is typically African, while Carnival time is vibrant with the compulsive music of the *passo*, an emotionally and physically exacting dance. Other popular entertainments include the fandango dance, the *bumba-meu-boi* (a pageant with dancing), the *pastoris* (open-air plays), and the *lapinhas* (Nativity scenes).

Beaches

The city is well endowed with open spaces—squares, parks, and gardens. At Dois Irmãos, seven miles from the city centre, there is a nature reserve in a beautifully wooded landscape well appointed for visitors. The region of hills outside the city is included in the National Historic Park of the Guarapes. The beaches, extending for about seven miles in all, offer a special attraction: bright sunlight with Atlantic breezes to temper the heat, placid emerald-green water, and ever-present palm trees providing coconut milk to slake thirst. The beaches nearest to central Recife include Pina, Piedade, and fashionable Boa Viagem, with its lights and playgrounds and an adjacent residential area that is practically a self-contained community. Candeias, Olinda, Rio Doce, and Pau Amarelo, farther away, are served by good roads.

Numerous clubs cater to social life and sports. Some are renowned for their festivities during Carnival time.

(J.L.de L.)

BIBLIOGRAPHY. Reference works providing information on the city and metropolitan area include: JOSUE DE CASTRO, *A Cicade de Recife* (1954); ROBERT M. LEVINE, *Pernambuco in the Brazilian Federation, 1889–1937* (1978); GOVERNO DE ESTADO, FUNDAÇÃO DE DESENVOLVIMENTO DA REGIÃO METROPOLITANA DO RECIFE, *Região Metropolitana do Recife, Plano de Desenvolvimenro Inregrado* (1976); MINISTERIO DA SAUDE, CENTRO REGIONAL DE ESTATISTICA DE SAUDE DO NORDESTE, *Indicadores de saude sobre os municipios das Capitais do Nordesre do Brazil 1977*, contains health statistics for the capitals of northeastern Brazilian states including Recife.

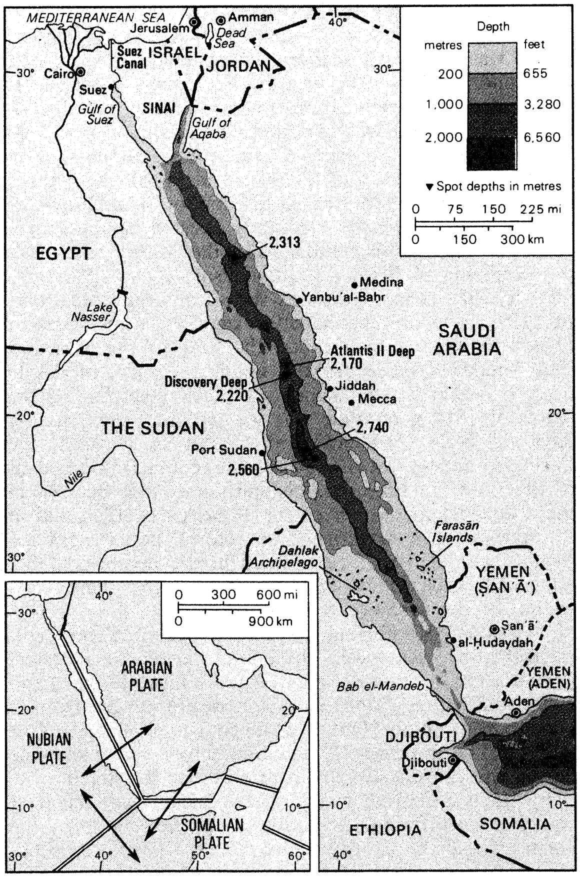
(Ed.)

Red Sea

The Red Sea is a narrow strip of water extending southeastward from Suez for about 1,300 miles (2,100 kilometres) to the Straits of Bāb el-Mandeb, which connects with the Gulf of Aden and thence with the Indian Ocean. The sea separates the coasts of Egypt, The Sudan, and Ethiopia to the west from those of Saudi Arabia and Yemen (San'a') to the east. Its maximum width is 190 miles; its greatest depth 9,580 feet (2,920 metres); and its area approximately 169,000 square miles (438,000 square kilometres). The Red Sea contains some of the world's hottest and saltiest seawater. When the Suez Canal is open, it is one of the most heavily travelled waterways in the world, carrying maritime traffic between Europe and Asia. Its name derived from the colour changes observed in its waters. Normally the Red Sea is an intense blue green; occasionally, however, it is populated by extensive blooms of the algae *Trichodesmium erythraeum*, which, upon dying off, turn the sea a reddish-brown colour. (For associated physical features, see ARABIAN DESERT; ARABIAN SEA; ADEN, GULF OF; and SINAI DESERT; see also SUEZ CANAL.)

Physiography and submarine morphology. The Red Sea lies in a fault depression that separates two great blocks of the Earth's crust—Arabia and North Africa. The land on either side, inland from the coastal plains, reaches heights of more than 6,560 feet above sea level, with the highest land in the south.

At its northern end the Red Sea splits into two parts, the Gulf of Suez to the northwest and the Gulf of Aqaba to



Inset adapted Irom D P McKenzie, D Davies, and P Molnar, "Plate Tectonics of the Red Sea and East Africa," Nature vol 226, p 244 (April 18, 1970)

Red Sea area. Inset shows the relative motions of the three plates that make up the Red Sea area.

The Gulfs of Suez and Aqaba

the northeast. The Gulf of Suez is shallow—approximately 180 to 210 feet deep—and it is bordered by a broad coastal plain. The Gulf of Aqaba, on the other hand, is bordered by a narrow plain, and it reaches a depth of 5,500 feet. From approximately 28° N, where the Gulfs of Suez and Aqaba divide, south to a latitude near 25° N, the Red Sea's coasts parallel each other at a distance of about 110 miles apart. Here the sea floor consists of a main trough, with a maximum depth of some 4,100 feet, running parallel to the shorelines.

South of this point, and continuing southeast to latitude 16" N, the main trough becomes sinuous, following the irregularities of the shoreline. About halfway down this section, roughly between 20" and 21° N, the topography of the trough becomes more rugged, and several sharp clefts appear in the sea floor. Because of an extensive growth of coral banks, south of 16° N only a shallow narrow channel remains. The sill separating the Red Sea and the Gulf of Aden at Bāb el-Mandeb is raised by this growth; therefore, the depth of the water is only about 380 feet, and the main channel becomes very narrow.

The clefts within the deeper part of the trough are unusual sea floor areas in which hot brine concentrates are found. These patches apparently form distinct and separated deeps within the trough having a north-south trend, whereas the general trend of the trough is from northwest to southeast. At the bottom of these areas are unique sediments, containing deposits of heavy metal oxides from 30 to 60 feet thick.

Most of the islands of the Red Sea are merely exposed reefs. There is, however, a group of active volcanoes just south of the Dahlak Archipelago (16° N), as well as a recently extinct volcano on Jebel Tier.

Relationship to the Great Rift System

Geology. The Red Sea occupies part of a large rift valley in the continental crust of Africa and Arabia. This break in the crust is part of a complex rift system that includes the East African Rift valley, which extends southward through Ethiopia, Kenya, and Tanzania for almost 2,200 miles, and northward for over 280 miles from the Gulf of Aqaba to form the great Wadi Aqaba–Dead Sea–Jordan Rift, also extending eastward for 600 miles from the southern end of the Red Sea to form the Gulf of Aden.

The Red Sea Valley cuts through the Arabian-Nubian Massif, a central mass of Precambrian igneous and metamorphic rocks (*i.e.*, formed deep in the earth under heat and pressure from 570,000,000 to 4,600,000,000 years ago), whose outcrops form the rugged mountains of the adjoining region. The massif is surrounded by Paleozoic marine sediments (from 225,000,000 to 570,000,000 years old). These sediments were affected by the folding and faulting that began in the Late Paleozoic Era; the laying down of deposits, however, continued to take place during this time and apparently continued into the Mesozoic Era (from 65,000,000 to 225,000,000 years ago). The Mesozoic sediments appear to surround and overlap those of the Paleozoic and are in turn surrounded by Early Tertiary sediments (from about 54,000,000 to 65,000,000 years old). In many places large remnants of Mesozoic sediments are found overlying the Precambrian rocks, suggesting that at one time a fairly continuous cover of deposits existed above the older massif.

The Red Sea is considered a relatively new sea whose development probably resembles the Atlantic Ocean in its early stages. The Red Sea's trough apparently formed in at least two complex phases of land motion. The movement of Africa away from Arabia began during the lower Eocene Epoch (50,000,000 years ago). The Gulf of Suez opened up during the lower Oligocene Epoch (about 35,000,000 years ago), and the northern part of the Red Sea in early Miocene time (25,000,000 years ago). The second phase began about 3,000,000 to 4,000,000 years ago, creating the trough in the Gulf of Aqaba and also in the southern half of the Red Sea Valley. This motion, estimated as amounting to 0.59 to 0.62 inches a year, is still proceeding, as indicated by the extensive volcanism of the past 10,000 years, by earthquake activity, and by the flow of hot brines in the trough.

Economic resources. Three major types of mineral resources are found in the Red Sea region: petroleum deposits, evaporite deposits (sediments laid down as a result of evaporation, such as salt, gypsum, and dolomite), and the newly discovered heavy metal deposits in the bottom oozes of the Atlantis II and Discovery deeps, which lie between 21°15′ and 21°30′ N. The oil and gas deposits are being exploited to varying degrees by the nations adjoining the sea. In the mid-1970s the evaporites were utilized only very slightly, primarily on a local basis. Of the heavy metal deposits, none of which had been touched, those contained in the sediments of the Atlantis II Deep alone were estimated as having a $25,000,000,000 value. The sediment of the Discovery Deep also has a significant metalliferous content but at a lower concentration than that in the Atlantis II Deep. These deposits are in the form of fairly fluid oozes, with an average of about 85 percent brine. The average analysis of the Atlantis II Deep deposit reveals an iron content of 29%; zinc 3.4%; copper 1.3%; lead 0.1%; silver 54 parts per million; and gold 0.5 parts per million. The total brine-free sediment estimated to be present in the upper 30 feet of the Atlantis II Deep is about 50,000,000 tons. These deposits appear to extend to a depth of 60 feet below the sediment surface, but the quality of the deposits below 30 feet is unknown.

Mineral deposits

The recovery of sediment located beneath 5,700 to 6,400 feet of water poses problems. But since most of these metalliferous deposits are fluid oozes, it is anticipated that it may be possible to pump them to the surface much the same way as oil. There are also numerous proposals for drying and beneficiating (treating for smelting) these deposits after recovery. It would indeed seem that exploitation is now feasible, provided international agreements can resolve legal difficulties.

Climate. The Red Sea region has very little precipitation in any form, although prehistoric artifacts seem to indicate that there were greater amounts of rainfall in times gone by. In general, the year-round climate makes an active life difficult, for the average temperature varies between 77" and 82" F (25" and 28° C), and there is a very high degree of relative humidity in summer. In the northern part of the Red Sea area, extending down to 19° N, the prevailing winds are north to northwest. Best known are the occasional westerly, or "Egyptian," winds, which blow with some violence during the winter months and are generally accompanied by fog and blowing sand. From latitudes 14° to 16° N the winds are variable, but during the months of June through August strong northwest winds move down from the north, sometimes extending as far south as the Straits of Bāb el-Mandeb; by September, however, this wind pattern retreats to a position north of 16° N. South of 14" N the prevailing winds are south to southeast.

Hydrography. No water enters the Red Sea from rivers, and rainfall is scant; but the evaporation loss, in excess of 80 inches a year, is made up by an inflow through the eastern channel of the Straits of Bāb el-Mandeb from the Gulf of Aden. This inflow is driven toward the north by prevailing winds and generates a circulation pattern in which these low salinity waters (the average salinity is about 36 parts of salt per thousand) move northward. Water from the Gulf of Suez has a salinity of about 40 parts per thousand due to evaporation, and consequently a high density. This dense water moves toward the south and sinks below the less dense inflowing waters from the Red Sea. Below a transition zone, which extends from depths of about 300 to 1,300 feet, the water conditions are stabilized at about 72° F (22° C), with a salinity of almost 41 parts per thousand. This south-flowing bottom water, displaced from the north, spills over the sill at Bāb el-Mandeb, mostly through the eastern channel. It is estimated that there is a complete renewal of water in the Red Sea every 20 years.

Salinity and temperature

In the 1960s it was discovered that below this southward flowing water, in the deepest portions of the trough, there is another transition layer, only 80 feet thick, below which, at some 6,400 feet, lie a number of pools of hot brine. The brine in the Atlantis II Deep has an average temperature of almost 140" F (60" C), a salinity of 256 parts per thousand, and contains no oxygen. There are similar pools of water in the Discovery Deep, and in the Chain Deep (at about 21°18′ N). Heating from below renders these pools unstable so that their contents mix with the overlying waters; they thus become part of the general circulation system of the sea.

Early use as a route to India

History. The Red Sea is one of the first large bodies of water mentioned in recorded history. It was important in early Egyptian maritime commerce (2000 BC) and was used as a water route to India by about 1000 BC. It is believed that it was reasonably well charted by 1500 BC, because at that time Queen Hatshepsut of Egypt sailed its length. Later the Phoenicians explored its shores during their circumnavigatory exploration of Africa in about 600 BC. A deep canal between the Mediterranean and the Red Sea was first suggested about 800 BC by Caliph Hārūn ar-Rashīd of Baghdad. Shallow canals were dug between the Nile and the Red Sea before the time of Christ, but were later abandoned. It was not until 1869 that Ferdinand Marie de Lesseps completed the Suez Canal (*q.v.*) connecting the Red and Mediterranean seas.

Navigation. Navigation in the Red Sea is difficult. The unindented shorelines of the northern half provide few natural harbours; in the southern half the growth of coral reefs has restricted the navigable channel and blocked some harbour facilities. At Bāb el-Mandeb the channel is kept open by blasting and dredging. Atmospheric distortion (heat shimmer), sandstorms, and highly irregular water currents add to the navigational hazards.

Prospects for the future. It seemed clear in the 1970s that the Red Sea would be subjected to extensive study for some time to come because of its unusual geology and its enormous untapped economic potential. Further, thorough understanding of the region would throw light on the study of worldwide continental drift and on the origin of the modern sea floor. Economically, further investigation of its mineral concentrations, in a mineral-hungry world, appeared to be imperative.

BIBLIOGRAPHY. T.D. ALLAN and C. MORELLI, "The Red Sea," vol. 4, pt. *2*, ch. 13, in A.E. MAXWELL (ed.), ***The Sea*** (1969), the most recent and most complete treatment of all aspects of topography, geology, and oceanography of the Red Sea; E.T. DEGANS and D.A. ROSS (eds.), ***Hot Brines and Recent Heavy Metal Deposits in the Red Sea*** (1969), an excellent series of papers in which every aspect of the geology, stratigraphy, paleontology, water circulation, and mineral deposits is reviewed clearly and in detail; C.L. DRAKE and R.W. GIRDLER, "A Geophysical Study of the Red Sea," ***Geophys.*** J., 18:473–495 (1964), the first definitive geologic study of the Red Sea region; R.W. GIRDLER, ***A Review of the Red Sea Heat Flow: Symposium*** (1969), a review dealing with the geophysics and unusual heat flow conditions observed in the Red Sea area; S.T. KNOTT, E.T. BUNCE, and R.L. CHASE, "Red Sea Seismic Reflection Studies," ***Geol. Surv. Pap. Can.***, 66:33–61 (1966), a basic work in which much of the structure and topography of the Red Sea was first delineated; D.P. MCKENZIE, D. DAVIES, and P. MOLNAR, "Plate Tectonics of the Red Sea and East Africa," ***Nature***, 226:243–248 (1970), a paper dealing with the relative motions of the continental plates adjacent to the Red Sea, on the basis of recent seismic data; C. TRAMONTINI and D. DAVIES, "A Seismic Refraction Survey on the Red Sea," ***Geophys. J.*** R. ***Astr. Soc.***, 17:225–241 (1969), useful studv in which the nature of the crust beneath the Red Sea is described.

(B.C.S./W.B.F.R.)

Reed, Walter

Walter Reed, a United States Army pathologist and bacteriologist, proved that yellow fever, for several centuries one of the great scourges of the Western Hemisphere, is transmitted by the bite of a mosquito.

Early career

Reed was born at Belroi in Gloucester County, Virginia, on September 13, 1851. He was the youngest of five children of Lemuel Sutton Reed, a Methodist minister, and his first wife, Pharaba White. In 1866 the family moved to Charlottesville, where Walter intended to study classics at the University of Virginia. After a period at the university he transferred to the medical faculty, completed his medical course in nine months, and in the summer of 1869, at the age of 18 graduated as a doctor of medicine. To obtain further clinical experience he matriculated as a medical student at Bellevue Medical College, New York, and a year later took a second medical degree there. He held several hospital posts as an intern and was a district physician in New York. He decided against general practice, however, and for security chose a military career. In February 1875 he passed the examination for the Army Medical Corps and was commissioned a first lieutenant.

Reed.
The Bettmann Archive

After marrying Emilie Lawrence in April 1876, Reed was transferred to Fort Lowell, Arizona, where his wife soon joined him. During the next 18 years—changing stations almost every year—Reed was on garrison duty, often at frontier stations. His letters provide vivid pictures of the rigours of frontier life. In 1889 he was appointed attending surgeon and examiner of recruits at Baltimore. He had permission to work at the Johns Hopkins Hospital, where he took courses in pathology and bacteriology. In 1893 Reed was assigned to the posts of curator of the Army Medical Museum in Washington and of professor of bacteriology and clinical microscopy at the newly established Army Medical School. During the Spanish–American War of 1898 he was appointed chairman of a committee to investigate the spread of typhoid fever in military camps. Its report, not published until 1904, revealed new facts regarding this disease. On the completion of the committee's work in 1899, he returned to his duties in Washington. Almost immediately he became involved in the problem of yellow fever. The result was a brilliant investigation in epidemiology.

The problem of yellow fever

During most of the 19th century it had been widely held that yellow fever was spread by fomites—*i.e.*, articles such as bedding and clothing that had been used by a yellow-fever patient. As late as 1898, a U.S. official report ascribed the spread to this cause. Meanwhile, other methods of transmission had been suggested. In 1881 the Cuban physician and epidemiologist Carlos Juan Finlay began to formulate a theory of insect transmission. In succeeding years he maintained and developed the theory but did not succeed in proving it. In 1896 an Italian bacteriologist, Giuseppe Sanarelli, claimed that he had isolated from yellow-fever patients an organism he called ***Bacillus icteroides.*** The U.S. Army now appointed Reed and army physician James Carroll to investigate Sanarelli's bacillus. It also sent Aristides Agramonte, an assistant surgeon in the U.S. Army, to investigate the yellow-fever cases in Cuba. Agramonte isolated Sanarelli's bacillus not only from one-third of the yellow-fever patients but also from persons suffering from other diseases. Reed and Carroll published their first report in April 1899 and in February 1900 submitted a complete report for publication. It showed that Sanarelli's bacillus belonged to the group of the hog-cholera bacillus and was in yellow fever a secondary invader.

Before this report had actually been published, an outbreak of yellow fever occurred in the American garrison at Havana, and a commission was appointed to investigate it. The members of the commission were Reed, who was to act as chairman, Carroll, Agramonte, and a bacteriologist, Jesse W. Lazear. In the summer of 1900, when the commission investigated an outbreak of what had been diagnosed as malaria in barracks 200 miles (300 kilometres) from Havana, Reed found that the outbreak was actually yellow fever. Of the nine prisoners in the prison cell of the post, one contracted yellow fever and died, but none of the other eight was affected. Reed thought it possible that this patient, and only he, might have been bitten by some insect. Reed therefore decided that the main work of the commission would be to prove or disprove the agency of an insect intermediate host.

Proof of the mode of transmission

On July 27, 1900, an infected mosquito was allowed to feed on Carroll, and he developed a severe attack of yellow fever. Shortly afterward Lazear was accidentally bitten, developed yellow fever, and died. In November 1900, a small hutted camp—Camp Lazear—was established, and controlled experiments were performed. Reed proved that an attack of yellow fever was caused by the bite of an infected mosquito, ***Stegomyia fasciata*** (later renamed ***Aedes aegypti),*** and that the same result could be obtained by injecting into a volunteer blood drawn from a patient suffering from yellow fever. Reed found no evidence that yellow fever could be conveyed by fomites, and he showed that a house became infected only by the presence of infected mosquitoes. In February 1901 official action in Cuba was begun by U.S. military engineers under Maj. W.C. Gorgas on the basis of Reed's findings, and within 90 days Havana was freed from yellow fever.

On his return to Washington in February 1901, Reed continued his teaching duties. He died following an operation for appendicitis on November 22, 1902. The army general hospital located in Washington D.C. was named in his honour.

BIBLIOGRAPHY. H.A. KELLY, *Walter Reed and Yellow Fever* (1906, 3rd ed. 1923), is the only satisfactory biography; see also W.B. BEAN, "Walter Reed," *Arch. Intern. Med.*, 89:171–187 (1952).

(E.A.U.)

Reformation

The Reformation was a religious movement beginning in the 16th century that attempted to purify the Christian Church morally and doctrinally on the basis of biblical norms, and which had far reaching effects in political, economic, and social spheres.

The article is divided into the following sections:

I. Interpretations of the Reformation

The Reformation shattered Christendom and saved the papacy. Such was the judgment of Jacob Burckhardt, the historian of the Renaissance. The outward framework of a society under a single church was gone, but the papacy, fast sinking into a secularized Italian city-state under the frivolous popes of the Renaissance, became again a religious institution under the austere popes of the Counter-Reformation.

The great upheaval of the Reformation has been variously assessed and explained. Since it began in Germany, German nationalists have regarded it as an upsurge from the profound depths of the German soul, but profound depths can be found in the souls of all peoples. After all, the Apostle Paul was a Jew. Others have suggested in broader terms that Protestantism was the form of Christianity suited to the temper of the northern peoples, Catholicism to that of the southern. Plausibility is leant to the thesis because, in the final stabilization, northern Europe was largely Protestant and southern largely Catholic. But the lines were not sharply drawn. In the Germanic lands the Rhine area, Bavaria, and ultimately Austria were Catholic; and Calvinism, the most militant and widely distributed variety of Protestantism, was Latin in origin. Moreover, the demarcation would have been less sharp save for the migration of minorities. In the Low countries 60,000 moved from the Catholic south to the Protestant north, leaving one-third of the houses in Ghent vacant. And, of course, the great example is the Huguenot dispersion from France to Germany, England, Ireland, the American colonies, and South Africa. The religious revolution cannot be explained in terms of race.

Ethnographic, social, and psychological interpretations

The social historians investigate the coincidence of the religious confessions with class structures. There were some. In France, the Huguenot movement had its strength among the nobility. The few Protestants in Spain and Italy were intellectuals. Later, during the Puritan revolution, Anglicanism had its following in the landed aristocracy and the peasantry, Puritanism among the merchants. But in Germany all classes were enlisted. The Marxists look for economic causes and see the basic reason for the upheaval in the restiveness of the disinherited, notably the peasants, who became Anabaptists. But the Anabaptists were not Anabaptists because they were disinherited, they were disinherited because they were Anabaptists. The earlier Marxist historians regarded the religious terminology of the social prophets as a device to enlist the masses. The more recent admit that the leaders believed their own slogans and were led by their illusions to become the unwitting agents of the social revolution. The psychiatrists have scarcely ventured to analyze the entire movement. They fasten rather on individual leaders, notably Martin Luther. But psychoanalysis of the dead is a precarious enterprise, and to explain Luther in terms of parental friction is to miss his sense of cosmic alienation, the terror of the holy. Some interpreters find the locus of Luther's religion in the stomach or the anus because of the misreading of an abbreviation in a manuscript to make it mean that Luther had his evangelical experience while attending to the needs of nature, as if, were it true, elimination would preclude illumination. And should any of these interpretations of Luther be correct there is still the question of how his message should have enlisted millions, most of whom did not have his physical or mental constitution.

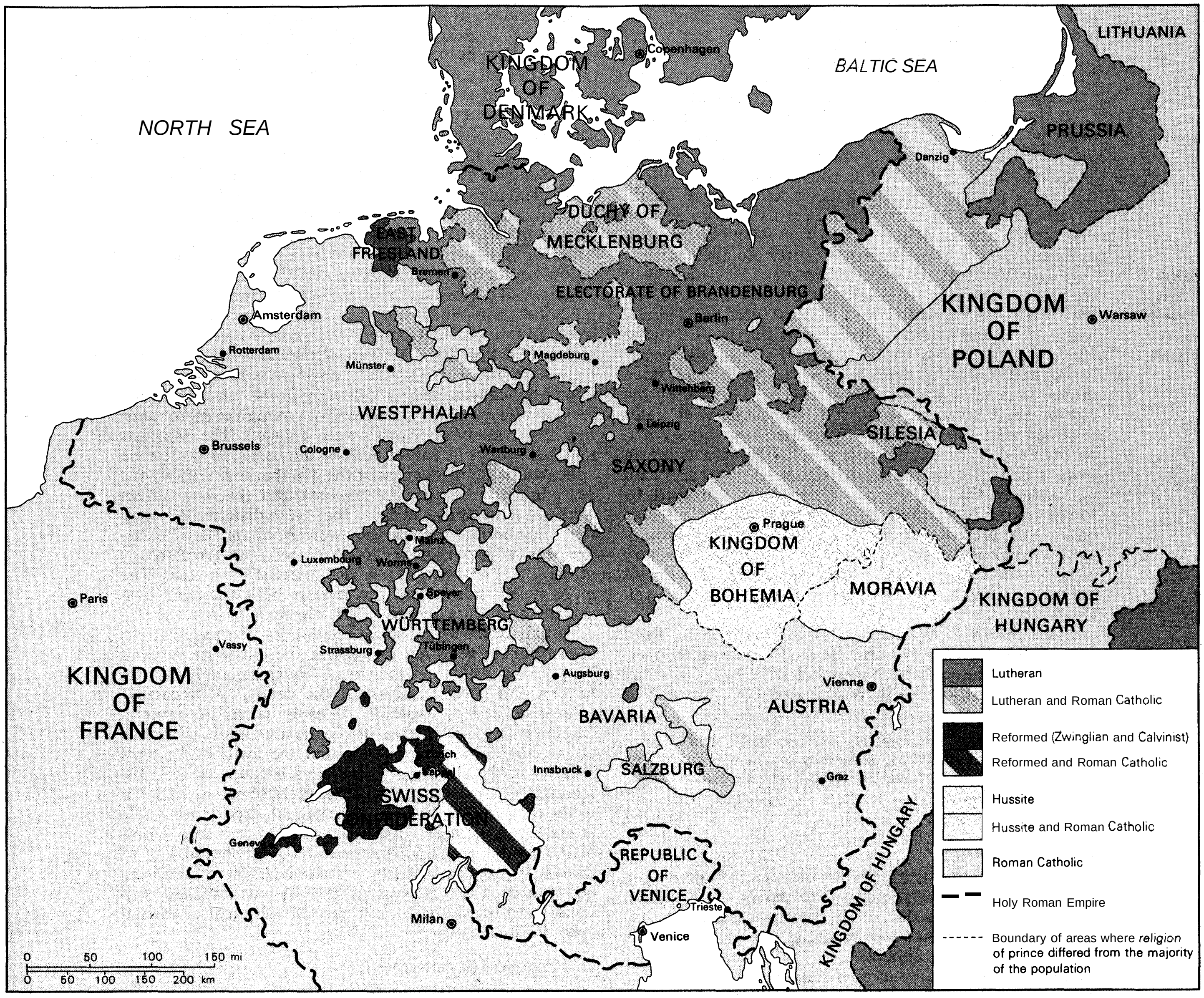
II. Historical development

THE LATE MEDIEVAL CHURCH

Recovery of the lost golden age

To understand the re-formation of the Christian Church one must begin with the de-formation. The Reformation was an attempt to recover a lost golden age of primitive purity as set forth in the Bible. Why and how did the fall occur? Partly because in life as in nature molten lava cools into pumice stone and new eruptions alone can revive the primal heat. But perhaps an even more fundamental reason in Christianity lay in the failures of success. The first success was in expansion. In order to gain converts the church had to accommodate herself to their modes of thought, language, and behaviour, and when they became converts they brought with them much of their former paganism. This was true to a degree in the Roman world, where Christianity was spread chiefly through the conversion of individuals, but much more in the Germanic lands, where whole tribes embraced the faith at the behest of rulers. The ancient gods then survived either as demons or by having their functions transferred to Christian saints. Fertility rites attached themselves to Christian feasts, for example eggs and rabbits to Easter. The Germanic custom of commuting a penalty for crime to a money payment affected the theory of indulgences. The Christian apostles, saints, and even archangels were militarized as Santiago, St. George, and St. Michael who led the hosts in battle.

Another success was in philanthropy. The church so devotedly served the poor that she was inundated with donations. Wealth accumulated and men decayed. Discipline also produced wealth. The great struggle of serious



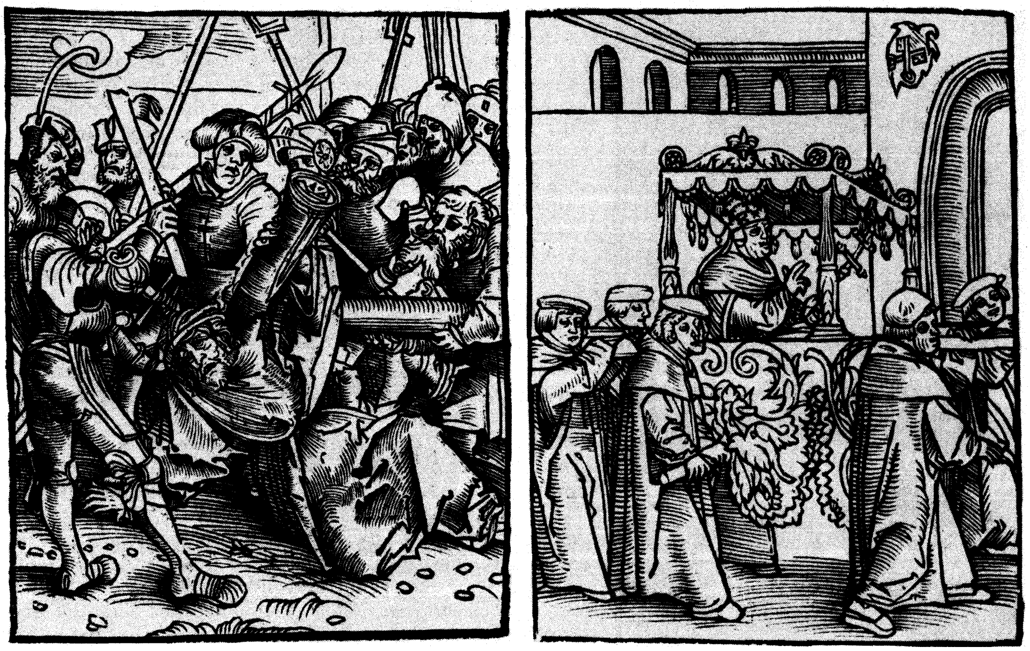
Religious situation in Europe in 1546.
Adapted from *Grosser Historischer Weltatlas;* Bayerischer Schulbuch-Verlag; Munich (1962)

monks in the Middle Ages was to stay poor. Their industry and thrift continually defeated their resolve. A monk formulated the rule that piety begets industry, industry creates wealth, wealth destroys piety, and piety in its collapse dissipates wealth.

A further success was in the realm of government. With the fall of the empire under the barbarian impact, the church took over many governmental functions and became thereby enmeshed in political complications. The great Gregorian reform movement of the 11th and 12th centuries achieved notable success, engendering disastrous failures. The churchmen who initiated the reform were resolved not to withdraw from the world, not to collaborate with the world but to dominate the world through the church after she herself had been reformed. The clergy should be as austere as monks. They must put away their wives. But the enforcement of clerical celibacy resulted in widespread clerical concubinage. The church took the lead in seeking to allay the wars of Christian princes, who should stop devouring each other and go against the common foe, the infidel. The great peace campaign thus ended in the Crusades, and the Crusades in the end attracted the dregs of Europe. The church undertook to direct all Christian rulers on the ground that those who administer the saving sacraments ministering to man's eternal salvation have an authority superior to that of rulers dealing with man's temporal welfare. The papacy in the 13th century succeeded to an amazing degree in regulating the behaviour of rulers, but only at the expense of pitting one power against another and of becoming so enmeshed in political manipulations as to approach the verge of secularization. The international role of the papacy called for immense financial levies on local churches, which the faithful resented, particularly when the money was squandered on wars and luxurious living.

In the realm of theology the great new synthesis of Christianity with congenial elements in Judaism, Islām, and Aristotle, and the surmounting of the dichotomy between reason and faith by a scale of ascents at the hands of Thomas Aquinas produced in the late Middle Ages a recession from universals to particulars and such a rift between reason and faith as to call for two systems of logic. These many failures led those who were actuated by the very ideals of the Gregorian reformers to initiate sects composed of those committed to the practice of the ideal, even in defiance of the church as an organized structure. For 300 years prior to the Reformation the sects swarmed in southern France and northern Italy. Bohemia had the Hussite movement and England that of John Wycliffe and his followers, the Lollards. The programs of these groups centred on the ethical, though some introduced theological ideas of great moment. One was the idea of predestination coupled with a test for identifying if not the elect, then at least the nonelect. The test was moral, with

Sectarian activity



The suffering of Christ (left) contrasted with the worldly glory of the pope. Woodcuts from Passional *Christi und Antichristi,* from the studio of Lucas Cranach, 1521.
By courtesy of the trustees of the British Museum; photographs, John R. Freeman & Co. Ltd.

the conclusion that scandalous popes were "the damned limbs of Lucifer," and the very Antichrist, a figure comparable to Satan. Some predicted and even set dates for the speedy return of Christ to overthrow so depraved a church and to set up the reign of the Spirit.

THE CONTINENTAL REFORMATION: GERMANY, SWITZERLAND, AND FRANCE

Luther, Luther said that what differentiated him from previous reformers was that they attacked the life, he the doctrine of the church. Whereas they denounced the sins of churchmen, he was disillusioned by the whole scholastic scheme of redemption. The assumption was that man could erase his sins one by one through confession and absolution in the sacrament of penance. Luther discovered that he could not remember or even recognize all of his sins, and the attempt to dispose of them one by one was like trying to cure smallpox by picking off the scabs. The whole man is sick. The church held that man is not too sick to make up for bad deeds by some good deeds. God gives to all a measure of grace. If a man lays hold of it and does the best he can, God will reward him with a further gift of grace with which he can perform deeds of genuine merit which will give him credit before God. He may even die with more than enough credits for his salvation. These extra credits constitute a treasury of the merits of the saints, from which the pope can make transfers to those whose accounts are in arrears. The transfer is called an indulgence and for this the grateful recipient makes a contribution to the church. This arrangement proved to be a popular way of raising money particularly because, unlike tithes, it was voluntary and could provoke no resentment. By this means crusades, cathedrals, hospitals, and even bridges were financed. At first the indulgence, according to the Germanic law of commutation of a physical punishment to a fine, applied only to penalties imposed by the church on earth. Then it was extended to penalties imposed by God in purgatory. In Luther's day immediate release from purgatory was being offered, and the remission not only of penalties but even of sins was assured. Thus the indulgence encroached upon the sacrament of penance.

The indulgence system and its effects

Luther was desperately in earnest about his standing before God and Christ. The woodcuts of Christ the Judge on a rainbow consigning the damned to hell filled him with terror. He believed the monastic life to be the way par *excellence* to acquire those extra merits which would more than balance his account. He became a monk and subjected himself to rigorous asceticism, but could never reach the assurance that a sinful pygmy like himself could ever stand before the inexorable justice and majesty of God. Continual recourse to the confessional simply convinced him of the fundamental sickness of the whole man. Then he began to question the goodness of God who would make a man so weak and then damn him for what he could not help. The advice of the confessor to love God brought the retort, "I hate him." Relief came through the study of the psalms. Luther found the 22nd Psalm particularly revealing, because it contains the words quoted by Christ upon the cross, "My God, my God, why hast thou forsaken me?" Evidently then, Christ, being without sin, so identified himself with sinful humanity as to feel himself estranged from God. Christ the Judge seated upon the rainbow had become Christ the Derelict upon the cross, and here the wrath and the mercy of God could find their point of meeting so that God was able to forgive those utterly devoid of merit. He could justify the unjust, and this required of man only that he accept the gift of God in faith. This was the doctrine of justification by faith which became the watchword of the Reformation.

Ninety-five Theses as the catalyst of the Reformation

What this insight meant for the doctrine of indulgences is at once apparent. The great offense was not the financial aspect, even though it stank, but rather the very notion that man could engage in bookkeeping with God. Luther by now had become a professor at the University of Wittenberg and also a pastor. His parishioners were obtaining the indulgences issued by Albert, the new archbishop of Mainz, half of the proceeds to be retained by him as reimbursement for his installation fee as archbishop, the other half to go to the pope for the building of the Basilica of St. Peter's at Rome. For this indulgence Albert made unprecedented claims. If the indulgence were on behalf of the donor himself, he would receive preferential treatment in case of future sin, if for someone else already in purgatory he need not be contrite for his own sin. Remission was promised not only of penalties but also of s$~, and the vendor of the indulgences offered immediate release from purgatory. Against these instructions Luther launched his Ninety-five Theses on All Saints Day of the year 1517. In the theses he presented three main points. The first concerned financial abuses; *e.g.*, if the pope realized the poverty of the German people he would rather that St. Peter's lay in ashes than that it should be built out of the blood and hide of his sheep. The second focussed attention on doctrinal abuses; *e.g.*, the pope has no jurisdiction over purgatory and if he does he should empty the

The sale of indulgences in church; woodcut from the title page of Luther's pamphlet "On Aplas von Rom," published anonymously in Augsburg, 1525.

By courtesy of the trustees of the British Museum: photograph, John R. Freeman & Co. Ltd.

place free of charge. The third attacked religious abuses; *e.g.*, the treasury of the merits of the saints was denied by implication in the assertion that the treasury of the church is the gospel. This was the crucial point.

When the papacy pronounced Luther's position heretical he countered by denying the infallibility of popes and for good measure of councils also. Scripture was declared to be the only basis of authority. Luther found support in many quarters. Already a widespread liberal Catholic evangelical reform sought to correct the moral abuses such as clerical concubinage, financial extortion, pluralism (*i.e.*, the holding of several benefices by one man), and ridiculed the popular superstitions associated with the cult of the saints and their relics, religious pilgrimages, and the like. This movement had representatives in all lands. One thinks of John Colet in England, Jacques Lefèvre in France, Francisco Jiménez de Cisneros in Spain, Juan de Valdés in Naples, and, above all, Erasmus of Rotterdam. Erasmus found nothing amiss in Luther's theses except that he had been too tart as to purgatory, and when the cry of heresy was raised against Luther, he wrote to the Elector Frederick III the Wise, Luther's prince, telling him that as a Christian ruler he was obligated to see to it that his subject should have a fair hearing. Under the impact of the Neoplatonic revival, Erasmus popularized a view destined to find radical implementation by the left wing of the Reformation, namely the view that the body is an impediment to the spirit and that in consequence the sensory aids to religion by way of art, music, and the sacraments should be inwardly surmounted. Others were to say they should be ruthlessly abolished.

Another party that rallied to Luther was that of the German nationalists led by Ulrich von Hutten, who aspired to convert the Holy Roman Empire into a German national state. This program would entail the suppression of the whole system of prince-bishops and could never be achieved without a war with the papacy. Luther was hailed because of his attack on the papacy, though he would not condone the program of violence.

Reaction of Pope Leo X

Yet despite the support from these parties, Luther would have been speedily crushed had Pope Leo X taken seriously the religious side of his office. The secularization of the papacy saved Luther, and he destroyed the secularization of the papacy. At the moment when Luther appeared to be foredoomed, an election for the office of Holy Roman emperor was pending. It was elective and any European prince was eligible, Henry VIII of England, Francis I of France, Charles I of Spain. The Pope wished none of them because the position gave control over Germany, and the augmentation of power to one of the three would destroy the balance. His preference was for a minor prince and none better fitted the role than Luther's protector, Frederick the Wise of Saxony, the senior member of the electoral college. In consequence the Pope dallied in the case of Luther and even after Charles was elected was willing to play Frederick against him. Not until June 1520, nearly three years after the Ninety-Five Theses, was Luther summoned to submit within 60 days. The time was reckoned from the date of the actual delivery of the bull to the person named. So great was the obstruction to Rome on the part even of German bishops that the bull was not handed to Luther until October 10.

Address to the Christian Nobility of the German Nation

He employed the summer of 1520 to bring out some of the great manifestos of the Reformation. The ***Address to the Christian Nobility of the German Nation*** called upon the ruling class in Germany, including the emperor, in whom Luther had not yet lost confidence, to reform the church in externals by returning to apostolic poverty and simplicity. This appeal to the civil power to reform the church was a return to the earlier practice of the Middle Ages when emperors more than once had deposed and replaced unworthy popes. Luther affirmed that the papacy of his day was only 400 years old, meaning that the Gregorian reform had given the church the lead in matters political, encroaching thereby on the sphere of the magistrate on the ground that the lowliest priest does more for mankind than the loftiest king. Luther countered with the doctrine of the priesthood of all believers, including Christian magistrates. Any layman is spiritually a priest, though not vocationally a parson. The Christian ruler, then, being himself a priest, may reform the church in externals, as the church may excommunicate him in spirituals. The liberal Catholic reformers could sympathize with this program except for the identification of the papacy with Antichrist. This savoured of the medieval sects.

The sacraments as defined in ***The Babylonian Captivity***

Another tract dealt with the sacraments. The title was ***The Babylonian Captivity,*** meaning that the sacraments themselves had been taken captive by the church. Luther reduced the number of the sacraments from seven to practically two. The seven are Baptism, the Eucharist or mass, penance, confirmation, ordination, marriage, and extreme unction. Luther defined a sacrament as rite instituted by Christ himself. By this token only Baptism and the Eucharist were strictly sacraments and penance only as confession. Extreme unction, that is anointing with oil those on the verge of death, was dropped entirely. Confirmation went out for a time but was later restored. Ordination continued as a rite of the church. Penance included contrition, confession, and satisfaction. Luther felt that none can be sure of genuine contrition, none can make satisfaction. Confession is wholesome but should be voluntary and may be made to any fellow Christian. Marriage is not a Christian sacrament, because it was not instituted by Christ but by God in the garden of Eden, and valid not only for Christians but also for Turks and Jews. Baptism is to be administered but once only and to babies on the ground of their dormant faith.

This leaves the mass, and at this point Luther gave the greatest offense.The wine, said he, should be given to the laity as well as the bread, as in the Hussite practice. No masses should be said for the dead by the priest alone without communicants, because the Eucharist involves fellowship not only with Christ but also with believers. The most drastic change was that Luther denied the doctrine of transubstantiation, according to which, at the pronouncement of the words of institution, the elements of bread and wine, though retaining their accidents of colour, shape, and taste, nevertheless lose their substance, which is replaced by the substance of the body of Christ as God. This Luther denied, saying that no change is wrought by the words of Christ.

Nevertheless, the body of Christ is physically present upon the altar because Christ said, "This is my body."

Luther and Hus distributing the sacramental bread and wine to the Elector of Saxony and his family. Woodcut by an unknown artist.
BY courtesy of the Lutherhalle, Wittenbero

The doctrine of Real Presence

Therefore, in some inexplicable manner, his body must be "with, in, and under" the elements. But if no change is wrought, how does his body come to be on the altar? Because his body is everywhere. But if everywhere, why especially here? Because in view of human limitations God has decreed two modes of self disclosure, the preaching of the Word and the administration of the sacrament. Here the eyes of the believer are opened. This view undercuts sacerdotalism, since the words of the priest do not bring the body of Christ to the altar. The undercutting of sacerdotalism destroys the hierarchical structure of society, culminating in the papacy.

Diet of Worms as the watershed of Luther's career

But now, what was to be done with Luther? On December 10, instead of submitting, he defiantly burned the papal bull together with a copy of the canon law. The normal course would then have been to excommunicate him outright, but Frederick the Wise insisted that he be given a fair hearing. The natural body to pass judgment would have been a council of the church, but the popes themselves were the greatest obstructionists with respect to the calling of a council because they feared the revival of conciliarism, which in the previous century bade fair to convert the church into a constitutional monarchy. There would have been no Council of Trent save for Luther. Only after another 20 years, when the spread of his teaching left no other expedients, was a council convened. Consequently, his hearing had to be before a secular tribunal, the Diet of the empire meeting at Worms in the winter and spring of 1521. Since this was a secular tribunal the attempt was made to prove that he was not simply a heretic but also a rebel whose views were more subversive of the civil than of the ecclesiastical order, because he was undermining the very principle of authority. Luther was brought before the Diet and given an opportunity to repudiate his books. Had he disclaimed the one on the sacraments the other points might have been negotiated. He acknowledged them all. Would he then disclaim some of their teaching? Who was he to reject the teaching of the ages? Let him give an answer without horns, to which he replied: "I will answer without horns and without teeth. Unless I am convicted by Scripture and plain reason—I do not accept the authority of popes and councils, for they have contradicted each other—my conscience is captive to the Word of God, I cannot and I will not recant anything, for to go against conscience is neither right nor safe. God help me. Amen." The emperor then placed Luther under the imperial ban. The bull of excommunication by the church was formally released only later. Frederick the Wise at this point intervened and wafted Luther away to a place of hiding.

He was concealed for a year at the castle of the Wartburg. During this enforced withdrawal he made perhaps his greatest contribution in that he translated the whole of the New Testament from the Greek text of Erasmus into an idiomatic, pungent, powerful German. In many respects his German helped to create the idiomatic. Nothing did so much to win popular adherence to his teaching as the dissemination of this translation. Earlier translations were by comparison wooden. Now the common man, on reading or merely hearing the simple, vigorous words of Jesus and of Paul would say of Luther's teaching, "By God, he's right."

But some were not so convinced. Many of the liberal Catholic reformers, like Erasmus, recoiled from Luther's paradoxes, from his confidence that his interpretation of Scripture was correct, from his acceptance of the doctrine of predestination which makes of God a tyrant when he elects some and damns others, regardless of their behaviour. The German national movement collapsed. Then in Luther's own circle variant forms of Protestantism arose, which in the aggregate are variously described as the left wing of the Reformation or as the radical Reformation. The terminology does not matter so much as the recognition that no neat classification is possible. The alignments vary in accord with the criterion. If the mainstream of the Reformation is that which made alliances with the state, then Huldrych Zwingli belongs with Luther, John Calvin, and Thomas Cranmer; but if the point is the spiritualizing of the Lord's Supper, then Zwingli belongs with the radicals. If the rejection of infant Baptism is the test, Zwingli moves back to the mainstream and the Anabaptists stand alone. If the denial of the doctrine of the Trinity is determinative, the Socinians are the only radicals. And there were also some individual and rationalist Reformers who formed no communities and dissolved the very structure of the church.

Radical Reformers related to Luther's reform. Two figures emerging in Luther's circle are significant by way of anticipation. One was Karlstadt, who drew the radical inference from the dualism of flesh and spirit that art and music should be abolished as external aids to religion, and the Presence of Christ's body on the altar should be interpreted in a spiritual sense. His program issued in iconoclastic riots. He extended Luther's doctrine of the priesthood of all believers to mean that all laymen are pastors. If one person is assigned the tasks of a parson he should dress like others and, like others, should work with his hands. The clergy not only might but must marry. The sabbath should be strictly observed. This program anticipates the Puritan movement. It entails a blending of spiritualism and legalism. The sensory aids to religion are to be discarded by those advanced in the spiritual life and then snatched away by laws from those still weak.

Karlstadt's and Thomas Miintzer's social and ecclesiastical reforms

A much more disquieting figure was Thomas Muntzer, a man of learning and a creative firebrand, who may be regarded not as the progenitor but as the first formulator of the concept of the Protestant Holy Commonwealth. He believed that the elect, those predestined by God for salvation, could be sufficiently identified to compose a definite group. Luther denied the possibility of distinguishing the elect from the nonelect. Muntzer's test was the new birth in the spirit. The test was not for him an absolute mark, and he recognized that among the wheat there might be some weeds, yet it was an adequate test for the formation of a community bound together by a covenant. The mission of this group was to set up the Kingdom of God on earth, the Holy Commonwealth, by wiping out the ungodly. In the attempt they would have to endure suffering, and here Miintzer drew from German mysticism the theme of walking in Christ's steps toward the cross. But the trial would end in triumph, for the Lord Jesus would speedily come to vindicate his saints and erect his Kingdom. There are obviously incompatibles here, the way of suffering and the infliction of suffering, the feverish activity of man to achieve that which will be established by God. But logical incompatibles fuse at high emotional temperatures. Miintzer appealed to the Saxon princes to implement his program, but they banished him. He found a hearing among the revolting peasants and led them at the Battle of Frankenhausen, where they were butchered and he captured and beheaded. Luther execrated his memory because he seized the sword in defense of the gospel. The Marxists have exalted him as the prophet of social revolution. He was the only one of the Reformers who had a deep feeling for the sufferings of the socially oppressed. In grasping the sword he did not essentially differ from Zwingli, Gaspard de Coligny, or Oliver Cromwell.

Zwingli. Zwingli, the great figure in Swiss Protestantism, was in fact if anything more military than Miintzer because he fell as a combatant with sword and helmet on the field of battle. He became a Reformer independently of Luther, with whom he was entirely in accord as to justification by faith and predestination. At certain points Zwingli drew from Erasmus and Karlstadt, notably with respect to the disparagement of the sensory aids to religion. Zwingli, though an accomplished musician, considered that the function of music is to put the babies to sleep rather than to worship God. The organ was dismantled and the images removed from the cathedral at Zurich. The Lord's Supper was understood by Zwingli in his most extreme period simply as a memorial of Christ's death and, on the part of the recipient, as a public declaration of faith with more significance for the members of the congregation who saw him take his stand than for his own spiritual life. Zwingli could the more readily retain the Baptism of infants because it was simply a recognition that the child belongs to the people of God as the child in the Old Testament belonged by circumcision to Israel. The analogy with Judaism applied at many points, for Zwingli regarded the Christian congregation as the new Israel of God, an elect people, reasonably identifiable, not as with Miintzer by the new birth but by adherence to the faith. This company could be called theocratic in the sense that it was under the rule of God, whom church and state should alike serve in close collaboration. The identification of the whole populace of Zurich with this elect people was the more tenable because those not in accord with the ideal were disposed to leave. As ancient Israel had defended herself by force of arms against the Canaanites and Philistines, so Zwingli approved of even an aggressive war to forestall interference from the Catholic cantons. In the second war of Kappel he fell in 1531.

Zwingli's theocratic ideals

Radical Reformers related to Zwingli's reform. In his circle arose the group who formed the mainstay of the radical Reformation. They shared with Zwingli, and with all the reformers to a degree, the desire to restore the church to the primitive pattern, but they were more drastic in their restitution. Manifestly the early church had not been allied with the state. Luther, Zwingli, and other Reformers saw no sense in forcing the church back into the period when the state was hostile and the Christians were persecuted. After the state became Christian there could very well be a close alliance, as indeed there had been in ancient Israel. The radicals restricted their biblicism to the New Testament and espoused three tenets therefrom that have come to be axiomatic in the United States: the separation of church and state, the voluntary church, and religious liberty. They were called Anabaptists on the ground that, having rejected infant Baptism, they rebaptized adults previously baptized. But they called themselves simply Baptists, denying that they repeated Baptism since the dipping of babies was no Baptism at all. Baptism does not itself regenerate but is only the outward sign of an inner experience, the rebirth in the spirit, of which only an adult is capable. The Anabaptists, so-called, also believed in the possibility of a Christian society whose members were marked both by the conversion experience and also by a highly disciplined deportment. In obedience to the New Testament they repudiated swearing oaths and recourse to violence, whether in war or at the hands of the magistrate. To be sure the rulers bear the sword from God, as the Apostle Paul said, but God instituted the sword because of sin, and its use should be left to sinners. The saints should withdraw from the wicked world.

Influence of Anabaptist tenets

This whole program obviously had political and social aspects and was a threat to that society or any other, for no society, save that of a small sect, has ever renounced the use of the sword. The Anabaptists were marked for extermination by Catholics and Protestants alike. One of their first leaders, Felix Manz, was drowned in Ziirich in 1527. The Diet of Speyer in 1529, at which the Lutherans protested, subjected the Anabaptists to the penalty of death with the concurrence of the Lutherans. Persecution in the first decade eliminated the leaders, most of them educated and moderate men. Less temperate spirits came to the fore, sustaining their courage by setting dates for the speedy coming of the Lord. One band, composed mainly of Anabaptists, took over the town of Munster in Westphalia in 1534 and, contrary to the tenets of their fellows, seized the sword and, in accord with Old Testament practice, restored polygamy. The town was captured by Catholics and Lutherans conjoined and the leaders were executed. Persecution everywhere intensified. In Holland Menno Simons, the founder of the Mennonites, repudiated violence, polygamy, and the setting of dates for the coming of the Lord and returned to the teaching of the early founders. The Mennonites have survived partly by reason of accommodation to military service in Holland, partly by migration first to eastern Europe and then to the Americas. Another group, named Hutterites from Jakob Hutter, was allowed to form communal colonies in Moravia on the estates of tolerant feudal nobles who were willing to drop the demand for military service in return for excellent craftsmanship in field and shop. Because of subsequent persecution these groups also migrated to the New World. The Swiss branch, which survives in the United States, is called the Amish. The entire pattern of ideas has reappeared in various combinations in subsequent history, not only among the Church of the Brethren and the Quakers but among all of the free churches disclaiming a state connection.

Role of Menno Simons

Calvin. Another form of Protestantism was Calvinism, named for John Calvin, a Frenchman educated in humanist and legal studies, who in consequence of a conversion to the Protestant reform had to flee France. In Basel, at the age of 27, he brought out the first edition of his ***Institutes of the Christian Religion,*** which in successive expansions became for centuries the manual of Protestant theology. Calvin was in basic agreement with Luther as to justification by faith and the sole authority of Scripture. On the sacrament of the Lord's Supper he took a mediating position between the radical Swiss and the Lutheran view. Whereas the body of Christ is not everywhere present, his spirit is universal and there is a genuine communion with the risen Lord. Calvin took a middle view likewise with respect to music and art. He favoured congregational singing of the psalms, and this became a characteristic mark of the Huguenots in France and the Presbyterians in Scotland and the New World. As to art he rejected the images of saints and the crucifix (that is, the body of Christ upon the cross), but allowed a plain cross. These modifications do not refute the generalization that Calvinism was alien to art and music in the service of religion, but not in the secular sphere.

As over against Luther there is a shift of emphasis in Calvin whose ***Institutes*** do not begin with justification by faith but with the knowledge of God. Luther found refuge from the terror of God's dispensations in the mercy of Christ. Calvin could the more calmly contemplate the frightfulness of God's judgments because they would not descend upon the elect. Luther, as noted, saw no way of knowing who were the elect. He could not be sure of himself and throughout his life had a continual struggle for faith and assurance. Calvin had certain approximate and attainable tests. He did not require the experience of the new birth which is so inward and intangible, though to be sure later Calvinism moved away from him on this point and agonized over the marks of election. For Calvin there were three tests: the profession of faith, as with Zwingli; a rigorously disciplined Christian deportment, as with the Anabaptists; and a love of the sacraments, which meant the Lord's Supper since infant Baptism was not to be repeated. If a person could meet these three tests let him assume his election and stop worrying.

If one could achieve such assurance, what anenormous release of energy to be directed to the glory of God and the erection on earth of some semblance of a holy commonwealth! The term became common in New England. Calvin's own statement was that "the Church reformed is the kingdom of God." Calvin saw more of a possibility of its realization through the efforts of the elect because he muted the expectation of the imminent return of the Lord. The service of the Kingdom did not require a particular vocation. Any worthy occupation is a divine calling demanding unremitting zeal. Luther had emphasized the secular callings as over against the monastic, which in the Middle Ages alone had been called a vocation. With Calvin the point was not so much that one should accept one's lot and rejoice in the assigned task, however menial, as that the work would contribute to the larger realization of the Christian society.

Geneva as the centre of Calvinistic reform

Calvin had a concrete opportunity for the realization of his ideal, albeit at first only on a small scale. The city of Geneva had recently thrown off the authority of the bishop and of the duke of Savoy and had not yet joined the Protestant Swiss Confederation, though aided in the fight for liberation by the Protestant city of Bern. Through the Bernese, Protestant preachers began to evangelize Geneva. The city was threatened by civil war. The bellicose preacher Guillaume Farel, unable himself to contain the violence he had helped to unleash, laid hold of Calvin merely passing through the city and impressed him into the unwelcome task of leadership. This was to be his vocation. After turbulent years, a banishment and a recall, he was able for the last two decades of his life—he died in 1564—to direct the city which John Knox considered "the most godly since the days of the apostles." There was actually scarcely a feature of Thomas More's *Utopia* that Geneva did not seek to realize.

The program, despite all the turbulence, was the more attainable because of a selective process with respect to the population. At the outset all the Catholics who would not submit to the new regime had to leave. Among those who remained, excommunication from the church, if not removed within six months, meant banishment from the city. Control over excommunication, after a long struggle, came to be entirely in the hands of the church. The state, having long suffered from the abuse of excommunication for political purposes, was loath to concede to the church exclusive control. Abortive attempts to achieve independence had been made by the Protestant churches at Basel and Strassburg. Calvin succeeded, with the result that one who was not in the graces of the church could not for long be a member of the community. A further factor ensuring a select constituency was the influx of 6,000 refugees fleeing from persecution in France, Italy, Spain, and, for a time, from England into a city of 13,000. Thus in Geneva, church, state, and community came to be one. The ministers and the magistrates with differentiated functions were alike the servants of God in the erection of this new Israel; and the comparison with ancient Israel was the more striking and the inner cohesion the more intensified because Geneva also was begirt by foes, the duke of Savoy and the duke of Alba, like the old Canaanites and Philistines.

Max Weber's thesis on capitalism and Calvin

Much discussion has been precipitated by the thesis of Max Weber that Calvinism fostered the spirit of capitalism. This is the reverse of the Marxist contention that economics determine religion. In this instance religion fashioned the course of economics. Weber had in mind not the structure of capitalism with bookkeeping and credit matured through papal finance during the Renaissance, but rather the building up of morale for economic endeavour by the removal of restrictions and the supplying of incentive. The restriction was the removal of the ban on usury (though as a matter of fact Catholic theologians had circumvented it by synonyms); the incentive was to work furiously for the glory of God in gainful employ as a vocation. Wealth ensued. What could be done with it? There should be no expenditure on pleasure, for Calvinism was as austere as monasticism. There remained only philanthropy or return to the business with the accumulation of capital.

There can be no question that capitalism developed mainly in the north, but Calvinist theology need not have been the reason. Minorities of any sort may develop economic power as, for example, the Armenians and the Jews. Spain declined, perhaps through overreaching in the colonial adventure, to be supplanted by England and Holland. And men will work simply to improve the standard of living. Witness in North America the Irish, the Italians, and the Poles, all Catholics. But there is no question that Calvinism engendered an enormous dynamism not only in economics but in many areas, least of all in the artistic, mostly in the political, first in France and the Low Countries, then in England, Scotland, and the New World.

Calvinism in France. The situation in France with respect to the Reformation was not altogether dissimilar to that in Germany because, although the decentralization of government was not so great, nevertheless some of the French provinces enjoyed a considerable autonomy, particularly in the south, and it was in the Midi and French Navarre that the Protestant movement had its initial strength. Then, too, noble houses were continually conspiring to manipulate or eviscerate the monarchy. The religious issues came to be intertwined with the political ambitions. The ruling houses, first the Valois from Francis I through Henry III and then the Bourbon, beginning with Henry IV, sought to secure the stability of the land and the throne by quelling religious strife either by the extermination or toleration of minorities.

The ground was better prepared for the reform of the church in France than in Germany because of the efforts of the liberal Catholics such as the scholar Lefèvre and the bishop of Meaux, Guillaume Briçonnet. King Francis I and his sister Margaret of Angoulême not infrequently intervened to save humanist reformers from the menaces of the obscurantists, and Margaret's daughter, Jeanne d'Albret, the queen of Navarre, a feudatory of France, provided an asylum for the persecuted in her domain, though she did not herself espouse the Huguenot cause until 1560. When Lutheran teaching first began to infiltrate France, Francis I, who would not abet positive heresy, fluctuated in his policy of repression depending on whether he desired a political alliance with the pope, the Turk, or the German Lutherans. The year 1534 precipitated a crisis when placards were posted in Paris savagely attacking the mass. Severe repression followed. Bishop Briçonnet made his submission. Farel fled to Geneva, Lefèvre to Strassburg, Calvin to Basel. Under Henry II, the son of Francis, repression was intensified, particularly when in 1559 France and Spain made peace and thus each was free to devote attention to the suppression of heresy at home. The persecution of the Huguenots, as the Protestants came to be called in France, would have been intense save for the death of the king in a tournament.

The Huguenots and the French religious wars

At this point the rivalry of the noble houses injected itself more overtly into the religious struggle. The crown, with its alternating policy of eradication or recognition, was flanked by two extreme houses for whom the religious issue was of intense concern. The house of Guise was so Catholic as to be willing to call in Spanish aid, and the family of Admiral Coligny so Huguenot as to be willing to court help from England and even from Germany. Under Francis II the Guises were in the ascendant, because the queen, later queen of Scots, was of that house. Some of the Huguenots, foreseeing the suppression in store, hatched the Conspiracy of Amboise, an attempted assassination of the leaders of the Guise party and transferral of power to the house of Bourbon.

This was plainly rebellion and acutely raised a problem with which the Protestants had long been wrestling. The Lutherans had had to face it earlier when the Diet of Augsburg in 1530 gave them a year in which to submit on pain of war. The Lutheran princes then had formed the Schmalkaldic League to resist arms with arms. Luther was loath to condone any use of the sword in defense of the Gospel and absolutely forbade any recourse to violence on the part of a private citizen against the magistrates. This was his reason for disapproval of the Peasants' War. But now the jurists pointed out to Luther that the emperor was an elected ruler and that if he transgressed against the true religion he might be brought to book by the electors, who also were magistrates. Thus arose the doctrine of the right of resistance of the lower magistrate against the higher. The concept lost its pertinence in Germany after the Peace of Augsburg in 1555, which granted toleration to the Lutherans in the territories where they were predominant. Minorities in Lutheran and Catholic lands were granted the right of migration without loss of goods.

But the Calvinists were not included in the peace, and the problem of armed resistance again became acute in France. Calvin would not condone the Conspiracy of Amboise because it was not led by a lower magistrate. The term was now applied to the princes of the blood in line for succession to the throne. This meant the house of Bourbon. The Conspiracy of Amboise failed. Francis II died, and was succeeded by his brother the young Charles IX. The queen mother, Catherine de Médicis, took the lead and sought to avert religious war by granting the Huguenots limited toleration in restricted areas in the edict of 1562. When François, duc de Guise, discovered the Huguenots worshipping outside the prescribed limits, as he claimed, he opened fire. The Massacre of Vassy set off the wars. The Huguenots now were led by a prince of the blood, Louis I, 1st prince de Condé, of the house of Bourbon. Calvin approved. There followed three inconclusive wars. Condé was killed in the first and François, duc de Guise, was assassinated. His son, now Henri, duc de Guise, believed in the complicity of Coligny, the new leader of the Huguenots. At the end of ten years of indecisive conflict, Catherine made another effort at a settlement to be cemented by the marriage of Henry of Navarre, a Bourbon, the son of Jeanne d'Albret and the hope of the Huguenots, and her own daughter Margaret (Marguerite de Valois), a Catholic. The leaders of all parties came to Paris for the wedding. The Duke of Guise made an attempt on the life of Coligny, which failed. Then the Guise, with the connivance of Catherine and her son Charles, who panicked, tried to wipe out all of the leaders of the Huguenot party in the Massacre of St. Bartholomew's Day in August 1572. Other massacres followed in the provinces.

The St. Bartholomew's Day Massacre and the subsequent Edict of Nantes

Charles IX was succeeded by his brother, Henry III, two years later (1574). Such was the revulsion against the massacre that the King could rule only by forming an alliance with the Huguenot Henry of Navarre. A fanatical Catholic was thereby so outraged that he assassinated the King. Both sides had abandoned the fiction of the inferior magistrate and had gone in unabashedly for popular revolution. Henry of Navarre then became Henry IV, but he was unable to take Paris and rule France so long as he was a Protestant. In order to pacify the land he made his submission to Rome and promulgated an edict of toleration for the Huguenots, the Edict of Nantes in 1598. It gave them liberty of worship again in limited areas but full rights of participation in public life. The edict remained in force until the revocation in 1685.

THE REFORMATION IN ENGLAND AND SCOTLAND

Henry VIII. In the meantime the Reformation had taken hold in England. The beginning there was political rather than religious, a quarrel between the king and the pope of the sort that had occurred in the Middle Ages without resulting in a permanent schism, and might not have in this instance save for the total European situation. The dispute had its root in the assumption that the king was a national stallion expected to provide an heir to the throne. England did not have the Salic law, which in France forbade female succession, but England had just emerged from the Wars of the Roses and the fear was not unwarranted that the struggle might be resumed if there were not a male succession. Catherine of Aragon, the queen of Henry VIII, had borne him numerous children of whom only one survived, the princess Mary, and more were not to be expected. The ordinary procedure in such a case was to discover some flaw in the marriage which would allow an annulment or, in the terminology of that day, a divorce. In this instance the flaw was not difficult to find, because Catherine had been married to Henry's brother Arthur, and the law of England, following the prohibition in the book of Leviticus, forbade the marriage of a man with his deceased brother's widow. At the time of the marriage the pope had given a dispensation to cover this infraction of the rule. The question now was whether the pope had the authority to dispense from the divine law. Catherine said there had been no need for a dispensation because her marriage to Arthur had not been consummated and there had been no impediment to her marriage to Henry. The knot would have been cut by some casuistry had Catherine not been the aunt of the Emperor Charles V, who was not prepared to see her cast aside in favour of another wife, and who controlled the pope. Clement VII, wishing neither to provoke the emperor nor to alienate the king, dallied so long that Henry took the matter into his own hands, repudiated papal authority, and in 1534 set up the Anglican Church with the king as the supreme head. The spiritual head was the archbishop of Canterbury, now Thomas Cranmer, who married the king to Anne Boleyn. She bore the princess Elizabeth. By still another wife Henry did have a son who succeeded as Edward VI.

England's problem of succession

Although the basic concern of Henry was political, the alterations in the structure of the church gave scope for a reformation religious in character. Part of the impulse came from the survivals of Lollardy, part from the Lutheran movement on the continent, and even more from the Christian humanism represented by Erasmus. The major changes under Henry were the suppression of the monasteries, the introduction of the Bible in the vernacular in the parish churches, and permission to the clergy to marry, though this was later revoked. The resistance to Henry's program was not formidable and the executions resulting were not numerous. Henry was impartial in burning some Lutherans who would not submit to his later reactionary legislation and toward some Catholics who would not accept the royal supremacy over the church, notably John Fisher and Thomas More.

Under Edward VI, the boy-king, the English Reformation took on a more distinctly Protestant character. Cranmer prepared a liturgy in English, The *Book of Common Prayer,* in two editions. The first used the formula in the Lord's Supper: "The body of our Lord Jesus Christ which was given for thee preserve thy body and soul unto everlasting life." This was susceptible of a Lutheran and even a Catholic interpretation. The second edition substituted: "Take and eat this in remembrance that Christ died for thee and feed on him in thy heart by faith with thanksgiving." This was susceptible of a Calvinist and even a Zwinglian interpretation as a memorial and an act of spiritual communion.

The Book of Common Prayer

After Edward a woman did succeed to the throne, the princess Mary who, married to Phillip II of Spain, reintroduced Catholicism. The leaders of the Anglican Church went to the stake, some 288 in number. Among them was Archbishop Cranmer. Mary was succeeded in 1558 by her half sister Elizabeth, who espoused a moderate Protestantism constructed on the idea of comprehension; that is, the inclusion of all the inhabitants of the land in one national church. To this end the doctrinal requirements were not rigidly pressed and the formulas for the Lord's Supper in the two prayer books of Edward VI were simply combined, thus inviting divergent interpretations. The number of executions was as great under Elizabeth as under Mary, but the reign of Mary was five years and that of Elizabeth 45, and the cause was chiefly political since the pope had the indiscretion to excommunicate Elizabeth and absolve her subjects from allegiance, so that obedience to the pope meant treason to the queen.

John Knox. In Scotland the Reformation is associated with the name of John Knox, who declared that one celebration of the mass is worse than a cup of poison. He faced the very real threat that Mary, Queen of Scots (formerly queen of France) would do for Scotland what Mary Tudor had done for England. Therefore Knox defied her to her face in matter of religion and, though a commoner, addressed her as if he were all Scotland. He very nearly was, because in the period prior to 1560 many an obscure evangelist had converted the lowlands largely to the religion of John Calvin. The church had been given a Presbyterian structure, culminating in a General Assembly, which had actually as great and perhaps a greater in-

John Knox's opposition to Mary, Queen of Scots fluence than the Parliament. Because of her follies, and very probably her crimes (complicity in the murder of her husband), Mary had to seek asylum in England. There she became the focus of plots on the life of Elizabeth until Parliament decreed her execution. Presbyterianism came to be established in Scotland, and this very fact alone made possible the union of Scotland with England. Union of Protestant England with a Catholic Scotland would have been unthinkable.

Knox is frequently reproached for his intolerance in regarding one celebration of the mass as worse than a cup of poison, but one must remember that the year 1560 marked the peak of polarization between the confessions. Similar intolerance had been mounting at Rome. Paul III, after an abortive attempt at reform, had introduced the Roman Inquisition in 1542. His successor, Paul IV, placed everything that Erasmus had ever written on the Index. The Council of Trent began its sittings in 1545, introducing rigidity in dogma and austerity in morals. The Protestant views of justification by faith alone, the Lord's Supper, and the propriety of clerical marriage were sharply rejected. All deviation within the Catholic fold was rigidly suppressed. When Carranza, the archbishop of Toledo, returned to Spain in 1559, after assisting Mary in the restoration of Catholicism in England, he arrived in time for the last great ***auto-da-fe'*** of the Lutherans. Himself under suspicion for ideas no more heretical than those of Erasmus, he was incarcerated for 17 years in the prison of the Inquisition. The liberal Cardinal Giovanni Morone was imprisoned during the pontificate by Paul IV, and under Pius V Pietro Carnesecchi, an Erasmian and onetime secretary of Clement VII, was burned in Rome. John Knox and Pope Pius V represent the acme of divergence between the confessions.

Council of Trent

Developments in the English Reformation. The real parallel in England to the continental Reformation followed only in the 17th century, for not until then did the populace come to be passionately concerned over the religious issues and broke into a welter of sects, many of them reviving various tenets of the Anabaptists. The English Baptists rejected infant Baptism; the Quakers rejected war and the oath; the Fifth Monarchy Men incited armed revolution, like the Anabaptists at Miinster; the Congregationalists and the Baptists demanded the "gathered" church consisting only of convinced believers. All of them, including the Presbyterians, objected to the episcopal structure of the Anglican Church. The Unitarians grew out of the Presbyterians, at first more as a revolt against the Calvinist doctrine of the depravity of man and the severity of God than over the doctrine of the Trinity, which was questioned only later. The Quakers went beyond all in rejecting a professional clergy and the outward sacraments. All of them rejected the authority of government if it contravened their principles, but the Presbyterians and the Congregationalists were willing to set up theocratic commonwealths if the government were favourable. The Quakers did so also in Pennsylvania, though without constraint of Dissenters. Such diversity raised acutely the problem of religious pluralism and contributed to the outbreak of the civil war in which the demand for independence in religion coincided with the curtailment of the monarchy in economics and politics. A further concomitant resulted from the union of Scotland and England under James I. As already remarked, the union would have been impossible if one of the parties had been Catholic, the other Protestant, but even varieties of Protestantism could fight before agreeing to tolerate diversity. In the sequel the political union did not require an ecclesiastical union.

The theory of revolution ran the same course in England as on the Continent. The king was at first said to be controlled by evil counsellors, then that he might be resisted by the inferior magistrate, in this case Parliament, finally that Parliament was a supreme court of judicature with authority to take the life of the king. After the execution of Charles I, successor to James I, Cromwell set up the establishment of a national religion rather than of a single confession. His establishment rested on three pillars, Congregational, Presbyterian, and Baptist. He excluded the Catholics and Unitarians on doctrinal grounds and the Quakers for refusal to swear. His scheme anticipated the pattern of the United States where, despite the separation of church and state, the government covets the sanction of three religious bodies, Jews, Catholics, and Protestants. The Puritan revolution was not as bitter as the wars of religion on the Continent because all of the contestants were Protestant and because the death penalty for heresy had been dropped early in the 17th century. The Restoration under Charles II began with a promise of toleration, which broke down because the sects were not willing to be tolerated if the Catholics were included. The definitive edict came only with the "Glorious Revolution" of William and Mary in 1688. The ideal of the comprehension of all Englishmen in one church was abandoned in favour of the establishment of the Church of England and the toleration of Dissenters, but Catholic emancipation did not come until 1829.

THE EXPANSION OF THE REFORMATION: SCANDINAVIA; THE BALTIC STATES; EASTERN, CENTRAL, AND SOUTHERN EUROPE

By the middle of the 16th century Lutheranism was dominant in northern Europe. Wiirttemberg, after the restoration of Duke Ulrich, adopted the reform in 1534. The outstanding Reformer was Johannes Brenz and the great centre Tiibingen. Brandenburg, with Berlin as its capital, embraced the reform in 1539. In that same year ducal Saxony, until then vehemently Catholic, changed sides. Elisabeth of Braunschweig, also in that year, became a convert, but only after long turbulence did her faith prevail in the land. Very significant for the north as a whole was the stand taken by Albert of Prussia, married to a Danish wife and a member of the Polish Diet. He secularized the Teutonic Knights and in 1525 acknowledged himself a Lutheran. In the Scandinavian lands Denmark toyed with Lutheranism as early as the 1520s, but not until 1539 was the Danish Church established on a national basis with the king as the head and the clergy as leaders in matters of faith. Norway followed Denmark. The Diet of Västerås officially declared what had for some time been true, namely, that Sweden was an evangelical state. The outstanding Swedish Reformers were the brothers Olaus and Laurentius Petri. Finland, under Swedish rule, followed suit. The Reformer there was Mikael Agricola, called "the father of written Finnish." The Baltic states of Livonia and Estonia were officially Lutheran in 1554. Subsequently ravished by the Russians, portions of these lands united with Sweden, Denmark, and Poland. Lutheranism survived. Toward the east, Austria under the Habsburgs could enjoy no state support for the evangelical movement, which nevertheless gained adherents. In Moravia, as noted, the Hutterites established their colonies under tolerant magnates.

Eastern Europe offered a seedbed for even more radical varieties of Protestantism, because kings were weak, nobles strong, and cities few, and because religious pluralism had long existed: in a given political area one section might be Roman Catholic, another Orthodox of one variety or another, and in Bohemia the Hussites had gained toleration. Poland acquired a large German Lutheran population when the Danzig area came under Polish control, and a considerable contingent of the Bohemian Brethren migrated to Poland when the Habsburg ruler attempted their extermination. Several of the Polish noblemen adopted their pacifism and would wear only wooden swords. To Poland also flocked the Italian anti-Trinitarians, having been granted an asylum, perhaps merely because they were Italian, by the Italian queen of Poland, Bona Sforza. Named Socinians from their leader Faustus Socinus, they flourished until dissipated by the Counter-Reformation. Much more extensive was the Calvinist influx not only into Poland but into the whole of eastern Europe. This variety of Protestantism appealed to those of non-German stock because it was not German and no longer markedly French, as well as because of its revolutionary temper and republican sentiments. The Compact of Warsaw in 1573 called the ***Pax Dissidentium*** ("The Peace of Those Who Differ") granted toleration to Catholics, Lutherans, Calvinists, and Bohemian Brethren, but not to the Socinians.

In Hungary, the Turkish victory at the Battle of Mohács in 1526 brought about a division of the land into three sections, the northwest ruled by the Habsburg Ferdinand, the eastern province of Transylvania under Zfipolya, and the area of Buda under the Turk. Even before this date Lutheranism had made inroads not only in the German but also in the Magyar sections. Subsequently Calvinism made even greater gains. The anti-Trinitarians found a permanent locus in Transylvania. The weakness of the government and the diversity of religion in this whole area made for a large degree of toleration.

The Inquisition

The Reformation notoriously gained no lasting hold in Spain and Italy. In Spain the main reason grew out of the conflicts of the previous century when the Christians were striving to achieve political, cultural, and religious unification by converting or expelling the unbelievers, the Jews and the Moors. The Inquisition was introduced in 1482 to root out all remnants of Jewish practices among the Marranos, the Jewish converts to Christianity. The non-Christian Jews were expelled in 1492. Then Granada fell and the same process was applied to the Moriscos, the Moorish converts, and the unconverted Moors, after a century, also were expelled. The process had thus far been successful so that early in the 16th century the pressures were relaxed and Spain enjoyed a decade of Erasmian liberalism in the 1520s, but with the infiltration of Lutheranism the machinery of repression again was brought into force.

Italy had no political unity. There were five great city-states, Florence, Milan, Venice, Naples, and Rome. Roman spiritual overlordship alone gave a semblance of unity, and Rome was never servilely followed. In the late Middle Ages sectarian and heretical movements had proliferated in Italy. One by one they had been crushed, and the Italians may well have felt that such rebellions were futile. Another factor was that the friars preached moral rather than doctrinal reform, as Luther had done. A further consideration was that the new monastic orders, the Capuchins, Theatines, and Jesuits, gained papal favour and became a mighty force in counteracting Protestant infiltration, which nevertheless did take place. Venice was a centre, with its branch house of the banking family of Fugger who brought in Lutheranism. Lucca was a centre At Naples the Spanish mystic Valdés, though not a Protestant, expounded a piety of the type of the liberal Catholic reform, and some of his followers were attracted to the movements coming from beyond the Alps. Calvinism gained a hold. But the Roman Inquisition, as above noted, was established in 1542, and those with Protestant leanings either made cloisters of their own hearts in silence, or went to the stake, or crossed the mountains into permanent exile. The most radical theological views of the Reformation were those propounded by the Spanish and Italian anti-Trinitarians.

III. Religious liberty

POLITICAL

Legal recognition of non-Roman Catholics

The question of religious liberty as a political problem has already received frequent notice. Toleration was achieved at first only by groups strong enough to imperil the peace if denied liberty of faith. The first to gain official recognition of the right were the Zwinglians of Ziirich, who after the second war of Kappel in 1531 were not to be molested but were not to expand. The Lutherans gained recognition by the Peace of Augsburg in 1555, the Huguenots by the Edict of Nantes in 1598, the English sects in 1688. In Holland legal restraints on Catholics and Anabaptists were not removed until late in the 17th century, but earlier practice was lax and a commercial city like Amsterdam did not allow dogmatic rigour to hamper trade. On the Continent, the Peace of Westphalia in 1648 was constructed on the same territorial principle as the Peace of Augsburg but included the Calvinists and revised the lines to correspond to what had come to be a fairly stabilized distribution of the confessions.

INDIVIDUAL

Servetus and religious dissent

Individual religious liberty came much more slowly. The political and economic arguments in favour of liberty were chiefly valid only for groups, but the pleas based on the nature of faith cut athwart all persecution. The great toleration controversy in Protestantism was precipitated by the burning at Geneva of Michael Servetus for anti-Trinitarianism and the rejection of infant Baptism. The following year (1554) Sebastian Castellio issued pseudonymously a tract ***Concerning Heretics*** in which he employed the rationalist argument that the doctrines for the rejection of which men were being put to death must be uncertain for the very reason that they were controverted. This left open the possibility of constraint with respect to those doctrines which in that day were assumed to be certain, such as the existence of God and immortality. A further point was that even if these controverted doctrines were true, they need not be believed in order to be saved. A distinction was thus drawn between the fundamentals and the nonessentials. Castellio, following Erasmus, was much more concerned to enumerate the nonessentials than to formulate the essentials. Another proponent of religious liberty, the Italian Jacobus Acontius, said that those doctrines are alone necessary to be believed which are stated in Scripture to be such. He could find only justification by faith and belief on the Lord Jesus. Excommunication should be the only penalty. His two points were taken to exclude the Catholics on justification and the anti-Trinitarians on the Lord Jesus. One recalls that these were the two excluded by Cromwell. Other arguments cut deeper, being rooted in the spiritual rather than the rational. Religious faith is inward and cannot be created by coercion. All that force can do is to provoke revolution in a group or to induce hypocrisy in an individual. Servetus was executed because he told the truth, that is, what he believed to be the truth. Had he lied and denied his conviction he could have been saved. There also were ethical considerations. God is better pleased by deeds than by creeds. In the case of groups, the argument ran that the toleration of error is better than the devastation of the land. Such was the line taken by the party in France called the ***Politiques.***

DOMESTIC

Notice has already been taken of the effects of the Reformation in economics and politics, but in no area were they so great as in the domestic sphere. Since marriage was no longer a sacrament it became a civil contract and special matrimonial tribunals were created. The degrees of consanguinity impeding marriage were repealed. The church in the Middle Ages had at one time excluded marriages to the seventh degree. This was reduced to the fourth in canon law. The Reformers left only incest. Spiritual impediments to marriage were removed. Hitherto, godparents could not marry their children to godchildren, that is to say, without a dispensation. Differences in religion led to a relaxation of the rule that divorce was to be allowed only for adultery. Luther held to this but the Calvinists ruled that if one partner refused to follow the other into exile for religion the bond was dissolved.

Revolution in the concept of marriage

The deepest revolution may well have been with respect to the concept of marriage, because monasticism died out and the home came to be the area ***par excellence*** for the practice of the gentler virtues. The church had solved the discrepancy between man's usual behaviour in society and the precepts of the gospel about nonresistance, love of enemies, sharing of goods, and the like by a vocational division, relegating these injunctions primarily to the monasteries. When these were gone the home may not have become actually more loving but it was exalted as the area for self-effacement, mutual forbearance, and the sharing of goods. An emphasis was placed also, notably in Calvinism, not so much on marriage as a remedy for sin or a device for progeny, but rather as a partnership in the vineyard of the Lord, in the rearing of children in the faith and the advancement of God's Kingdom.

BIBLIOGRAPHY

Brief popular histories: OWEN CHADWICK, ***The Reformation*** (1964); G.R. ELTON, ***Reformation Europe, 1517–1599*** **(1964)**, especially strong on political aspects; R.H. BAINTON, ***The Reformation of the Sixteenth Century*** (1952).

More extensive treatments: H.J. GRIMM, ***The Reformation***

Era, 1500–1650 (1954), on the Protestant and Catholic reform movements with attention to the social and political impacts; *The New Cambridge Modern History*, vol. 2, *The Reformation 1520–1559*, vol. 3, *The Counter-Reformation and Price Revolution 1559–1610* (1962, 1968), detailed accounts by specialists; E.G. LEONARD, *Histoire générale du protestantisme*, 2 vol. (1961, 1964; Eng. trans., *A History of Protestantism*, vol. 1, *The Reformation*, vol. 2, *The Establishment*, 1966–67), an account with extensive bibliographical references.

The Reformation in England and Scotland: T.M. PARKER, *The English Reformation to 1558* (1950); F.M. POWICKE, *The Reformation in England* (1941); G. CONSTANT *The Reformation in England*, 2 vol. (Eng. trans., 1934, 1941), a liberal Catholic account; A.G. DICKENS, *The English Reformation* (1964), based on new documentary material; E.G. RUPP, *Studies in the Making of the English Protestant Tradition* (1947); G. DONALDSON, *The Scottish Reformation* (1960), thoroughly documented.

The Reformation in France: J.T. MCNEILL, *The History and Character of Calvinism* (1954), an excellent survey; R.M. KINGDON, *Geneva and the Coming of the Wars of Religion in France, 1555–1563* (1956), from manuscript sources.

The radical Reformation and religious liberty: G.H. WILLIAMS, *The Radical Reformation* (1962); J. LECLER, *Toleration and the Reformation*, 2 vol. (Eng. trans., 1960).

Biographies: E.G. RUPP, *The Righteousness of God: Luther Studies* (1953); R.H. BAINTON, *Here I Stand: A Life of Martin Luther* (1950); H. BORNKAMM, *Luther's World of Thought* (Eng. trans., 1958); O. FARNER, *Zwingli the Reformer: His Life and Work* (Eng. trans., 1968); W. WALKER, *John Calvin: The Organiser of Reformed Protestantism, 1509–1564* (1906, reprinted 1969); A.F. POLLARD, *Thomas Cranmer and the English Reformation, 1489–1556* (1904); N.L. ROELKER, *Queen of Navarre: Jeanne d'Albret, 1528–1572* (1968), a detailed biography of the mother of Henry IV; R.H. BAINTON, *The Travail of Religious Liberty* (1951), nine biographies illustrating persecution and liberty from Torquemada to John Locke.

(R.H.B.)

Reformed and Presbyterian Churches

Sharing a common origin in the 16th-century Reformation in Switzerland, Reformed and Presbyterian churches are one of the major representative groups of classical Protestantism. Reformed is the more inclusive term identifying those Reformation churches that are regarded Calvinistic in doctrine, rather than Lutheran, the other major Protestant doctrinal system that arose in the 16th century (in Germany). The term presbyterian designates a collegiate type of church government by pastors and lay leaders who are called elders, or presbyters, from the New Testament term *presbyteroi.* Presbyters rule through a series of representative bodies from the local congregation to area and national organizations, commonly termed sessions, presbyteries, synods, and assemblies. Not all Reformed churches are presbyterian in their form of government, but the churches that use the name Presbyterian are a part of the Reformed tradition.

Meaning of Presbyterian and Reformed

In 1970 over 35,221,000 people identified themselves as being affiliated with Reformed and Presbyterian churches; of these, 15,960,000 were in Europe; 8,441,000 in the United States and Canada; 4,402,000 in Africa; 4,203,000 in Asia; 1,589,000 in Australasia; and 626,000 in Latin America and the West Indies. These figures include neither China nor Reformed groups that are not members of the World Alliance of Reformed Churches, which has about 130 member denominations.

A slogan for the Lutheran Reformation was "by faith alone." Reformed Christians added to that principle "to God alone the glory." Reformed Christians thus emphasized that God's word alone and no mere human opinion should be the norm for faith. Apprehension that something less than God would take the place of God determined Reformed attitudes toward church government and worship, the furnishing and design of church buildings, and secular authority.

HISTORY

The Reformation (1517–1660). The Swiss Reformation leaders were Renaissance Humanists who were inspired by Desiderius Erasmus, a Dutch Humanist, to regard the Greek New Testament as the basis for Christian faith and action. They shared his concern to reform the life of the church and make the Bible available to all.

Zwingli in Ziirich. Erasmian ideas were put into practice in Ziirich in 1519 by Huldrych Zwingli. Later, after Zwingli recovered from the plague with a sense of God's providence in his life, he confided to a friend that the church that had come into life by means of blood could only be resurrected through the spilling of blood. This sentiment went beyond the views of Erasmus, but it also appeared in the views of later Reformers.

A difference of spirit with the Lutherans became apparent in 1525, when Zwingli published a letter of Cornelius Hoen, a Dutch Humanist, regarding the Lord's Supper, interpreting the "is" in the statement "This is my body" to mean "signifies." John Oecolampadius, the Protestant leader of Basel, joined Zwingli in a pamphlet controversy with Luther over the nature of the sacrament. The controversy between Luther and Zwingli led to the Colloquy of Marburg of 1529, in which Philip the Magnanimous, landgrave of Hesse, sought to bring about a consensus between Saxon and Swiss theologians. They reached agreement on 14 out of 15 key points, but, on the issue as to whether the body and blood of Christ are physically present in the bread and wine, they disagreed. The divergence of the Reformed and Lutheran Protestant traditions stems from this issue. When Zwingli was killed in a war between Roman Catholic and Reformed forces at the Battle of Kappel in 1531 and Oecolampadius died shortly thereafter, Luther considered his own position to be vindicated.

Bucer in Strassburg (Strasbourg) and Calvin in Strassburg and Geneva. In the 1530s Martin Bucer, the Reformation leader in Strassburg, saw a need for lay elders chosen from the congregation to assist the pastor in church discipline. Oecolampadius had earlier identified such officials with the presbyters of the New Testament congregations. On the basis of this identification, the presbyterian form of church government was initiated on a congregational level. In debates with Anabaptists, radical Protestant Reformers who stressed a rigid congregational discipline, Bucer conceded that, unless discipline could be maintained, there was no real community.

The civic leaders of Strassburg resisted Bucer's disciplinary measures in 1534. Later, a similar reaction took place in Geneva against John Calvin, who had become the Reformation leader in that city. Forced to leave Geneva in 1538, Calvin became pastor to the French-speaking congregation in Strassburg. There Calvin expanded his famous *Institutes of the Christian Religion,* the first edition of which had been published in Basel in 1536. He also experimented with new forms of worship and discipline. When he returned to Geneva at the invitation of the civil and religious leaders in 1541, Calvin brought with him a developed idea of doctrine, worship, and discipline, which he sought to realize in Geneva. Though his model for the Christian Church and commonwealth was opposed, by 1555 the Reformation in Geneva had been formulated according to Calvinistic principles. Geneva became a centre for Protestant refugees and developed an international training school for pastors. John Knox of Scotland was representative of many who saw in Geneva "the most perfect school of Christ that ever was in earth since the days of Apostles."

Calvin's doctrinal and discipline accomplishments

Calvin sought to bring about a consensus between Lutherans and Zwinglians on the matter of the Lord's Supper. He found value in both of their positions, though he favoured Luther, who, on occasion, spoke favourably about Calvin. When Calvin's friend Philipp Melanchthon lost influence among Lutherans, however, hope of uniting the two traditions diminished. Calvin and Heinrich Bullinger, Zwingli's successor in Ziirich, reached an understanding in the Ziirich Consensus (Latin Consensus Tigurinus) of 1549. The Second Helvetic Confession of 1566, written by Bullinger, became the most widely held Reformed confession of faith, just as Calvin's final edition of the *Institutes* in 1559 became the basic theological text for Reformed Christians throughout the world.

The Reformation in France. In 1555 Calvin turned his attention beyond Geneva. He sent Geneva-trained ministers into the Piedmont, as well as into France, to contact the remnants of the Waldenses, a medieval reform group active mainly in Italy. He also sent pastors trained in Geneva into France, where many new congregations were formed. By 1562 the Huguenots, as French Protestants came to be known, claimed 2,150 congregations with a membership estimated at 3,000,000. After discussions between Catholic and Reformed representatives at the Colloquy of Poissy in 1561, the Huguenots expected toleration from the queen regent Catherine de Médicis. The massacre of a Protestant congregation by the Duc de Guise, however, led the Huguenots to take up arms. The ensuing religious wars reached a climax with the St. Bartholomew's Day Massacre of August 23–24, 1572, when thousands of Huguenots were killed. As a result of the massacre, Philip du Plessis Mornay, a French Protestant diplomat and writer, developed a theory of resistance. He held that, when a monarch persecutes the true religion, he has become a rebel against God, and his people are absolved from loyalty to him.

The Edict of Nantes

A third party, known as the Politiques, included representatives from both Catholics and Huguenots. They believed religious uniformity to be desirable, but, if this was not practical, the welfare of the state should be put above religious interests and the religious minority tolerated. This policy took effect on the accession of King Henry IV, who on April 13, 1598, proclaimed the Edict of Nantes. This edict permitted the profession of Protestantism in many towns and castles south of the Loire River.

Peace ended when Huguenots rose to protest the restoration of Roman Catholicism in the territory of Béarn. Resistance ended in 1628 with the fall of La Rochelle. The Huguenots were no longer a military power.

The Huguenots made a lasting contribution to the development of Presbyterianism. The presbyterian form of church government was organized for the first time on a national basis in France. The national synod of the Reformed Church in France in 1559 planned consistories on the congregational level, colloquies on the regional level, and provincial synods, thus setting the pattern for national Presbyterian organizations in other countries.

The Reformation in the Rhineland. The Reformed cause was advanced in Heidelberg when Frederick III, elector of the Palatinate, took as advisers Geneva-trained Caspar Olevianus (1536–87) and Zachary Ursinus (1534–83). Heidelberg became a flourishing centre of Reformed theology, and there the Heidelberg Catechism of 1562, one of the great international Reformed confessional documents, was produced.

In 1568 the Erastian controversy broke out in Heidelberg, showing that differences over discipline still existed within Reformed churches. A young English student, George Withers, in defending his doctoral theses at Heidelberg, maintained

> that in every church . . . there ought to be a government or discipline observed, whereby the ministry, in conjunction with elders . . . should have the right and authority to excommunicate any vicious liver, even princes themselves.

Thomas Erastus, a Swiss physician living in Heidelberg, insisted that law and order should be maintained by the state and not the church. From his arguments the term Erastianism was coined.

The religious and political struggle in the Netherlands. In 1559, when William I of Orange decided to resist Spanish power, a number of Reformed congregations appeared in the Netherlands. These congregations soon became associated with the cause of Dutch nationalism, especially during the struggle for independence from Spain between 1566 and 1578. Because of this association, the Reformed Church became established once freedom for the Netherlands had been obtained. The early national synods of the Reformed Church of the Netherlands, resembling those of France, adopted the Belgic Confession and the Heidelberg Catechism.

The Arminian controversy

Controversy engulfed the Reformed Church of the Netherlands, however, as a result of the writings of Jacobus Arminius (died 1609), a professor of theology at the University of Leiden who objected to the doctrine of predestination taught by William Perkins, an English theologian. Arminius held that God's offer of grace to man was universal rather than relegated to the elect few and that man was free to accept or reject God's offer. After Arminius' death, men of his party presented a ***Remonstrance*** to the State in 1610, requesting a synod to settle the controversy. The debate that followed polarized elements on all levels of Dutch society until a-synod met at Dort from November 13, 1618, to May 28, 1619. Representatives came from the churches of England, Scotland, and eight of the Reformed churches of Germany and Switzerland. The decision went against the Remonstrants, but, even though forced out of their positions in the official church, the Remonstrants continued as an independent church. The Synod of Dort became the touchstone of Reformed orthodoxy in the 18th century. In Holland the primary spokesman for that orthodoxy was Gisbertus Voetius (died 1676), who did his best to stand firm against every new idea that appeared in the 17th century. He was an important link between English Puritanism and, later, continental Pietism (see also PURITANISM; PIETISM).

In spite of controversy, the 17th century was a period of the flowering of the Dutch republic in culture and in international trade. This commercial expansion makes plausible the Weber–Tawney thesis (formulated by Max Weber, a German sociologist, and Richard Tawney, an English economist) that Calvinism has given rise to the spirit of capitalism.

Reformed influence on the English Reformation. England received its major influx of Reformed influence during the reign of Edward VI (1547–53). Martin Bucer became professor at Cambridge; Zurich-trained John Hooper was named bishop; Dutch and French refugees established the Church of the Strangers along Reformed lines in London; and constant counsel was sought and received by the English from Heinrich Bullinger, Philipp Melanchthon, and John Calvin.

The rise of the Puritan controversy

With the death of Edward and the restoration of Catholicism under Mary I, exiles from England took refuge in Zurich, Basel, Strassburg, and Geneva. In 1558—after Queen Elizabeth I assumed the throne—the exiles returned, with intentions of reforming the Church of England beyond the desires of the Queen. Yet, when the Puritan controversy over the extent of the reform of the Church of England broke out in earnest in the 1570s, men of Reformed persuasion did not necessarily take the Puritan side, which intended to rid England of all vestiges of "popery." The son of John Knox at Cambridge and the grandson of Huldrych Zwingli at Oxford, for example, opposed the Puritans. Bishops siding with Elizabeth considered themselves Reformed theologians, quoting Calvin's works that favoured retaining bishops in the church. Heinrich Bullinger of Ziirich supported the English bishops, many of whom were his personal friends.

The Puritans, during the last years of Elizabeth's reign, were subdued, but they never lost hope of further reformation. They wanted to modify the power of bishops in favour of presbyterian government and the enforcement of discipline in local congregations. Their opportunity came with the outbreak of civil war in 1642 and the calling of the Westminster Assembly of divines (Puritan clergymen) in 1643.

Shocked by the radical developments in the Civil War, Presbyterians and Erastians worked toward the restoration of Charles II as king of England in 1660. They hoped for a restored episcopate that would be broad enough to comprehend the "tender consciences" of Presbyterians and Independents. This did not occur, however, and the latter continued to exist only as Nonconformist sects.

The Reformation in Scotland. Scotland's Reformation dates from the return of John Knox from Geneva in May 1559. The following year the Scottish Parliament adopted the Scots Confession, which was drafted in part by Knox, but it did not ratify the ***Book of Discipline,*** which applied Genevan ideals to a national system. The lay lords accepted those parts of the ***Book of Discipline*** designed to improve the behaviour of the populace but not those limiting their own power. Knox reshaped the medieval episcopal dioceses into ten districts, each with a superintending pastor, but he did not envision a developed Presbyterianism.

Scottish Presbyterianism was promoted by Andrew Melville (died 1622), a theologian and educational reformer who had taught in Geneva. The episcopal system had been restored in Scotland by 1572. In opposing episcopacy backed by the civil powers, Melville fought for church authority parallel to but not subject to the state. Between 1576 and 1584 he made progress with his proposals, but King James VI (James 1 of England) exiled him in 1607. When James became king of England as well as of Scotland, he sought conformity between the churches of the two countries. Episcopacy was re-established in Scotland but with the apparatus of presbytery.

Even under an episcopal form of church government, the Reformed Church of Scotland followed the liturgical ideas of John Knox. Charles I continued the policy of James in using the episcopacy to further royal power. His misunderstanding of the importance of the Reformed liturgy to the Scottish church precipitated the overthrow of episcopacy in Scotland. A prayer book, which many considered a "popish book," was introduced in Edinburgh on June 23, 1637, resulting in an uprising that Charles I was unable to subdue. He then had to concede a presbyterian government for Scotland.

When English Puritans sought Scottish aid in 1643 during the Civil War, the price of the Scots was the adoption of the Solemn League and Covenant, committing English leadership to a church government such as that which existed in Scotland. The beheading of Charles I by the English Independents, who had purged Presbyterians from the government, shocked the Scots, but they could do little against the forces of Cromwell, the leader of the Commonwealth. With the restoration of the kingdom in 1660, Scotland became episcopal once again. Those who remained loyal to the Solemn League and Covenant faced what was called "the killing times," during which the Covenanters were virtually wiped out.

Reformed Churches in Hungary, Bohemia, and Poland. Reformed Protestantism gained headway in Hungary in the 1550s. In 1566 and 1567 the Geneva Catechism and the Second Helvetic Confession were adopted by Hungarian synods, and in 1576 the government of the Reformed Church of Hungary emerged, with superintending bishops chosen by the church councils of pastors and elders. István (Stephen) Bocskay, prince of Transylvania, secured recognition of the rights of the Reformed Church in both Habsburg and Turkish territories in 1606.

Development of Czech Reformed churches

In Bohemia two dissenting church bodies (the Utraquists and Unity of Brethren) remained from the days of Jan Hus, the 15th-century Reformer burned at the stake at the Council of Constance in 1415. In aligning themselves with the views of the 16th-century Reformation, the Utraquists favoured Lutheranism, and the Unity of Brethren preferred the Reformed position. As early as 1540 the Czech Brethren had been in contact with Bucer and Calvin in Strassburg. In 1575 the Brethren and the Utraquists adopted a Bohemian Confession, and the Habsburg emperor Maximilian II guaranteed religious liberty to those who held this confession. Because Calvinist leaders in 1617 feared a revocation of the guarantee, they decided upon resistance. When they threw three agents of Archduke Ferdinand out of a window in Prague on May 23, 1618, they precipitated the Thirty Years' War. This war was absolutely devastating to Czech Protestantism, which was able to survive only in exile or underground.

Poland made a contribution to the Reformed churches of western Europe through the nobleman and Humanist Jan Łaski the Younger (died 1560), who was with Erasmus in Basel in 1523. Refusing a bishopric in Poland, he became superintendent of the church in Emden in 1542 and developed the Frisian Church on Reformed principles. In England, from 1549 to 1553, he was superintendent of the Church of the Strangers in London and in 1556 became a superintendent of the Polish Reformed Church, supervising the translation of the Bible into Polish.

Three Protestant groups attempted to unite in Poland. The Reformed and the Czech Brethren held a joint synod in 1555, and in 1570 the Lutherans joined these in the adoption of the Consensus of Sandomir. As a result of the Counter-Reformation, however, the Reformed Church in Poland was reduced to a small sect in the early 17th century.

The Reformed Confessions of the 16th and early 17th century. Because Reformed churchmen believed that new situations required new responses, a large number of manifestos of the Reformed faith were written. A harmony of confessions, prepared in 1581, indicated that there was agreement both among Reformed churches and with the Lutheran Confession of Augsburg. Some of these confessions were theses for debate, such as Zwingli's *Sixty-seven Articles* of 1523. Others, such as the Zürich Consensus of 1549, sought unity between groups on controversial doctrines. Still others, such as the Heidelberg Catechism of 1563, were intended for the nurture of people through teaching and preaching. The very names of the Geneva, Helvetic, French, Belgic, and Scots confessions indicate a relationship of Reformed confessions to the rising sense of nationhood in the 16th century. Many of these national confessions took on international significance.

Perhaps the most elaborate statement was the Westminster Confession of faith of 1648, produced by the Assembly of divines meeting under the authority of the English Parliament. It became a standard of the English-speaking Presbyterian churches and, on a modified basis, of Congregational churches.

After the Reformation in Europe (1660 to the present). ***France.*** On October 18, 1685, the Edict of Nantes was revoked, and 250,000 Huguenots emigrated from France. After the suppression of the Camisard (French Protestant peasant) revolt in 1715, Protestantism in France seemed at an end. For the survivors, toleration finally came in 1787 and with the Revolution equality under the law. Napoleon placed Reformed congregations under state control, with pastors on state salary.

Revocation of the Edict of Nantes

A national synod did not meet again until 1848. At that time the free Evangelical Synod was organized, separating from the state-recognized church over the issue of state support. In 1905 state support of the old synod was withdrawn, and the two synods were united in 1938. The Reformed Church in Alsace–Lorraine has a different history and remains separate from the Reformed Church of France. Recently, French Reformed Christians have played a notable role in the development of the World Council of Churches, in liturgical and theological renewal, in relating the church to technology and urbanization, and in Catholic–Protestant and Communist–Christian dialogue. Outside of French-speaking Switzerland, they remain the largest Protestant group in the Latin countries of Europe, each of which has a small Reformed Church.

Reformed churches in Germany. The Peace of Westphalia in 1648, ending the Thirty Years' War, established the legality of Reformed churches in German states, according to the pleasure of the ruling prince. The Reformed Christians in the Palatinate faced a systematic attempt at their destruction at the beginning of the 18th century, and many fled to the Netherlands and to America. The Hohenzollern elector of Brandenburg converted to Calvinism in 1609; thereafter, Reformed congregations developed in Brandenburg, and Reformed territories came under the control of its electors. Refugees from France and the Palatinate were invited to settle in Brandenburg.

The Hohenzollerns worked toward union between Reformed and Lutheran churches, and King Frederick William III of Prussia in 1817 proposed completion of this union. The Reformed theologian Friedrich Schleiermacher led ministers in support of the Prussian Union. The Evangelical Reformed Church of Northwest Germany, a result of the Prussian Union, is the largest Reformed Church in Germany today and is a constituent part of the Evangelical Church in Germany.

In 1884 a Reformed League was organized in Germany to preserve the Reformed heritage. Its synod, held at

Barmen in 1934, helped to form the Confessing Church in Germany. The Barmen Declaration, inspired by the theology of Karl Barth, a 20th-century Reformed theologian, became the manifesto for Reformed, Lutheran, Union, and Free churches in opposition to Nazi corruption of the gospel. Though after World War II the Confessing Church ceased to exist, the Reformed League from which it originated is still active.

The Synod of Barmen

Reformed churches in England, Scotland, Holland, and Switzerland. The Glorious Revolution of 1688–89, which expelled the Roman Catholic sovereign James II, left English Presbyterians and Independents with limited toleration outside the state (Episcopal) church. The Toleration Act of 1689 granted protection to all Protestants who accepted the doctrine of the Trinity. In Scotland, however, the state church became Presbyterian. Because of subsequent state interference in the appointment of pastors, several secessionist churches arose in Britain. The Evangelical awakening, a religious revival of the 18th century, brought about a renewed sense of purpose among Christians both within and without the state churches. Membership of churches increased, and new denominations came into being. The free-church spirit changed the Reformed emphasis from concentration upon parish discipline to the support of voluntary societies dedicated to the uplift of the community: to preserve the sabbath, to suppress vice, to abolish slavery, and to work for moral reform.

Evangelicalism in The Netherlands, Switzerland, and Scotland had popular support but could not move ecclesiastical bureaucracies. The resulting conflicts within the churches caused all three countries to experience separatist movements that eventually formed their own ecclesiastical organizations. In The Netherlands, conservative separatist elements formed the Christian Reformed Church in 1869. Migrants to the United States from various Dutch separatist groups united to form the Christian Reformed Church in 1890. A Free Church—based on evangelical principles—was constituted in 1843 in Scotland. Within 15 years it had 1,000 ministers and 800 churches. In Switzerland, in 1845, a Free Church Synod was organized when the state forbade pastors to hold revival meetings.

Reunions of separated churches have occurred in the 20th century, but there has also been a marked decline of influence of these churches among the masses. A theological revival, led by the Swiss theologians Karl Barth and Emil Brunner, enlivened Reformed churches between 1920 and 1960.

Churches elsewhere in Europe. The Reformed Church of Hungary remains the largest Reformed group in eastern Europe. It survived an attempt at extermination under the Habsburgs; toleration came in 1781, and in 1884 equality under the law. Divisions of land after the world wars resulted in large Hungarian-speaking Reformed churches in Czechoslovakia, Yugoslavia, Romania, and the Soviet Union.

In Czechoslovakia some toleration for Protestants came in 1781, and an Evangelical Church of the Czech Brethren was reconstituted on the basis of both Lutheran and Reformed confessions.

The surviving Reformed community in Poland has nine congregations. There are five congregations in Lithuania and one in Latvia.

Reformed and Presbyterian churches in the United States. *In the colonial period.* Reformed influences in what became the 13 original United States were quite noticeable. Persons of Reformed background and Reformed institutions were prominent in directing the course of the colonies in matters of religion, politics, and other areas. Alexander Whitaker, who established churches in Virginia, was the son of a famous English Reformed theologian. On Manhattan Island a Dutch Reformed Church was organized in 1628. Elder William Brewster, of the Plymouth Colony of 1620, used the writings of the English Presbyterian Thomas Cartwright as his guide in church government, and the Massachusetts Bay Colony of 1628 was an attempt at a new model Reformed Church and commonwealth. During the 17th century Waldensian refugees came to Staten Island, and Huguenots settled in New York and New England. These were followed by a Scots-Irish migration that ranged throughout the colonies and by German Reformed refugees from the Palatinate.

Influence of Reformed churches

Varieties of Reformed and Presbyterian churches whose lives were based on the experiences of these various colonists contributed to developing a diversity of religion in the American colonies.

The 18th-century Great Awakening—led by such Calvinist preachers as Jonathan Edwards, Theodore Frelinghuysen, George Whitefield, and Gilbert Tennent—encouraged an evangelical Christianity that was often at odds with traditional church attitudes, and the revival-seasoned leaders learned to fight for the free expression of religious differences. For differing reasons, evangelistic churchmen were able to join with Deists (Rationalists) in supporting religious liberty as a part of the constitutional foundations of the United States.

The 19th century: the Calvinist mold of U.S. life. The new nation's largest denominations were Congregationalist, Presbyterian, Baptist, Episcopalian, and Dutch and German Reformed. Each of these groups had a measure of Reformed tradition and often worked in the same cause. Presbyterians and Congregationalists cooperated in the settlement of the frontier, but Baptists and Methodists adapted better to the frontier and thus became the largest U.S. Protestant denominations. New denominations, such as the Cumberland Presbyterians and the Disciples of Christ, developed among the Western settlements.

A Calvinist viewpoint and pattern of life, favouring constructive activity rather than idle enjoyment, helped shape U.S. habits and ethics. Art, music, literature, and recreation were approved only if edifying. Sunday was viewed as a quiet day with freedom from business cares, minimal attention to farm chores, Sunday school and church attendance, and quiet conversation among friends. Reformed concern for a disciplined community, as interpreted by the Puritans, became a major influence on United States local, state, and national concerns. By its attendance to a proper discipline, the nation might receive the blessing of God and thereby enjoy peace and prosperity.

Presbyterian revivalist Charles G. Finney (1792–1875) saw the fruits of Christian revival in movements for women's rights, abolition of slavery, and temperance. The saving of souls for Christ and the building of a better world were two aspects of the same vision of the Kingdom of God in the United States.

In the 20th century: the breakdown of Calvinist dominion in U.S. Protestantism and U.S. life. After the U.S. Civil War (1861–65) there appeared new kinds of divisions in U.S. religion that made the older denominational distinctions obsolete. Conflict broke out between those who adapted Darwinism (evolutionism) to theology and those who viewed evolution as a threat to biblical authority, between the champions of biblical criticism and those who opposed it. These eventuated in a fundamentalist–modernist controversy that reached its peak in the 1920s.

Conflicts and divisions

In the latter part of the 20th century, the issues producing tensions apparently are focussed on the disagreements between Christians who would relate the churches to the restructuring of society and those who would work for the salvation of individuals. Reformed Christianity traditionally has held the individual and social emphases together, but its ability to continue to stress and to affirm both has been challenged. Another factor affecting Reformed and Presbyterian churches in the United States is the rise in numbers and importance of Roman Catholics, Lutherans, black religious groups, and non-Christian religions. The rising importance of these other religious groups has caused Calvinists to take a more humble place in the U.S. religious procession.

Reformed and Presbyterian world mission. *Asia.* In 1622 a seminary was instituted in Leiden (the Netherlands) to prepare missionaries for the Dutch Asian colonies (now Indonesia). Building upon work begun by

Catholic missionaries, they established a church in Indonesia that now numbers one-third of all Asians belonging to Reformed and Presbyterian churches.

The great century of world mission

The 19th century was the great century for the world mission of Reformed and Presbyterian churches. An ecumenical mission was designed by the U.S. Board of Commissioners for Foreign Missions, which was originally formed by Congregationalists, Presbyterians, and Dutch Reformed. It was decided, after a survey of the Middle East, that the primary task of missions in Muslim lands was to contact the minority Orthodox and independent Eastern churches that did not accept the decrees of the Council of Chalcedon of 451 (*e.g.*, Armenian, Nestorian, and Coptic), in order to strengthen them through fellowship.

The Presbyterian churches of Korea are second in numbers to Indonesia among Asian churches. The work there of missionaries became identified with the cause of Korean nationalism.

In other Asian countries there were missionary heroes, such as Robert Morrison in China and Alexander Duff in India. Churches in those countries, as well as in Japan, the Philippines, and Thailand, formed united churches with other Protestants. The Asian Reformed churches since World War I1 have shared the difficulties of being identified with Western institutions. Anti-colonialism, combined with Communism, had a devastating effect upon the churches of China.

In Africa. Reformed churches in Africa date from Dutch settlement in South Africa in 1652, soon followed by Huguenot and German Reformed refugees. With British occupation of South Africa in 1806, Scots brought Presbyterianism. In the 20th century, half of Presbyterian and Reformed church membership in Africa is in the Republic of South Africa. The Dutch Reformed churches have been closely identified with Boer (Dutch settler) nationalism and with its recent policy of apartheid (separating blacks and persons of mixed blood from whites). Churchmen opposed to this policy have been suppressed by the state.

Other Presbyterian and Reformed churches in Africa date from the early 19th century and are based on the work of such organizations as the London Missionary Society, which sponsored the missionary-explorer David Livingstone. The largest church outside South Africa is in the Malagasy Republic, followed by those in Ghana, Malawi, Cameroon, and Congo.

In other areas. Large Presbyterian churches of Australia, New Zealand, and Canada resulted from British colonization. In Canada the majority of Presbyterians united with Methodists and Congregationalists in 1925. There are many Presbyterians in former British colonies of the West Indies, and Presbyterian churches of Brazil include over half of the Latin American Reformed Christians.

Reformed Christians in the ecumenical movement. Since the time of Martin Bucer and John Calvin, the Reformed movement has had leaders who were untiring in their efforts toward church unity. In the 17th century the Scot John Dury and the Bohemian John Amos Comenius were notable for their ecumenical efforts. Dury's ability to work with both Archbishop William Laud (Anglican) and Oliver Cromwell (Puritan) demonstrated his love of unity.

Efforts toward church unity

The pietism of the 18th century helped to divide churches, but it also tended to encourage people to put aside their differences in pursuit of common goals. Missionary societies received support and sent out missionaries from diverse groups.

The union of individual denominations has proceeded gradually. A Consultation on Church Union in the United States began discussions in 1961 in response to a sermon by Eugene Carson Blake, a Presbyterian leader who became a leader of the World Council of Churches. In 1970 a plan for union known as The Church of Christ Uniting was sent to the member denominations for study, but was rejected in 1973. Included in this consultation are Reformed, Presbyterian, Congregational, Methodist, Episcopal, and Disciples churches.

In 1948, at Amsterdam, most Presbyterian and Reformed churches became members of the World Council of Churches. There are also many national and local councils of churches in which Reformed and Presbyterian churches participate. Since the second Vatican Council (1962–65), called by Pope John XXIII, there has been increased dialogue with Roman Catholics.

THE TEACHING OF THE REFORMED CHURCHES

Doctrines shared with most historic Christian communions. The Reformed churches consider themselves to be the Catholic Church, reformed. Calvin in his *Institutes* speaks of the holy Catholic Church as mother of all the godly. Bullinger in the Second Helvetic Confession makes it clear that Reformed churches condemn what is contrary to ecumenical creeds. The interpretations of the early Church Fathers and the decrees and canons of councils are not to be despised, "but we modestly dissent from them when they are found to set down things differing from, or altogether contrary to, the Scriptures." Universal articles of Christian faith, such as the doctrines of the Trinity, the humanity and divinity of Christ, and the sin of man and the saving work of Christ, are affirmed in Reformed faith.

The Catholic Church reformed

Doctrines shared with other Protestant communions, especially Lutherans. Reformed churches share with Lutheran and other Protestant communions the concept of justification by grace through faith. The essence of faith is trust in God's forgiving love coming as a gift through Jesus Christ. For Reformed Christians, as for Luther, the true treasure of the church is the good news of the grace of God. Scripture is the authoritative witness to the good news, but, as the Westminster Confession states, "authority thereof is from the inward work of the Holy Spirit, bearing witness by and with the word in our hearts." Calvin states: "There is no doubt that faith is a light of the Holy Spirit through which our understandings are enlightened and our hearts are confirmed in a sure persuasion." This is in accord with the Confession of Augsburg of the Lutheran churches.

The church and the sacraments. When Calvin tried to mediate between Luther and Zwingli on the Eucharist, he believed that Zwingli had been more concerned to show how Christ was not present than how he was. Calvin thought that Christians are raised by the Spirit into the presence of the risen Christ in the Lord's Supper.

Both Calvin and Bucer were concerned to keep the profane from unworthily receiving Communion. This influenced the development of church discipline. Some Lutherans developed disciplinary bodies within the church, but Luther himself was opposed to such coercion. The use of elders to oversee discipline within the parish became a prominent feature of Reformed church life. Calvin called such discipline the very sinews of the church.

In the struggle to maintain and extend that discipline after the death of Calvin, his successor, Theodore Beza, asserted that the presbyterian form of government was ordained by Christ. This led to men speaking of the divine right of presbyteries.

Scripture and tradition. Before the Reformation, Humanists rejected arguments that appealed to the authority of church tradition. They raised up the authority of Scripture within the church. Following them, Reformed churchmen insisted that nothing in terms of authority was on a level with Scripture; by Scripture all tradition is to be judged. They were free to quote copiously from early Church Fathers and medieval theologians but only when consistent with Scripture.

Church and state. The position worked out in Basel and Strassburg and put into practice in Geneva was that church and state should render reciprocal service yet remain separate and distinct. The church invisible consisted only of God's elect, and the membership of a particular visible church approximated the actual population of the corresponding state. Beyond their borders such national churches kept communion with each other in spite of differences of custom.

Separation of church and state

Obedience was required of Christians, even to kings viewed as less than good, unless the ruler commanded disobedience to God. On such occasions, man must obey God rather than man, but, even then, the private individual should not actively resist the ruler. The lesser magistrates were responsible to bring such rulers into line. The early resistance of Huguenots in France, John Knox in Scotland, and Puritans in England was justified on this basis.

English Puritans asserted that the government of the state should be patterned after the government of the church. This was one of the doctrinal sources of modern constitutional government, and another source was the Reformed belief that no man should be trusted with unlimited power.

There has been a constant Reformed belief that the kingdoms of this world can be brought closer to the will of God so that there might be a better justice for humankind, and this view involves churchmen in politics.

The sovereignty of God and double predestination. There has been little argument in Reformed theology about the positive side of the doctrine of predestination concerning the election of those whom God wills to save. Difference of opinion arose over whether God himself determines who is to be reprobated. Bullinger did not believe that it was God's will that "one of these little ones should perish." He therefore maintained that Christians should always hope for the best for all. John Calvin affirmed "double" predestination, meaning that both reprobation and election are within the active will of God. His reason found this appalling but scriptural. To call God, thereby, unjust was to judge the very standard of justice.

In his *Institutes* Calvin discussed predestination in the context of the love and grace of Jesus Christ. Later theologians expounded predestination more abstractly as an aspect of God's sovereignty. Arminianism rose in protest to this. The defenders of double predestination thought that Arminianism would cut the nerve of the Protestant doctrine of justification by grace alone and lead men back to popery (Roman Catholicism). Hence, at Dort, double predestination was affirmed as orthodoxy.

Man and society. The sociologist Max Weber and the economist R.H. Tawney attributed the economic vigour of Reformed countries to the Calvinist doctrine of double predestination. Calvinists were thought to be uneasy about their election and looking for outward signs of it. This uneasiness led to a this-worldly asceticism that gradually produced persons such as Benjamin Franklin, for whom pursuit of profit was a way of life. This thesis concerning Calvinism and capitalism has been criticized in detail. Capitalism can also be seen as a late medieval phenomenon that was able to develop in Reformed territory, while the Counter-Reformation stifled it in Catholic lands.

This-worldly asceticism

Real advance took place in Reformed countries not only in the economic area but also in science and technology, democratic institutions, and public education. Present scholarship is looking more to the actual social, political, and economic ideas of the Reformers. The zeal for a disciplined community may have had more to do with the social impact of Reformed Christianity than the doctrine of predestination.

WORSHIP AND ORGANIZATION

Liturgy. Reformed liturgies modified earlier forms of worship by using the vernacular, removing anything that implied the doctrine of sacrifice in the mass (as in Roman Catholicism), providing for congregational confession, and emphasizing the preaching of the word. Following Erasmus' recommendation, the singing of psalms became a characteristic feature of Reformed worship. To this day, there are some Reformed churchmen who are opposed to singing anything else.

Stress on preaching reached its peak among English Puritans. Some clergymen preached two hours on an Old Testament text on Sunday morning, two hours on a New Testament text in the afternoon, and devoted the evening to discussion of the day's sermons with the congregation.

Calvin held that the Eucharist should be celebrated weekly, though others believed that it was too sacred for frequent use. Care was taken to prepare participants for instruction and confession, and they were served around a table.

In the 20th century fresh attention has been given to the relevance of worship to the actual social, psychological, and material needs of human beings, as well as to the need for communicating the word to human hearts and minds. At the Iona Community in Scotland, where worship is directed to those who would work in economically deprived human situations, and at the Taizé Community in France, new forms of worship are being developed. In recent years there has been more emphasis upon celebration in response to the good news of God in Reformed worship and a greater appreciation of the various arts in worship than was true in the past.

Religious education. The requirements of Reformed life have demanded an educated clergy and an informed laity. Besides academic training for pastors, the practice has been for them to meet together frequently, at which meeting a pastor would interpret Scripture and the others join in critical discussion. Queen Elizabeth suppressed the custom in England, because she believed that four sermons a year were quite enough and that gatherings of pastors might be subversive.

The education of the laity was accomplished through preaching the word and teaching the catechism, such as Calvin's ***Little Catechism,*** which was designed for instructing the very young. Others, such as the Larger Westminster Catechism, were to be used to instruct pastors and teachers.

More recently, catechetical instruction has given way to inductive forms of education with emphasis on the age level at which instruction is taking place. There is also a greater concern to relate the Christian faith to the total community in which individuals are involved so that what is learned has direct meaning and relevance to daily life.

Present organization of Reformed and Presbyterian churches. In Presbyterian churches each local congregation is ruled internally by a session moderated by the pastor and composed of laymen (elders) elected from the congregation. A presbytery formed of ministers and elders representing each congregation rules over the local congregations on a district level. In other Reformed churches the district association has less power and the local congregation more than in Presbyterian churches. In a few Reformed churches there is a presiding bishop who moderates the presbytery.

Beyond the district association are regional synods or conferences and national assemblies. These bodies are usually composed of an equal number of clergy and laity. Since 1877, there has been a World Alliance of Reformed Churches, which was joined in 1970 at Nairobi, Kenya, by the International Congregational Council to form the World Alliance of Reformed Churches (Presbyterian and Congregational).

Although a few Reformed organizations are still state establishments, their ties with the state are increasingly tenuous. In practice, there is now little difference between established and free Reformed churches.

Social ethics. Reformation leaders were much involved in the total life of their communities. Calvin's relation to the education, health and welfare services, industry, finance, and politics of Geneva is well documented; and R.H. Tawney, impressed by this, has called Calvin a "Christian socialist." The English Puritans believed that, if they could discipline the nation, God's blessing would come upon the land, instead of war, famine, and pestilence.

Concern to reshape society in the direction of a more just situation for mankind has been normative among Presbyterian and Reformed churches. Sometimes, the actualizing of this concern in the past has resulted in petty rules and harsh administration, but it has also had humane dimensions. Such concern is still a living force among Reformed churchmen, even though they do not intend it in the same way as Calvin or the Puritans.

Types of Reformed piety. In Zwingli, Calvin, William I of Orange, and Cromwell, a classic type of Reformed piety is manifest. Each man viewed himself as God's instrument in redeeming human affairs, even at great cost to himself. In obedience to God, these men disciplined their lives, and they expected others to do the same. Living under God's mercy, they were not afraid of the powers of this world, were men of prayer and action, and were remarkably pragmatic in their choices.

In a less heroic mold was the Reformed churchman who personally did not expect to change history but who encouraged the development of godliness in those about him, beginning with himself. The increasing emphasis in the late 16th century upon the personal experience of saving faith caused the Reformed tradition to become a nursery for pietism in the late 17th and 18th century.

Along with a more confessional orthodoxy and a more rationalistic liberalism, such pietism remains to the present day. A new style of worldly Christianity now seems to be emerging, with Christ, the man for others, as its model.

Current Reformed doctrines and practices. For those who interpret Reformed Christianity as a strict adherence to old confessional standards, there is only a small group of "true Calvinists" left. More influential than these have been Karl Barth, Emil Brunner, John and D.M. Baillie, Hendrik Kraemer, and Reinhold and H. Richard Niebuhr (all 20th-century theologians in the Reformed tradition). They have shown that Reformed doctrine can be stated in a fresh way so that it is alive and active in the 20th century.

CONCLUSION

World civilization gradually has become secularized in the past five centuries: constitutionalism in politics has replaced the divine right of kings; the economic process has been rationalized, and no one advocates a theocratic regulation of commerce and banking; and new scientific models of the universe have supplanted earlier mythological models. Reformed Christianity has not only lived through this secularization, but it bears partial responsibility for it.

The problem of the churches is to relate relevantly to secular men in a secular world. There are still human beings who need a new sense of confidence and hope, and there are institutions that need to be modified. Reformed churchmen continue to work at both tasks: communicating the good news of the Christian faith to those who seek meaning in their lives and attempting to reshape society so that individuals can live more tolerably.

BIBLIOGRAPHY. JOHN T. MCNEILL, ***The History and Character of Calvinism*** (1954), a comprehensive treatment of the rise and development of Presbyterian and Reformed churches, with an excellent bibliography; W. FRED GRAHAM, ***The Constructive Revolutionary: John Calvin and His Socio-Economic Impact*** (1971), the best current assessment of the discussion generated by the Weber–Tawny thesis concerning the economic impact of Reformed doctrine; JAMES HASTINGS NICHOLS, ***Corporate Worship in the Reformed Tradition*** (1968), an overview of the variety of forms developed in the Reformed tradition of the public worship of God; HEINRICH HEPPE, ***Reformed Dogmatics Set Out and Illustrated from the Sources***, rev. and ed. by ERNST BIZER, foreword by KARL BARTH, trans. by G.T. THOMSON (1950), enables the reader to get beyond Calvin's ***Institutes*** to some acquaintance with other Reformed theologians of the 16th and 17th centuries; ROBERT MCAFEE BROWN, ***The Spirit of Protestantism*** (1965), a popular interpretation of contemporary Reformed theology; ARTHUR C. COCHRANE fed.), ***Reformed Confessions of the 16th Century***, with historical introductions (1966), 12 classic confessions of the 16th century together with the ***Heidelberg Catechism*** and, oddly, ***The Theological Declaration of Barmen*** in 1934 in the appendix (the introduction is quite helpful); THOMAS F. TORRANCE (ed. and trans.), ***The School of Faith: The Catechisms of the Reformed Church*** (1959), ten catechisms of the 16th and 17th centuries; ***Reformed World*** (quarterly), a periodical, published by the World Alliance of Reformed Churches (Presbyterian and Congregational), Geneva, Switzerland, that reports on the life and work of Reformed and Presbyterian churches throughout the world.

(J.C.S.)

Refrigeration Equipment

Refrigeration implies the development and use of temperatures lower than those existing in ambient space, from the cooling of a beverage with ice to the maintenance of large chilled or frozen storage space, or even the production of the extremely low temperatures of the cryogenic region.

Artificial refrigeration has been known only since the early 19th century and has been in widespread use in homes and industry only since the early 20th. Without refrigeration, most perishable foods would have to be consumed where they are produced, and the transportation of foodstuffs over long distances in chilled or frozen form would be impossible. The manufacture, operation, and maintenance of refrigeration equipment is a major world industry.

Distribution as an indicator of technological development

The distribution of refrigeration equipment throughout the world reflects the varying levels of technological development. In the industrially advanced countries, refrigeration is accepted as an indispensable part of contemporary life. In the less developed countries, refrigeration facilities are often available for the affluent but almost completely lacking for most of the population. Even in these countries, however, limited central refrigeration must be provided for many perishable products that cannot be consumed as soon as they are produced.

Foodstuffs deteriorate less rapidly as the storage temperature is lowered; and, at temperatures below 40° F (4° C), many food products can be kept for days or even weeks without serious loss in quality and appearance. If perishable products must be kept for long periods, it may be possible to freeze them with little loss in character and maintain their quality for periods of six months or longer. The most important causes of food deterioration are the destructive action of bacteria, yeast, and molds, the growth of which is inhibited by low temperature. Refrigeration has made it possible for such meat-producing nations as Australia and Argentina to export much of their output to Britain and western Europe, usually in frozen form.

Refrigeration for comfort cooling has been developed extensively in the United States and is gaining wider acceptance in other portions of the world (see HEATING, VENTILATING, AND AIR CONDITIONING).

HISTORY

Before the advent of mechanical refrigeration, man had to depend on nature for means of cooling. The Greeks and Romans, among other ancient people, transported snow and ice from the high mountain levels to the cities. They also made use of the snow cellar—a cavity dug in the ground, its sides stabilized with boards or logs behind which heavy layers of straw provided insulation. Snow was packed into the cavity, and the top of the snow was covered with straw. So insulated, the snow, usually packed down to solid ice, kept for long periods of time and could be used for refrigeration, at least by those wealthy enough to afford the cost of transportation and of storage.

During the 19th–early 20th centuries, in Europe and the United States, ice was sawed in large blocks from northern lakes in the winter and stored for use during the summer in icehouses constructed either above or below ground. Such ice could be successfully shipped, insulated by sawdust, to the southern United States and the tropics.

Evaporative cooling

In India and Egypt, evaporative cooling combined with night radiation to outer space was long used to manufacture ice. Under clear skies and in a dry climate, if water is placed in shallow earthenware trays on a straw bed, the rapid evaporation from the water surface and from the wet sides of the trays, combined with nocturnal radiation, can result in ice forming in the trays, even though the night air temperature does not fall below the freezing temperature of water. Under certain conditions, only a film of ice is formed; but under other conditions, it is possible to freeze the water into solid blocks.

Prior to the advent of mechanical refrigeration, natural ice or tray ice as described above represented the only means available for the creation of refrigeration. Ice is useful for refrigeration because it has the unique characteristic of always melting and freezing at a fixed temperature, 32" F (0° C); furthermore, when it melts it absorbs an unvarying amount of heat energy per unit of weight. Thus, an area surrounded by or containing melting ice will be refrigerated, or held at a lower temperature. Man learned at an early date that temperatures much lower than that of normal ice melting could be created by mixing salt with the ice. Ordinary salt (sodium chloride), when mixed with ice, can drop the temperature (and the freezing point) of the mixture to −6° F (−21° C); calcium chloride can lower the temperature to −67" F (−55° C). Some liquid organic compounds, such as ethyl alcohol, methyl alcohol, and certain glycols, also lower the freezing point of water.

The first man-made refrigeration, produced by the evaporation of ethyl ether into a partial vacuum, is credited to William Cullen at the University of Glasgow in 1748. Although his procedure involved vapour refrigeration, development of a successful refrigeration machine of this type was not achieved until Jacob Perkins, in 1834, obtained a British patent on a volatile-liquid closed-cycle system using a compressor. Perkins made at least one successful ice machine but did not actively promote his invention.

In 1844 a U.S. physician, John Gorrie, in Apalachicola, Florida, developed a machine to provide ice and air cooling for his hospital. His machine, which operated successfully, was similar to one proposed (but never constructed) by the inventor Oliver Evans in Philadelphia in 1805. Gorrie's machine consisted of a compressor that compressed air, which was then cooled by circulating water. As the air re-expanded in an engine cylinder, it dropped to a sufficiently low temperature to create useful refrigeration for the production of ice or other cooling. The expanded air was then drawn back to the compressor cylinder for compression and recirculation.

The first ice machine

In 1856 another American, Alexander C. Twinning of Cleveland, produced what is believed to have been the first commercial ice by means of a vapour-compression machine. His first patent was taken out in 1850. James Harrison, who emigrated from Scotland to Australia in 1837, became interested in refrigeration. After surveying the machines of Gorrie and Twinning, he developed the first vapour-compression machines for use in the brewing industry and for freezing meat for shipment. Harrison's machines, which were produced for several decades, used ethyl ether as the refrigerant.

A second type of refrigeration machine was developed in France in the 1850s by Ferdinand Carré. In Carré's system the refrigerant, normally a vapour, as in the vapour-compression machines, is absorbed in a suitable liquid. This solution is heated, driving off the refrigerant as a vapour, which is then condensed. Evaporation of the liquid, in a manner very similar to that in the vapour-compression system, produces the desired cooling. This refrigerant vapour is again absorbed in the liquid, thus completing the cycle.

Carré's first machines employed water as a refrigerant with sulfuric acid as an absorbent, but in 1859 he introduced ammonia as a refrigerant with ammonia–water as an absorbent. This successful combination was used widely throughout the world. Some of Carré's machines passed through the Federal blockade during the American Civil War and supplied the Confederacy with much-needed ice when the supply of natural ice from the northern United States was cut off.

The basic principles on which refrigeration machines operate were thus developed by 19th-century inventors. Subsequent inventions involved only modifications and improvements in the machines and processes.

Improvements in vapour-compression refrigeration using ammonia hinged largely on the development of successful compressors in the last half of the 19th century. Efficient ammonia compressors gradually edged out the air-cycle machine, and vapour-compression refrigeration came to predominate. As a vapour-compression refrigerant, however, ammonia had a number of disadvantages, the most important being that it was toxic and extremely objectionable when breakage or other sources of leakage occurred. This meant that it was undesirable for use on shipboard and in places where air cooling for comfort was required. The search for refrigerants that were nontoxic, had no odour, and could perform effectively in conventional equipment led to the use of carbon dioxide; a major objection, however, was the fact that the pressures required for its use were extremely high. In addition, carbon dioxide, although not toxic itself, is lethal in high concentrations, because it excludes oxygen and interferes with the regulatory breathing system of man. In the 1920s a group of synthetic refrigerants known as halogenated hydrocarbons was developed. It was found that by starting with a hydrocarbon such as methane (CH,), it was possible to substitute two chlorine and two fluorine atoms for the four hydrogen atoms and produce a new chemical (CCl_2F_2) with highly desirable properties as a refrigerant. This new chemical was called dichlorodifluoromethane and was named refrigerant-12. It was essentially nontoxic, had no odour, and could exist in the air in concentrations as high as 18 percent, by volume, without lethal effects. It was also found to have refrigerant properties resembling ammonia and to perform well in compressors. The early halogenated-hydrocarbon refrigerants were licensed by one company and carried the name of Frwn, with an appropriate numerical designation.

Once the basic patents expired, these refrigerants were made freely throughout the world by many chemical-manufacturing companies. These substitution-type halogenated-hydrocarbon refrigerants can be based on molecular structures other than methane, and a broad range of special-purpose refrigerants is now available. The two additional ones that are most important at the present time are monochlorodifluoromethane, $CClHF_2$, also known as refrigerant-22, and trichloromonofluoromethane, CCl_3F, known as refrigerant-11. Refrigerant-22 operates at higher pressure than refrigerant-12 and is suitable for the production of lower temperatures because of its high-pressure characteristics. Refrigerant-11 is extremely desirable for use with centrifugal compressors in which it operates in a low-pressure range, often in the vacuum region. Refrigerant-22 and refrigerant-1 1 are slightly more toxic to humans than refrigerant-12 but are used extensively.

Refrigerant number

The numbers used with refrigerants represent an internationally accepted designation system, originally adopted by the American Society of Heating, Refrigerating, and Air-Conditioning Engineers.

TYPES OF REFRIGERATION SYSTEMS

Modern mechanical refrigeration. For continuous mechanical refrigeration, the same refrigerant must be repeatedly used for an indefinite period. Three patterns for doing this are now in common use: the vapour-compression system, the gas-expansion cycle, and the absorption system. Of these, the vapour-compression system is by far the most effective and is used more extensively than any other arrangement. It consists basically of three elements: an evaporator, a compressor, and a condenser. The vapour-compression-system cycle is very similar to the air-compression-system cycle described later.

Referring to Figure 1, in the evaporator the refrigerant boils (evaporates) at a temperature sufficiently low to absorb heat from a space or from a medium that is being cooled. The boiling temperature is controlled by the pressure maintained in the evaporator, since the higher the pressure, the higher the boiling point. The compressor removes the vapour as it is formed, at a rate sufficiently rapid to maintain the desired pressure. This vapour is then compressed and delivered to the condenser. The condenser dissipates heat to circulating water or air. The condensed liquid refrigerant, now ready for use in the evaporator, is then sharply reduced in pressure by passing through an expansion valve: Here, the refrigerant's pressure and temperature drop until they reach the evaporator pressure and temperature, thus allowing the cycle to be repeated.

During expansion some of the liquid flashes into vapour so that a mixture of liquid and flash vapour enters the evaporator. In a refrigeration system the low pressure in the evaporator is set by the refrigeration temperature which is to be maintained. The high pressure maintained in the condenser is determined ultimately by the available cooling medium; *i.e.*, the temperature of circulating water or the atmospheric air temperature. The process is one in which the refrigerant absorbs heat at a low temperature and then, under the action of mechanical work, the refrigerant is compressed and raised to a sufficiently high temperature to permit rejection of this heat. Mechanical work or energy supplied to the compressor as power is always required to raise the temperature of the system.

Compressors. The compressor, the key element of the system, can be powered by electric motor, steam or internal-combustion engine, or steam or gas turbine. Most compressors are of the reciprocating (piston) type and range from the fractional-horsepower size, such as those found in domestic refrigerators or in small air-conditioning units, to the large multicylinder units that serve large industrial systems. In these large multicylinder units, capacity can be controlled with automatic devices that prevent the valves in certain cylinders from closing. For example, in a six-cylinder unit, if the valves are held open on two of the cylinders to keep them inoperative, the capacity of the machine is reduced by one-third when operating at normal speed.

Centrifugal compressors are used for large refrigeration units. These employ centrifugal impellers that rotate at high speed. Centrifugal compressors depend for their compression largely on the dynamic action of the gases themselves as they flow in the diffusion passages of the compressor (see also PUMP). These compressors can be large centrifugal compressors made with a single impeller, or with two to four or more impellers in series, to compress the gas through the range required. Centrifugal compressors are built for capacities ranging from approximately 600,000 BTU per hour to over 12,000,000 BTU per hour (50 to 1,000 tons). These are used extensively for large air-conditioning installations, and also find usage in the industrial field when gases are compressed for liquefaction or for transportation, such as in the natural-gas industry, and when air is compressed to produce liquid oxygen or nitrogen.

Evaporators. The evaporator is the part of a refrigeration system in which the cooling is actually produced; the liquid refrigerant and vapour from the expansion valve are introduced to it. As it vaporizes, the liquid absorbs heat at low temperature and cools its surroundings or the medium in contact with it. Evaporators may be direct expansion—that is, acting directly to cool a space or product---or they may operate as indirect-expansion units to cool a secondary medium, such as water or a brine, which, in turn, is pumped to a more distant point of utilization. A domestic refrigerator, for example, is a direct-expansion unit in that its evaporator directly cools the air in the food compartment and also directly contacts the water trays used for making ice. Evaporators vary greatly in design, with those used for cooling air often made as continuous pipe coils, with fins mounted outside the pipes to give greater surface contact to the air being chilled. For cooling liquid, such as a brine or water, the shell and tube arrangement is common. In this, the brine passes through tubes surrounded by the boiling (evaporating) refrigerant, which is contained in a larger cylindrical shell. The brine tubes, in turn, are welded or rolled into tube sheets at the shell's end to prevent leakage of the refrigerant from the shell or into the brine circuit.

The expansion valve that feeds the evaporator must control the flow so that sufficient refrigerant flows into the evaporator for the cooling load but not in such excess that unevaporated liquid passes over to the compressor, with the possibility of causing damage to it.

Condensers. The condenser of a vapour system must dissipate heat from the hot vapour it receives from the compressor and condense this vapour to liquid for reuse by the evaporator. Condensers either dissipate heat to the ambient atmosphere through externally finned surfaces or by a shell and tube arrangement in which the vapour delivers heat to circulating water that passes through tubes contacting the refrigerant vapour. The temperature of the vapour is kept above that of the circulating water or air by compression to insure that heat is transferred to the coolant; thus, when the vapour is allowed to expand, its temperature drops well below that of the coolant.

Double-pipe condensers are also used. In such units the refrigerant vapour and condensate pass in one direction through the annular space between the two tubes, while the water, flowing in the opposite direction through the central tube, performs the cooling function.

Air-cycle refrigeration. This type of refrigeration came into extensive use before the nontoxic halogenated-hydrocarbon refrigerants were developed. In this system the refrigerating medium, instead of operating at two essentially fixed temperatures, operates over a range of temperature. In the conventional arrangement the same air is continuously circulated. This makes possible a plant of smaller physical size, because the air is compressed at all stages of the cycle, and, with dry gas charged into the system, moisture does not create a problem.

Adapted from B. Jennings, *Air Conditioning and Refrigeration*

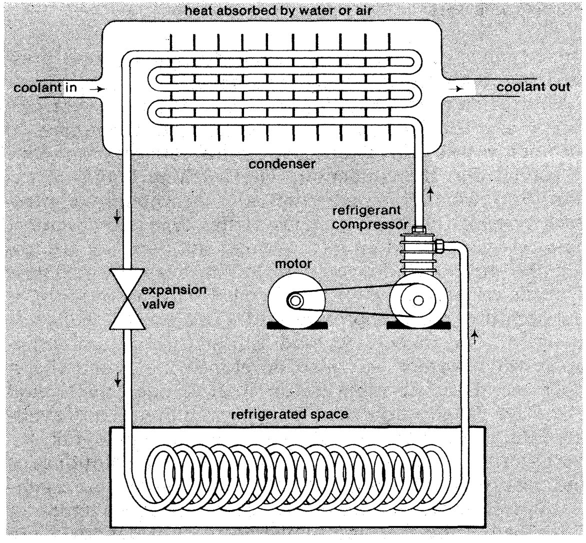


Figure 1: Elements of a vapour-compression refrigeration system.

Operation of the closed-air-cycle system somewhat resembles that of the mechanical system just described. Air from the chilled space enters the compressor, where it is compressed; this process also raises the temperature. The hot compressed air then enters the water-cooled heat exchanger, which corresponds to the condenser of a vapour system. Here the air is cooled to **a** temperature only a few degrees warmer than the inlet-water temperature. The cooled compressed air then enters the expander engine, where, in expanding, both the temperature and the pressure are lowered.

Systems used in aircraft

Air-cycle refrigeration systems of the nonclosed type are used in many modern commercial aircraft for air-conditioning the cabin. Here, a small part of the air, partially compressed by the main power compressors of the aircraft, can be cooled by circulation of high-altitude ambient air over the compressed-air coils. The compressed air, when expanded in an expander turbine, drops enough ın temperature to be used to chill the cabin air. Such a system for aircraft is effective and lightweight because it makes use of air already compressed by the power compressors. Representative values for an aircraft system show that the expander turbine can readily lower the temperature of the compressed air supplied to it by 100" F or more, so that with 100" F (38" C) air entering the turbine, 0° F (–18° C) air becomes available. This cold air is mixed with sufficient reduced-pressure warm air from the compressor system to temper it, and the

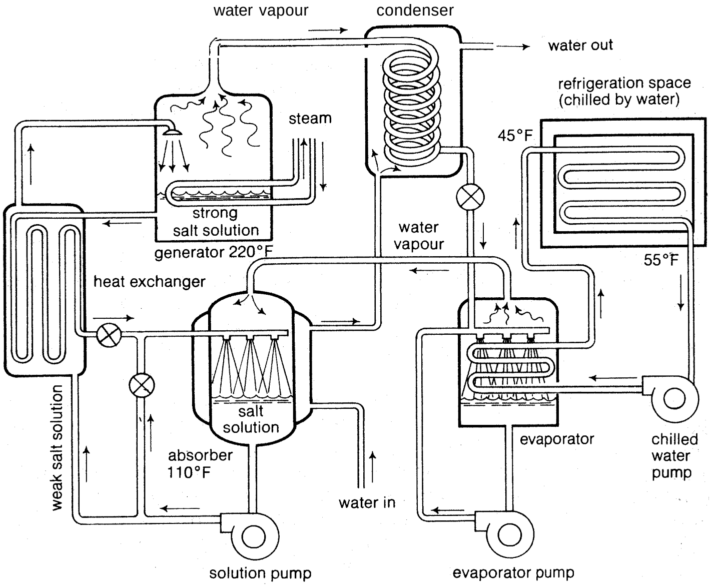


Figure 2: Absorption system using aqueous solution of lithium bromide with water as the refrigerant.

mixed air is delivered to the cabin for cooling and pressurizing to a satisfactory value for the flight altitude.

Absorptionrefrigeration. In this method the refrigerant is absorbed in a chemical solution, then pumped to another vessel. There, on heating, the refrigerant vapour is forced into the condenser. The resulting liquid on expansion lowers its temperature as in the vapour-compression system. Industrial systems of this type use ammonia as a refrigerant and ammonia–water solutions for absorbing the ammonia delivered from the evaporator. These systems can operate over an extended range of evaporator temperatures from approximately —40" to 40° F (—40" to 4° C) or higher. A modification of the ammonia-absorption system was also developed for use in heat-operated domestic refrigerators and, at one time, found extensive usage where heat energy was more readily available than electricity. Another type of unit employing water as a refrigerant, with lithium bromide salt solutions for the absorption medium, has found wide use in air conditioning. Even though these units cannot produce temperatures below the freezing temperature of water, they are very effective for the 38" to 58° F (3° to 14° C) range required for air-conditioning use.

An ammonia–water system is effective in its operation because water will absorb large quantities of ammonia vapour, the amount absorbed increasing with the external pressure and decreasing with rising temperature. The absorber operates at about evaporator pressure and is supplied with a cooled solution of water–ammonia having a low concentration of ammonia. This weak solution absorbs ammonia gas until the liquid becomes almost saturated with ammonia. The heat created during the absorption process is then removed by cooling water. The ammonia-laden solution is pumped through a heat exchanger into the generator. The generator employs steam or another heat source to boil off ammonia until the solution is reduced to a saturated condition (low in ammonia) at generator pressure and temperature. The ammonia vapour from the generator moves over to the condenser, where it is condensed; and then, as liquid, it passes through the expansion valve into the evaporator. The hot, weak solution from the generator passes through the heat exchanger, where it cools in warming the strong liquid and is again throttled into the absorber to absorb ammonia vapour from the evaporator.

Absorption units in air conditioning

Absorption systems for air conditioning use water as the refrigerant and a water–salt solution as the absorbent, with lithium bromide salt having the widest acceptance. In a vacuum system, the vapour pressure of an aqueous solution of concentrated lithium bromide is so low that, when an evaporator supplied with water is connected to a vessel in which such a salt solution is being circulated, the water evaporates to produce cooling. For example, the vapour pressure of a 61 percent (by weight) lithium bromide solution has a vapour pressure of 0.27 inch (6.86 millimetres) of mercury (Hg) at 110" F (43" C). This vapour pressure is sufficiently low to cause water to evaporate (boil) at 43° F (6" C).

The diagrammatic layout of such a system is shown in Figure 2. The absorber is supplied with a slightly cooled solution, high in salt concentration (about 65 percent), that will absorb the water produced in the evaporator as the cooling takes place, ultimately producing a chilled product at approximately 45° F (7° C). The salt solution in the absorber, when it absorbs the water vapour, changes it to liquid and in so doing releases the latent heat of condensation along with some heat of mixing. Such cooling must be provided for the absorber to keep it below the 110" F mentioned above for effective operation. The salt solution from the absorber, now greatly diluted by the water added, is pumped through a heat exchanger on its way to the generator. In this vessel the temperature of the solution is increased to about 220" F (104" C), and the surplus water is driven out of solution by heat from the steam supply. The water vapour that is given off passes to a condenser in which it is condensed back to water, and from the condenser the water can then be throttled into the evaporator. A circulation pump is required at the evaporator to distribute the liquid water over the coils to promote better evaporation. Similarly, recirculation is also used in the absorber to improve the absorption process. In the generator, with the temperature of the solution at 220" F, the vapour pressure is approximately three inches (75 millimetres) of mercury absolute.

Thus, the system described operates at a generator–condenser pressure of three inches of mercury and an evaporator–absorber pressure of 0.27 inch of mercury.

The Peltier effect

Thermoelectric refrigeration. Thermoelectric refrigeration came into prominence only during the 1960s, when the unusual characteristics of semiconductor materials showed a sufficiently satisfactory performance to justify commercial usage. The useful effect for refrigeration is the so-called Peltier effect, observed in France in 1834 by Jean Peltier, who noticed that, if an electric current was passed through junctions of two dissimilar metals, one junction cooled. If the current was reversed, the junction heated; or, to complete the picture, if two such junctions were included in a circuit, one junction always cooled, while the other junction heated. The use of properly doped semiconductor materials, such as bismuth telluride (Bi_2Te_3), for the junction materials gave this method of refrigeration commercial significance (see also THERMOELECTRIC DEVICES).

Many such junctions can be placed in series and attached to a plate in such a manner that heat is absorbed on the cold side, producing useful refrigeration, while heat is dissipated on the other side. Modules can be designed in functional forms and shapes to meet the needs of the particular application. The performance of such a system is far from ideal, however, because the junctions must be made short to decrêase electrical resistance and to decrease the cost of materials. Such shortening increases the transfer of heat through the assembly and thereby offsets some of the useful cooling. In any specific application, a compromise must be made to achieve optimum characteristics. Although the Peltier effect is the important phenomenon in cooling, the closely related Seebeck effect is more easily measured and is used to indicate the thermoelectric performance of semiconductor materials.

APPLICATIONS

Household refrigerators and freezers. Household refrigerators normally employ a vapour-compression system, with the motor and compressor mounted in a hermetically sealed container. In such a unit, refrigerant leakage from the system is minimized; most domestic units go through their complete life without the need of additional refrigerant charge or change of lubricant. The evaporator serves to cool the storage space for food supplies and usually is set to maintain this space at a range of 33° to 45° F (1° to 7" C). The evaporators of most units are also equipped to freeze water into ice cubes. Many domestic units also provide, in the evaporator section, a low-temperature storage compartment for holding foods in frozen condition, with the temperature controlled between −10" and 20° F (−23" and −7° C).

Domestic home freezers are usually designed to maintain a temperature range of −10" to 10° F (−23" to −12" C), a range in which foodstuffs remain in a frozen state and can be kept for extended periods. The home freezer usually employs a vapour-compression system, with a compressor delivering the refrigerant to an air-cooled condenser. The liquid refrigerant from the condenser passes through an expansion valve to the evaporator, which carries out the useful cooling for the freezer. One problem with domestic refrigerators and freezers is that ice (frost) collects on the coils and must be removed periodically. This requires the application of sufficient heat to the coils to melt the ice, which can be done automatically at timed periods, usually by bypassing hot condenser gas to the evaporator coils or by the rapid addition of heat from electrical resistors. Usually, the time required for melting (defrosting) is so short that the product does not warm up to the point of damage.

Defrosting

Cold-storage warehouses. Cold-storage warehouses usually have rooms that store frozen food and others in which the food is only chilled. The storage life of nonfrozen products is improved by keeping the temperature as close to 32" F (0° C) as possible without actually freezing. Some vegetable products can tolerate temperatures slightly below 32° F and not suffer damage; but others, such as lettuce, are extremely sensitive to freezing and must always be kept slightly above 32° F. Bananas must not be chilled below 55° F (13" C), or the ripening process stops. Cold-storage warehouses usually have a central refrigeration plant of the vapour-compression type in which the evaporator cools a salt or glycol brine; the brine is pumped to the various storage rooms, where it flows through coils to hold the room at a required temperature. Frozen-food rooms operate frequently in the range of −20" to 0° F (−29" to −18° C). It is more desirable, however, to maintain a temperature with a minimum of variation.

Refrigerated transport. Refrigerated products are moved over long distances by specially insulated trucks and railroad cars. Many of these involve the operation of a complete refrigeration system driven by a gasoline or diesel engine. Others employ liquid nitrogen or Dry Ice (frozen carbon dioxide) to provide the cooling.

Ice making. During the early part of the 20th century, the commercial manufacture of ice was a major industry. The ice was usually made in cans holding about 300 pounds of water, from which the ice was removed in blocks for cutting and distribution. Some commercial ice is still made in blocks, but most frequently it is broken up into small pieces—for example, for distribution over green produce and other food products prior to shipment. Large amounts of ice for this purpose are also made in small cube for·m or in cylindrical tubes that are broken into useful sizes. Ice is also made in chip, flake, or slush form by freezing it onto a surface and scraping it off.

Air conditioning and other uses. Air-conditioning systems use refrigeration equipment to produce the required cooling. Refrigeration for public buildings can be produced in large central plants from which the cooling is distributed either as chilled air or as chilled water, pumped to the points of utilization. Residence air conditioning may also make use of a central refrigeration plant and distribute the cooling to various areas of a house, either by air or chilled water. In addition to central systems, extensive use is made of unit air-conditioners, which, when placed in a window, dissipate the heat load from the compressor to the outside and, using a separate fan, distribute chilled air from the evaporator to the inside space. Unitized-space conditioners are also used, placed completely within the air-conditioned space. These house the whole refrigeration system but always require a means of heat dissipation, such as a circulating water supply or an air circuit to the outside (see also HEATING, VENTILATING, AND AIR CONDITIONING).

Refrigeration is also used for many other purposes, such as chilling drinking water by means of a unit mounted in a drinking fountain. In medical practice, refrigeration is required to store biological and other temperature-sensitive materials and is also employed for direct chilling of patients in cases in which, under certain circumstances, by lowering the deep-tissue temperature of the body, the life-process rates can be slowed down to permit carrying out certain types of surgical operations or to reduce the speed with which a disease might spread throughout the body.

Freeze-drying

Lyophilization (freeze-drying) is used to preserve certain delicate chemicals, biologicals, or tissues. In this process the material is rapidly frozen, and under high-vacuum conditions the water is removed by subliming it to the vapour phase while the material is still frozen. During lyophilization the cell structures of many materials are maintained essentially intact, and the character of the basic product is preserved; while in others the form changes to that of a powder, but the character is unaltered. The process is necessary for certain extremely delicate materials, but the process is also being applied to foodstuffs for commercial use. Meat, for example, can be lyophilized into a material with spongelike appearance that, when reconstituted by the addition of water, has much the same character, taste, and form as the original material.

BIBLIOGRAPHY. CHARLES SINGER *et al.* (eds.), *A History of Technology,* vol. 5, *The Late Nineteenth Century, c. 1850 to c. 1900,* pp. 45–51 (1958), a brief account of refrigeration history during the period of its most active development; W.R. WOOLRICH, "The History of Refrigeration: 220 Years of Mechanical and Chemical Cold, 1748–1968," *ASHRAE Jl.,* 11:31–39 (July 1969), an article that chronicles in detail the people and machines that contributed to the creation of modern-day refrigeration; BURGESS H. JENNINGS, *Environmental Engineering: Analysis and Practice* (1969), *a* college-level text that covers in depth engineering aspects relative to the equipment needed and methods employed in controlling the indoor environments through cooling, heating, and cryogenics; *ASHRAE Handbook of Fundamentals,* ch. 2 (1967), and *ASHRAE Equipment Guide and Data Book* (1969), two large volumes that represent the most extensive coverage of current refrigeration practice and application that has ever been written; A.F. IOFFE, *Semiconductor Thermoelements, and Thermoelectric Cooling* (1957; orig. pub. in Russian, 1956), a masterful summary of the author's lifelong studies on thermoelectric refrigeration; H.J. GOLDSMID, *Thermoelectric Refrigeration* (1964), a treatment of the subject of thermoelectric cooling from the viewpoint of solid-state physics.

(B.H.J.)

Refugees

A refugee, according to the 1951 statute of the Office of the United Nations High Commissioner on Refugees (UNHCR), is any person who, "owing to well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group or political opinion, is outside the country of his nationality and is unable or, owing to such fear, is unwilling to avail himself of the protection of that country; or who, not having a nationality and being outside the country of his former habitual residence, is unable or, owing to such fear, is unwilling to return to it." The definition offered in 1951 by the Intergovernmental Committee for European Migration (ICEM) included an additional broad category: any person who "has been the victim of a war or a disaster which has seriously disadvantaged his condition of living."

This article will apply the more circumscribed and concise definition offered in Elfan Rees' ***We Strangers*** and ***Afraid*** (1959) : that a refugee is "anyone who has been uprooted from his home, has crossed a frontier—artificial or traditional—and looks for protection and sustenance to a government or authority other than his former one." The operative terms of this definition are first, "uprooted," indicating a forced departure from the home country, as distinguished from a migratory movement that is essentially the result of differential population or economic pressure, and, second, having "crossed a frontier," denoting the difference between a refugee and a person displaced (mostly temporarily) within his own country (an internal refugee). The misery of internal or national refugees is real and urgent, but they belong to a category of their own whose origin, plight, problems, and prospects are different from those of international refugees.

THE NATURE OF THE REFUGEE PROBLEM

Historic forces. Man's story is in large measure the story of his wanderings. Throughout recorded history, invasions, wars, and conquests were followed by large-scale movements of populations. The so-called Great Peoples' Migration in the wake of the fall of Rome involved multitudes of human beings driven away from their home countries by invading hordes. By AD 900 the chain of "great migrations" had come to an end, and Europe's various tribes had settled in their respective areas. The continent entered a "sedentary" era. But within the new ethnoterritorial framework, the forcible removal and flight of religious groups has, since medieval times and especially since the Reformation, constituted a constant and important element in social and economic history of the European continent. Expulsions of Jews from western Europe, Catholics from Protestant countries, and Protestants from Catholic states are examples.

At a time when state frontiers were less clearly defined and jealously guarded than they are now, the movement from one country to another did not require passports and visas; the right to asylum, which for a refugee is a corollary to the right to life, was commonly recognized and honoured. As a result, although there were numerous waves of refugees in earlier centuries, there was no refugee problem. Avenues of escape were narrowed only with the emergence of fixed and closed state frontiers characteristic of the late 19th and of the 20th century and with the complete control of the globe by sovereign nation-states. The era of open frontiers was over.

The era of intolerance

The lines of sovereignty have become drawn ever sharper. The value apparently placed on national and racial homogeneity has inspired intricate networks of immigration restrictions in areas in which freedom of movement had been unimpeded for centuries. This "new order" has created a huge category of uprooted people facing immense, often insuperable difficulty in finding a temporary asylum, let alone a new home, outside of their regular habitat. Ethnic, religious, and political minorities have been persistently—and increasingly—singled out as objects of intolerance, discrimination, and persecution on the part of the government or public of the countries they have lived in, often irrespective of how ancient and deep were their roots in these countries and how willing they were to assimilate completely into the cultural and social fabric. A classical case in point were the 600,000 Jews in Germany who were deprived of citizenship and property, deported, or forced to flee Hitler's Third Reich.

By the 1920s and 1930s the long-standing tradition of political asylum had broken down completely, partly because of growing insensitivity to human suffering and partly because of unprecedented numbers of refugees. Existing mechanisms of immigration and naturalization, which had been adjusted to a smaller, "normal" influx of people, ground to an actual standstill.

The worldwide nature of the refugee problem. In modern times each great upheaval casts up its own living jetsam—the fugitives, the expellees, the homeless ones, more than 40,000,000 of them since World War II. The "Age of the Uprooted" and the "Century of the Homeless Man" have become common labels to describe the first half of the 20th century. The concurrent assumption that the problems of the world's uprooted and homeless multitudes are so great as to be altogether insoluble does not, however, appear to be justified. It is a fact that by 1966 the number of uprooted persons the world over had been reduced to 11,000,000. But between the years 1967 and 1970 there was a new upsurge; by the end of 1970, a total of 18,173,011 refugees were counted by the United States Committee for Refugees; they were spread over more than 80 countries of Africa, Asia, Europe, Latin America, and North America. Their numbers declined by some 2,470,000 during 1971, however, and despite the enormous numbers displaced by the India-Pakistan war, the total was reduced to about 15,700,000. These millions became refugees in stages, group by group. Each refugee entity, while it bears some similarity to other groups, is unique. The reasons for displacement differ in each case, as do the circumstances of asylum and the opportunity for resettlement and integration.

TYPES OF REFUGEES

Refugees from advancing armies. The military, political, and ethnic convulsions of the 20th century have sent millions of uprooted and dispossessed swarming over the earth. World War I (1914–18), the Chinese-Japanese War (1937–45), World War II (1939–45), and their aftermaths triggered movements of populations far exceeding in scope the sufferings caused by mass displacements in previous centuries. Some of them were of a temporary and transient nature, such as flight from advancing or occupying enemy armies: Belgians fleeing from the German invaders; refugees from German-occupied provinces of France, Italy, and Romania; Turks from Greek-conquered eastern Thrace and Smyrna; and so forth. When peace was restored, most of these wartime refugees, as a matter of course, returned to their former homes.

Greek and Turkish refugees of 1923

Refugees expelled or evacuated by political authorities. The collapse of the Greek invasion troops in Turkish Asia Minor in September 1922 caused 1,000,000 of the area's indigenous Greek population to flee to Greece. A Greco-Turkish convention concluded in 1923 at Lausanne legalized the newly created situation, providing at the same time for the compulsory transfer to Greece of the surviving 189,000 Greeks in Asia Minor and for a compulsory transfer of the 355,000-strong Turkish minority in Greek Macedonia and Ípiros (Epirus) to Turkey. This exchange of minorities—a border case between a mass flight of desperate refugees and organized, government-supervised transfer—served as a pattern for other similar arrangements.

World War II refugees

During World War II and after, many peoples in eastern and central Europe were thus exchanged: Bulgarians, Romanians, Czechs, Slovaks, Ukrainians, Russians, White Russians, Germans, Poles, Hungarians, and so on. In February 1946, for instance, Hungary and Czechoslovakia agreed on a voluntary exchange of their respective ethnic minorities involving the transfer of 31,000 Magyars to Hungary and of 33,000 Slwaks to Czechoslovakia. The inhabitants of the Karelian Isthmus in Finland had the sad distinction of experiencing the refugee plight three times. When, in March 1940, defeated Finland was forced to cede Karelia to the Soviet Union, 415,000 Karelians left for Finland proper; in June 1941, Finnish troops recaptured the region, and some 305,000 started the return trek; Soviet Russia reconquered and incorporated Karelia in 1944, and the returnees for the third time became refugees, moving to their hard-tested homeland.

Political refugees. Politically motivated refugee movements, frequent in modern times, have been occurring intermittently since the earliest development of state entities powerful enough to oppress nonconformist minorities. The Russian Revolution of October 1917 and the postrevolutionary civil war (1917–21) caused the exodus of 1,500,000 opponents of the regime. Scattered throughout France (400,000), Germany (300,000), Poland (400,000), the Baltic States (100,000), parts of the Far East (100,000), and the Balkan countries, they refused to return to their Communist-ruled home country.

More than a million Armenians were deported or fled from Turkish Asia Minor to Russia, the Middle East, the Balkan States, and North and South America between 1915 and 1923; a million and a half are estimated to have been massacred by the Turks or to have perished as a result of forcible deportation.

Four to five hundred thousand Spanish Loyalists poured into France to escape the newly established regime of General Franco in the wake of the 1936–39 civil war. At first, able-bodied men of military age were sent back; later, they were admitted and interned in refugee camps. Those who had fled primarily from military action were repatriated to Spain after the end of hostilities. Many migrated to North Africa, Mexico, the Dominican Republic, and Latin and South American countries.

Following the abortive Hungarian Revolution of October 1956 and the subsequent Soviet occupation, some 203,000 Hungarians, mostly members of the liberal professions and skilled technicians, fled the country. In an unusual display of international concern and assistance, 172,000 of them were speedily and successfully resettled by the end of 1958 (18,000 eventually chose to return to Hungary). Between 1945 and 1961, when the Communist regime erected the Berlin Wall, a total of 3,700,000 political refugees from East Germany found asylum in West Germany. From 1961 until June 30, 1969, only 128,000 crossed the border illegally. Now, scarcely 20 people a month escape to West Berlin.

After 1960, some 750,000 Cubans—nine percent of Cuba's population—left their country and the regime of Fidel Castro; some 608,000 now live in the United States. Of the 450,000 registered as immigrants to the U.S., 380,000 had been, by the end of 1971, resettled in communities in all 50 states of the Union and Puerto Rico, 280,000 as resident aliens and nearly 100,000 had acquired U.S. citizenship. Some 48,000 Cubans found refuge in other countries, mostly Spain, where they are trying to obtain visas to the U.S.

When the People's Republic of China was established in 1949, some 2,000,000 Chinese, remnants of Chiang Kai-shek's army and civilians, escaped to the island of Taiwan, 121 miles off the mainland. Other sizable groups of refugees, totalling 340,000, went to such places as Burma, Vietnam, Japan, Laos, India, and Macau. The overwhelming majority, however, found sanctuary in the British Crown Colony of Hong Kong, just across the borders of mainland China. Together with early Chinese refugees who had left southern China before the advance of the Japanese armies, their number amounts to 2,065,000. The take-over of Tibet by People's Republic of China in 1959 resulted in the flight to India, Bhutan, Sikkim, Nepal, and Macau of 76,000 Tibetan refugees.

Ethnoreligious refugees. For many centuries, the refugee phenomenon was a corollary of religious and racial intolerance. Entire ethnoreligious entities were uprooted, exiled, or deported by secular or religious authorities in an effort to enforce conformity. In 1492 Spain expelled 150,000 Jews, who found asylum mainly in Turkey, Holland, and North Africa; 400,000 Huguenots had to leave France after the revocation of the Edict of Nantes in 1685 and found refuge in England, Germany, Holland, and South Africa. In the 1930s, as a result of the Third Reich's anti-Jewish policies, 400,000 Jewish refugees in all fled (1) from Germany proper (283,000), (2) from Austria (95,000), and (3) from the Sudetenland (23,600).

Israeli Jews

More than 50,000 Jewish refugees from Yemen were transported to Israel by "Operation Magic Carpet" in 1949–50, which had constituted almost the entire Jewish community of Yemen; all 7,700 Jews of the British colony of Aden fled, mostly to Israel, in 1947–60; 123,000 Jewish refugees were airlifted from Iraq to Israel in 1950–51; of the 29,770 Jewish nationals of Syria (as of 1943), more than 80 percent left the country as refugees to Israel, the United States, and Latin America. It was announced in mid-1969 that 97 percent of the Jewish population of Yemen, Aden, Iraq, and Syria had found asylum in Israel.

Refugees from territorial realignment. Several major refugee movements were caused by the partition of existing territorial entities. Defeated in World War II, Germany was compelled to disgorge all the territories she had gained since 1938; in addition, 117,000 square kilometres (about 44,300 square miles) of prewar German territory, with a prewar population of nearly 10,500,000, were detached and placed under Polish and Soviet administration. The Potsdam Conference in 1945 authorized the compulsory "transfer to Germany of German population or elements thereof" from Poland; authorization for the compulsory transfer of their respective German minorities was also given to Czechoslovakia and Hungary. Without applying for Allied approval, the Romanian government deported 165,000 ethnic Germans, and the government of Yugoslavia expelled 150,000. A total of some 12,000,000 refugees and expellees were dumped on the truncated territory of Germany, split into the Federal Republic of Germany (West Germany) and the German Democratic Republic (East Germany).

Hindus and Muslims

The partition of the Indian subcontinent in 1947 resulted in the two-way flight of 18,000,000 Hindus from Pakistan and Muslims from India—the greatest population transfer in history. With a dual religious and communal background, it is still continuing: between 1962 and 1971, 1,032,000 Hindus were reported to have been evicted from Pakistan, while 260,000 Muslims had reportedly fled or were evicted from India and Kashmir. Eight to ten million persons were temporarily made refugees by the creation of Bangladesh in 1971.

Palestine's partition in 1948 triggered an almost wholesale exodus of Palestinian Arabs in the wake of the military-political confrontation between the newly established state of Israel and the five neighbouring Arab states. The number of Arabs who had actually left the territory of Israel at that stage was variously estimated as between 419,000 (by Jewish sources) and one to two million (by Arab spokesmen at the United Nations).

Another major determinant of recurring refugee waves resulting from territorial realignments was the liquidation of European colonial possessions in Africa and Asia. The stream of British subjects from all parts of these continents—of French refugees from Morocco, Algeria, Tunisia, and Indochina; of Italians from Libya; of the Dutch from Indonesia—was a by-product of the progressive disintegration of the vast British, French, Italian, and Dutch colonial empires. The emerging independent regimes were intent on the speediest possible elimination of what they consider personified "vestiges of colonial domination." In one way or another, life became, or was made, unbearable for the European "colonials," even though many were second- or third-generation settlers, with deep roots in the local economy. In the North African countries, this trend engulfed not only nationals of the former colonialists but all resident minority groups of European origin. Tens of thousands of Greek and French refugees from Egypt, of Greek refugees from Sudan, of Italian refugees from Tunisia indicate that in the Arab-Muslim states along the Mediterranean coast and in parts of black Africa no room is left for communities that are neither Arab nor Muslim. A similar phenomenon has recently become noticeable in regard to the millions of people of Indian ancestry in several countries of Asia and Africa whose forebears had lived there for generations but who are now being compelled to "return" to India, leaving behind their family businesses and possessions.

Africa entered the orbit of refugee movements in the 1960s, becoming by 1963 a major centre of refugeeism, deserving particular attention because of the novelty and diversity of its refugee problems. In the past, black Africa's peoples moved freely within their often ill-defined tribal areas in search of game, more fertile soil, or forage for their herds. Existing national boundaries were no handicap for such intermittent or seasonal movements. But with the emergence, in the late 1950s and early 1960s, of several sovereign African states, these boundaries—arbitrarily traced lines on the map, often cutting across traditional tribal grounds — became barriers to population movements. They were jealously watched and often aggressively defended. To most Africans the notion of a modern nation-state with often artificial boundary lines is, however, still alien and incomprehensible. It is extremely hard for them to adjust to the new realities imposed by modern political developments. Many African refugees, cast adrift by this underlying incongruity, were abruptly thrust from an almost prehistoric way of life into new patterns of society totally unfamiliar to them. The situation in most African countries, the "to and fro" of warfare and revolution, is so volatile and precarious that additional refugee waves have to be considered a continuing probability. The power struggle in some new states, generated by an interplay of internal and external forces and tribal animosities has to be seen against the background of hunger, hope, remnants of colonialism, and great power tensions. Refugees in Africa increased from 860,000 at the end of 1968 to 1,858,000 by 1971.

Black Africans

The states represented in the 38-member Organization of African Unity (OAU) have shown a remarkable degree of goodwill and understanding for the rights and interests of their respective refugees. The Conference of Heads of State and Government of OAU, meeting in Addis Ababa in September 1969, approved a Convention Governing the Specific Aspects of the Problem of Refugees in Africa. One of the convention's key provisions established the principle that "no person shall be subject by a Member State to measures such as rejection at the frontier, return or expulsion which would compel him to return or to remain in a territory where his life, physical integrity or liberty would be threatened." It was the first time that such a principle had been enunciated in a binding international legal instrument. A second valuable innovation was the stress on the necessity of handling refugee problems in a peaceful and humanitarian spirit. The convention specifically enjoined a refugee from engaging in any subversive activity against any member state of the OAU while obliging the countries of asylum to prohibit refugees from attacking any member state through subversive activities, "especially through arms, press and radio." It also provided that "for reasons of security, countries of asylum shall settle refugees at a reasonable distance from the frontier of their country of origin."

THE CONDITIONS OF REFUGEE MOVEMENT AND SETTLEMENT

Refugees on the move. Only in a few exceptional cases have displaced refugee groups been able to reach their respective destinations speedily and with a minimum of hardship. In general, most refugees are traditionally on the move for long periods of time and under the most trying circumstances.

In April–May 1945, in the final phase of World War II, for example, more than five million Germans rushed from the provinces of Silesia, East Prussia, East Brandenburg, and the German-occupied West Polish areas to escape the advancing Soviet and Polish armies. They dragged along with them horses, cattle, and cars, packing the roads and fighting their way across the Oder River. In the summer of that year, after the conclusion of military operations, large numbers of these refugees began remigration through the still open frontiers to their abandoned homes. In the winter months of 1945 and in 1946–47, the entire German population of the former German eastern provinces was forcibly expelled in accordance with the Potsdam decision of the Big Three. The aggregate number of casualties during this three-stage, back-and-forth mass movement was variously estimated as between one and two million.

Similar, in 1947, was the plight of the 18,000,000 Hindu and Muslim refugees fleeing to their respective national countries. Mostly on foot, they were on the move for extremely long periods, flooding every highway, road, and cow path linking Pakistan with India. Some of their caravans were giant, unwieldy columns of hundreds of thousands of people, crawling for weeks at the slow pace of the bullock carts and herds of cattle. Casualties — killed outright or dying from starvation and exhaustion — were estimated to be nearly 10 percent of the total.

Displaced persons' camps. In the wake of World War II, hundreds of thousands of refugees and displaced persons, deported by the Germans or fleeing from the advancing Soviet armies, were housed in Allied-supervised camps, mostly in Germany, Austria, and Italy. By March 31, 1946, the UN Special Committee on Refugees and Displaced Persons estimated their number at 1,100,000. That the total figure remained the same in September 1947, in spite of repatriation and an initiated resettlement, was due to the influx of Jewish refugees from the east. The 100,000 Jewish displaced persons who lived in the camps at V-E Day became 250,000 in 1947; departures were offset by a high birth rate.

The situation in the camps grew increasingly strained. Morale declined and patience ran thin among people uncertain about the possibility of their resettlement. Many Jewish camp inmates took recourse to illegal immigration into British-occupied Palestine. Of those intercepted by the Royal Navy, 51,500 were put behind barbed-wire in hastily improvised new special camps on the island of Cyprus. The emergence of the state of Israel in May 1948 opened the channels of migration to Jewish displaced persons. By the spring of 1959, only 21,000 hard-core refugees remained in the European camps.

Vietnamese

But such camps are still very much in evidence in Asia. As late as June 1969, the large majority of refugees in South Vietnam were housed in temporary shelters, which are loosely described as "camps." Their number was estimated by a committee of the U.S. Senate to approach 4,500,000 in mid-1972. Many of them, having been among the 900,000 who left North Vietnam at the time of the partition in 1954, have become refugees for a second time. As a result of Arab-Jewish hostilities in June 1967, some 350,000 Arab refugees were newly displaced and fled to east Jordan, Syria, and the United Arab Republic. Some 130,000 of this total were already refugees from the 1948 conflict and were uprooted for the second time; new tented and hutted camps had to be hastily established to house them.

Refugee resettlement. The bulk of the refugees resettled since the end of World War II were integrated by their respective conational governments. In nearly all cases, the theme was basically "do-it-yourself." The U.S. Marshall Plan aid to Germany and the massive UN financial assistance to South Vietnam and South Korea were applied to the general recovery of these countries devastated by war; nonetheless, such aid has but indirectly facilitated refugee integration. In 1960 Pakistan's President Mohammed Ayub Khan pointedly noted that his country had received 9,000,000 Muslim refugees from India as compared with only "three-quarters of a million refugees from Palestine" in Arab lands; his people had had to face the problem themselves without "substantial support from Muslim brethren over the world." Similarly had India, while accepting some aid from many private organizations the world over, borne the major burden of rehabilitating its millions of refugees. As a rule, there was little or no direct contribution on the part of the international community for the specific purpose of refugee rehabilitation and resettlement.

The sole exception is the Palestinian refugees in Arab countries who have been the wards of the international community for more than two decades. They are commonly considered to be the only refugee group firmly resisting integration. Backed by the governments of the Arab host countries, they insist on return to Israel as the only solution to their problem.

Rehabilitation and integration of refugees

Integration of refugees in the countries of their resettlement, no matter how advanced, is, however, in no way equivalent to the restoration of their previous economic, occupational, and social status. The very idea of "integration" is by definition relative. Economic recovery does not equal genuine rehabilitation, any job does not necessarily mean integration. Among those employed are many whose new status is lower than their predisplacement level and whose income compares unfavourably with what it used to be. Most refugees had to start working in poorly paid jobs in fields not to their liking; many who had worked independently have had to accept salaried positions. This was especially the case of middle class refugees —the fortunate exceptions being members of the free, liberal professions such as medicine, law, and pharmacy.

The character and pace of the integration process largely depend on the general standard of life in the receiving country. In such highly developed countries as The Netherlands, Great Britain, the United States, West Germany, and France, even a successfully integrated refugee group usually needs much time and luck to reach a position approaching that of the natives. In underdeveloped countries, on the other hand, where the majority live in poverty, the lowest common denominator is attained more speedily and easily, the initial discrepancy between refugee and native being almost unnoticeable.

Refugees in a new culture. Refugees resettling in countries with a culture different from their own have always faced problems of adaptation to an unfamiliar way of life and to an alien language and culture.

Many post-World War II refugees in Europe were survivors of concentration camps and slave labour. Their emotional state was a mixture of bewilderment, anxiety, and shock, combined with gratitude for deliverance and an intense desire to shake off the past and make good. Resettlement took two principal forms. One was group resettlement under organized programs conducted by individual governments and by the International Refugee Organization; during the IRO's first year of operation (July 1, 1947 to June 30, 1948) 156,925 persons were brought to 73 countries, mostly to Great Britain, Canada, Belgium, and the United States. The other form was individual resettlement, whereby the IRO, in consultation with voluntary agencies, arranged immigration for individuals and small groups of refugees (a total of 47,652 for the same period). Those arriving as individuals usually felt utterly lost in the new surroundings; they had to begin rebuilding their lives on their own, from scratch. For many, the process of acculturation was painfully slow. In the initial stages most of them understandably tended to settle in the neighbourhoods of earlier immigrants from the same countries of origin and to seek employment in identical or similar professions. Organizations of former fellow countrymen were often instrumental in finding them jobs or dwelling accommodations. This attachment to existing ethnocultural communities mitigated the new arrivals' feeling of loneliness and smoothed the first stages of their economic adjustment, yet it also had a retarding influence on their assimilation.

REFUGEES AND THE WORLD COMMUNITY

International and intergovernmental agencies. Paul Weis, a UN expert on refugee matters, has drawn attention to the fact that though the refugee problem is as old as history, international action for refugees did not start until after the Russian Revolution.

League of Nations

In 1921 Fridtjof Nansen of Norway was appointed by the League of Nations as high commissioner for Russian and Armenian refugees and devised a so-called League of Nations Passport ("Nansen Passport"), a travel document recognized "in principle" by 53 states. It restored their identity to its beneficiaries and made them somewhat less vulnerable to arbitrary action by the authorities of the countries of their sojourn. After Nansen's death in 1930, the protection of refugees was entrusted by the League to the Nansen International Office for Refugees, established as an autonomous body with a mandate to wind up its functions by 1938. In 1933, consequent to Nazi racial persecution, James G. Macdonald was appointed the League's high commissioner for refugees, largely concerned with Jewish and other refugees from Germany. He resigned after two years because of lack of cooperation from the League's member states; his successors were restricted to purely legal protection of the refugees, and their efforts were just as fruitless.

In an attempt to substitute planned migration of refugees for their chaotic dispersion, President Franklin D. Roosevelt of the United States, in July 1938, invited 32 governments to a conference at Évian, France. Only the Dominican Republic, however, offered to admit 100,000 refugees for rural settlement; no other country responded. Nevertheless, an Intergovernmental Committee on Refugees (IGCR) was organized, with a seat in London, to look for new areas of settlement for nonrepatriable refugees; its efforts proved to be unsuccessful.

A conference of 44 nations, convened in Washington, D.C., in November 1943, resulted in the signing of the charter of a new international body, the United Nations Relief and Rehabilitation Administration, whose main task was postwar reconstruction of devastated areas, as well as help for and repatriation of uprooted people, both refugees and displaced persons. By the time of its dissolution in 1947, UNRRA had repatriated nearly 7,000,000 anti-Nazi and anti-Fascist refugees, but there still remained, in 920 refugee camps of Europe, a hard core of 1,600,000 unrepatriables of 52 nationalities; a count made in 1964 revealed that 95 percent of the camps' remaining small populations did not want to be repatriated.

IRO

In 1947, both the IGCR and the League's high commissioner were replaced by a nonpermanent specialized agency of the United Nations, the International Refugee Organization (IRO), whose mandate expired in 1952, while there were still in Europe alone some 15,000,000 unsettled refugees. On the initiative of the United States, 16 nations met in Brussels in December 1951 and set up the Intergovernmental Committee for European Migration (ICEM), which was not to be a UN agency but would be "responsible for the movement of migrants, including refugees, for whom arrangements could be made with the governments of the countries concerned." By 1971 the number of sponsoring nations in ICEM reached 31. Since the beginning of its operations in 1952, ICEM has processed and moved 1,775,000 European refugees to resettlement countries of their choice. Member governments have contributed nearly $400,000,000 for ICEM's activities. Particularly successful was its handling of the sudden refugee influx in the wake of the Hungarian Revolution in 1956: within six months after the first refugee crossed the Austrian border, ICEM had moved over 160,000 Hungarians to all parts of the world.

UNHCR

In 1950, shortly before ICEM was created to take care of the refugees' transportation and resettlement, the Office of the United Nations High Commissioner for Refugees (UNHCR) was established to ensure the legal and political protection of refugees previously under IRO mandate and to promote permanent solutions to their problems. By ensuring that he is not sent back to the country of his origin against his will, the UNHCR is enabling the refugee freely to choose either to return voluntarily to his country of origin, to remain permanently in the country of first asylum, or to resettle in another country depending on the possibilities made available by governments. Sixty-seven of the UN member-states support the UNHCR financially, but the agency budget is still inadequate.

In addition to these bodies dealing with the general problems of refugeeism, several regional agencies have been established. To provide for the needs of Arab refugees from Palestine, the UN General Assembly in December 1949 established the United Nations Relief and Works Agency for Palestine Refugees (UNRWA), which began operations in May 1950. In 1951 the United Nations Korean Reconstruction Agency (UNKRA) was founded, and in March 1952, the U.S. Escapee Program (USEP), with a central office at Geneva, was set up to assure for escapees from behind the Iron Curtain adequate opportunities in the free world.

Nongovernmental voluntary agencies. All international or intergovernmental agencies established on behalf of refugees were conceived as temporary organizations to deal with each refugee problem as it arose; consequently, international agencies have succeeeded each other frequently. The only ***permanent*** relief work has been conducted by the numerous voluntary agencies, national and international, which have become increasingly active in the last several decades. Their permanency enabled them to pursue and attempt to conclude the tasks that the ephemeral intergovernmental bodies were unable to complete. By the end of 1947, the IRO had concluded agreements with 128 voluntary organizations belonging to three major categories. First were denominational organizations that extended support to refugees of their own faith, such as the National Catholic Welfare Conference, the World Council of Churches (for Protestant refugees), the American Jewish Joint Distribution Committee, the Jewish Agency for Palestine, and the Hebrew Immigrant Aid Society (HIAS). Second were the worldwide nondenominational organizations such as the International Red Cross, the International Social Service, the International Rescue Committee, the Organization for Rehabilitation and Training, and the YMCA. Third were a number of private organizations, mostly national groups, working almost exclusively on a local scale. In Korea alone in 1970, there were 77 private organizations from 13 countries spending the equivalent of $13,000,000 a year to sustain programs of wide variety.

In 1962 the International Council of Voluntary Agencies (ICVA) was formed. By 1969 it had a membership of some 100 organizations and a central bureau in Geneva. In November 1969, there convened in Washington, D.C., a National Conference on World Refugee Problems, which attempted (1) to arrive at a common definition of the term refugee and of the criteria for determining needs and services; (2) to design patterns of interagency cooperation in refugee assistance programs; and (3) to elicit citizen participation in overseas assistance programs.

The problems of assimilation. In pre-World War II Europe and during the early postwar years the world over, a refugee's two major predicaments were (1) his practically extralegal status as a "stateless person," living on sufferance in his countries of refuge, and (2) his sojourn in an ethnically and culturally alien milieu. This twin disadvantage is now endured by only a minority, though a sizable one, of the refugee mass; as late as June 1962, a tabulation by the United States Committee for Refugees established that about 200,000 persons were still stateless in Europe, as against nearly a million in 1957; the bulk of the stateless in Asia comprised the 1,270,000 Chinese from mainland China, the 70,000 Tibetans, and the 532,625 Palestinian refugees in Syria, Lebanon, and the Gaza Strip; there were 320,000 stateless Africans in the Congo, Burundi, Tanganyika, and Togo; 175,000 Cubans were at that time stateless in the Western Hemisphere—a total of 2,567,625.

UN Convention relating to the Status of Refugees

The 1951 United Nations "Convention relating to the Status of Refugees," which granted a minimal legal status to refugees falling under the terms of that instrument, by January 1972 had been ratified by 62 states; in a number of other countries, the question of accession was under consideration; the United States was still withholding ratification.

The rights granted to refugees under the convention can be divided into (1) those relating to the refugee's needs as a member of the community of his country of residence and (2) those relating to his specific status as a refugee. The first group of rights involves the exercise of religion, economic pursuits in general, wage earning and independent employment, taxation, public relief, the acquisition of property, education, civil rights (access to courts), and freedom of movement. To the second group belong the personal status of the refugee (legal capacity, family rights, succession and inheritance), the delivery of documents relating to his personal status (which cannot be obtained from his native country), the issuance of identity and travel documents, treatment of refugees who enter or reside unlawfully in a country, the possibility of their expulsion, the question of their naturalization, and the exceptional measures taken in time of war or other emergency against nationals of the country of which the refugee was, or formally may still be, a citizen. Despite all the recent improvements, however, the situation of the residual stateless refugee is still abnormal and unenviable.

The predicament of statelessness was spared the great majority of the 40,000,000 people who became refugees during the period after 1945. The 18,000,000 Hindus and Muslims involved in the postpartition exodus from India and Pakistan suffered greatly; yet they enjoyed from the very beginning, in both countries, the security of full citizenship and the protection and encouragement of the respective conational state in which they found a new life, not on sufferance but as of right. Full citizenship rights were also granted to the 12,500,000 expellees in Germany, the 400,000 Karelians in Finland, the 250,000 Indonesian Dutch repatriates in Holland, the millions of refugees from North Vietnam and North Korea, the more than a million Jewish refugees in Israel, the 1,060,000 refugees from North African Arab-Muslim states in metropolitan France, and 300,000 Italian repatriates from the provinces of Dalmatia and lstria ceded to Yugoslavia. Of the Palestine Arabs on UNRWA's relief rolls, 989,000 were in 1967 full-fledged citizens of Jordan.

Apart from statelessness, the other common feature of the refugees' classic plight was that they usually arrived defenseless and helpless in a strange country with a strange language and other customs, the bewildered guests of embarrassed hosts who had no obligation toward them except that dictated by common humanity. Yet every refugee group enumerated above went to a country with whose people they were linked by common language, religion, and customs. In the words of Elfan Rees, they have all found asylum among their own kind:

> They are strangers but not among a strange people. Community of language, faith, culture, and social organization provides a cushion that both minimizes the shock of exile and increases the chance of assimilation.

Major waves of refugees have rarely been officially welcomed in the countries of refuge, however—the only exception being Israel, for whom the "ingathering of the exiles" is a national policy. The more than a million refugees from mainland China were a nearly unbearable burden for overcrowded, British-governed Hong Kong, and sporadic attempts were made to stem their influx forcibly. But even when the newcomers were ethnic brethren of the resident population (as was the case with the eastern Germans expelled to Germany, the Algerian French returned to metropolitan France, and the Dutch repatriates from Indonesia in The Netherlands), there was noticeable initial friction between the incoming refugees and the local residents. The West German government strongly resented the imposed necessity of accepting and resettling millions of forcibly expelled Germans from Czechoslovakia, Poland, and Hungary; it considered that the decision of the Potsdam Conference made the Big Three, which endorsed this compulsory transfer, morally and politically coresponsible for the refugees' predicament. There was also initial bitterness in both India and Pakistan in the wake of the frantic two-way stream of refugees, although both governments eventually accepted the new states of affairs created by the de facto exchange of population. After the first shock subsided, there was no talk about the return of the Hindus and Muslims to their previous homes. The national efforts were fully concentrated on their resettlement and integration, on abolishing their refugee status. "Do you intend to go about all your lives with the word 'refugee' inscribed on your foreheads?" Prime Minister Jawaharlal Nehru admonished India's new citizens. "This word connotes dependence and helplessness—the sooner you get rid of it the better."

A refugee who has reached his chosen country of resettlement legally has ceased to be a "refugee," both in law and in fact: his status and problems, for all practical purposes, have become identical with those of an immigrant admitted on a regular consular visa. Only refugees who have settled among people of their own culture and ethnic nationality have sometimes been singled out for special preferential treatment. In countries of mass immigration (such as the United States, Canada, and Australia) they have been treated, on the whole, not better and not worse than the average immigrant.

BIBLIOGRAPHY. E.M. KULISCHER, *Europe on the Move: War and Population Changes,* 1917–47 (1948), discusses the role of migration in world history, especially its connection with war, with examples; M.J. PROUDFOOT, *European Refugees:* 1939–52 (1956), a study in forced migrations before, during, and after World War II; J.B. SCHECHTMAN, *The Refugee in the World: Displacement and Integration* (1963), traces the origin, development, and attempts at solutions of each significant Old World movement between 1945 and 1963; F.D. SCOTT (ed.), *World Migration in Modern Times* (1968), essays dealing with migrations in all parts of the world; J.G. STOESSINGER, *The Refugee and the World Community* (1956), an analysis of the problem and the response of the world community to it, particularly the role of the International Refugee Organization; L.W. HOLBORN, *Refugees: A Problem of Our Time,* 2 vol. (1975), a description of the activities of the United Nations High Commission for Refugees from 1951 to 1972 with historical background. For current data, see U.S. COMMITTEE FOR REFUGEES, *World Refugee Survey Report* (annual).

(J.B.Sc.)

Refuse Disposal Systems

The term refuse refers to solid wastes, and the two are used more or less synonymously to describe those discards of society that are not liquid or gaseous in nature. Solid-waste management is the development and operation of refuse disposal systems designed to handle community refuse in a healthful, economic, and conserving manner. The amount of solid wastes produced can be correlated to the output of goods and services; in the late 1970s the United States, which consumed nearly half of the world's annual production of industrial raw materials, produced more than 600,000,000 tons of solid wastes each year, excluding wastes from agriculture and mining. Inclusion of wastes from agriculture and mining increases the total more than sevenfold (see Table 1).

A comparison of the average output of refuse per family (3–4) per week in eight countries is shown in Table 2.

Table 1: Solid Wastes in the United States (1977)

waste source	amount (lb/capita/year)*
Municipal	
Residential, commercial, and institutional	1,340
Sewage sludge	46
Automobiles and construction demolition	410
Industrial	3,330–3,680
Radioactive	0.4
Mining and milling	21,210
Agricultural	23,030–30,640
Utility (electrical)	710
Total	50.076–58,036

*1 pound = 0.45 kilogram
Source: U.S. Environmental Protection Agency, *Solid Waste Data,* 1981.

These figures indicate the effect of degree of industrialization on family refuse, entirely apart from the waste contributed by industry itself.

The character of refuse varies considerably. The ash content of refuse in the United Kingdom is 20–30 percent, whereas in the United States it ranges from 5 to 20 percent. Paper and paper products constitute by weight about 30–50 percent of the U.S. and Canadian household refuse, and in Europe they reach as much as 40 percent.

Table 2: Generation of Household Refuse per Family per Week

country	weight (pounds)*
Canada	48
Czechoslovakia	46
France	37
Israel	31
Poland	26
Spain	29
United Kingdom	35
United States	53

*1 pound = 0.45 kilogram

There are three environmental depositories for society's discards—the air, the water, and the land. Air and water pollution control has historically been based on control or treatment of effluents discharged into the air or water. There has been little effort in the past to control the production of solid waste. Rather, control programs have centred on methods to salvage that part economically feasible and to reduce the remainder by burning or compaction to minimize transport costs and the environmental space required to harbour it.

Problems of solid-waste management

Refuse collection and disposal is nearly universally regarded as a responsibility of local government and as a major public health and welfare service. Improperly handled refuse serves as a breeding ground and food supply for flies and rats. The former practice of feeding raw garbage to swine has been shown to be a link in the transmission of trichinosis to humans. Outbreaks of the virus disease of swine, vesicular exanthema, which is similar to foot-and-mouth disease, have resulted in the requirement that garbage fed to swine be heat-treated and that the practice of raising hogs on garbage dumps be discontinued. Heat treatment of garbage has been required in the United Kingdom since the early 1900s.

Most communities in Europe and the United States no longer permit leaf burning. Smoke and odours from open burning dumps and particulate matter from stacks of improperly designed or operated incinerators are sources of pollution. Furthermore, rubbish has been found to be a significant factor in causing fires in buildings.

Improper disposal of refuse has resulted in pollution of surface and ground water. Indiscriminate dumping in pits or on river banks is not uncommon. A problem particularly acute in the United States is the littering of highways and roadside areas. Materials jettisoned in the United States have been estimated to comprise a bulk of 20,000,000 cubic yards (15,300,000 cubic metres) annually.

Besides population growth, a number of other factors have contributed to the world's increased solid-waste problem: design of products leading to accelerated obsolescence; higher real income resulting in the increased manufacture and sale of consumer items; and packaging to improve sales appeal, prolong shelf life, and reduce marketing cost per unit.

The refuse collection and disposal function is ordinarily grouped with other municipal sanitation operations as a major department of local government. Traditionally, collection and disposal service has been oriented to a single-city institutional arrangement. In most metropolitan areas political fragmentation is a serious problem. Regional groupings have been suggested as an avenue of improvement, particularly for the disposal function.

MODERN COLLECTION AND DISPOSAL METHODS

Most cities require household garbage be well wrapped and stored in durable, easily cleaned containers with tight-fitting covers. Ashes are stored in metal containers. Plastic or paper bags as container liners have come increasingly into use, particularly in commercial food preparation areas and institutions. Many new multiple-dwelling units are equipped with refuse compaction systems. A pneumatic pipeline refuse transport system, first installed in a high-rise development in Stockholm in the late 1960s, has come to be used widely in Europe.

Dustless systems

Collection and disposal in urban areas. In some urban commercial areas bulk containers are mechanically lifted into large compactor-equipped trucks for transport to the disposal site. In Europe, because of the ash content of household refuse, service provided by some municipal authorities features dustless systems, wherein heavy metal containers are mechanically dumped into the collection vehicle through a tight-fitting portal designed to accommodate standard bins. A decrease in the ash content, however, has resulted in greater use of smaller, lighter containers, including plastic bags.

Household refuse collection in both the United Kingdom and the United States is generally characterized as a combined collection of household food wastes and rubbish from containers placed at the house line, alley, or curb. Loading is done by hand, but research has been undertaken to reduce labour requirements through increased mechanization. Household refuse collection vehicle design incorporates a rear-, front-, or side-loading closed metal body equipped with a mechanically or hydraulically operated compaction plate to increase the load capability. This has become more important as travel distances to disposal sites have increased, and in many large cities transfer of refuse from neighbourhood collection vehicles to long-haul vehicles, which can carry 60 or more cubic yards of compacted refuse to disposal facilities, is not uncommon. In most of the world household refuse-collection systems are operated by local government authorities. By contrast, in the United States private contractors and private systems provide almost half of the household and the majority of the commercial and industrial service. Frequency of collection varies from daily to less than once per week. Reported collection costs vary widely, depending on such factors as haul distance, type of service, climate, and wage levels.

Incineration. Combustible wastes may be reduced to inert residue by high-temperature burning. An incinerator is composed of a furnace into which the refuse is charged and ignited, a secondary combustion chamber in which burning at a high temperature is continued to complete the combustion process, and flues wherein the gases of combustion are cleansed as they are conveyed to a chimney and thence to the atmosphere. Incinerator plants also include facilities for unloading and storing refuse for short periods, to permit uniform charging of the furnaces, and a building to house the incinerator appurtenances.

The growth of large cities has been primarily responsible for the development of refuse incinerators because it was not practical from nuisance, public health, and aesthetic viewpoints to permit refuse dumps in their midst. The burning of refuse provided an efficient means of inoffensively reducing the bulk of the refuse (about 95 percent by volume) to a readily transportable and often usable ash. Municipal refuse contains a great variety of combustible items; some require special treatment before processing through a conventional incinerator, and air pollution agencies have come to regulate the use of such incinerators. Bulky refuse may be preprocessed by size reduction through such equipment as shredders, hammermills, and impact mills. Many special incinerators have been designed for bulky refuse. Hazardous and obnoxious wastes, such as highly volatile dusts and flammable liquids, are also usually disposed of in specially designed plants.

Size reduction

On-site incineration of combustible refuse is an economical method for many industries, apartment dwellings, and large commercial establishments, as well as in certain instances for householders. The basic furnace design for such units is similar to that required for their larger municipal counterparts. Apartment house incinerators, which are often flue-fed with a chute directly to the furnace, have been found to be a significant source of air pollution; their design and operation is stringently regulated.

Waste heat utilization

Heat from incinerator furnaces may be used to generate steam for use in industrial processes, for space heating and power generation, or in sewage disposal plants and other public facilities. The economics of the use of waste heat for steam generation depend primarily on the steam requirements of users within a reasonable distance of the plant and the market rates. A number of European cities such as Paris, Munich, and Frankfurt am Main operate steam-generation plants in conjunction with municipally-operated power plants. Montreal has a steam-generating plant that began operations in 1970. In such arrangements the price of coal is a crucial factor.

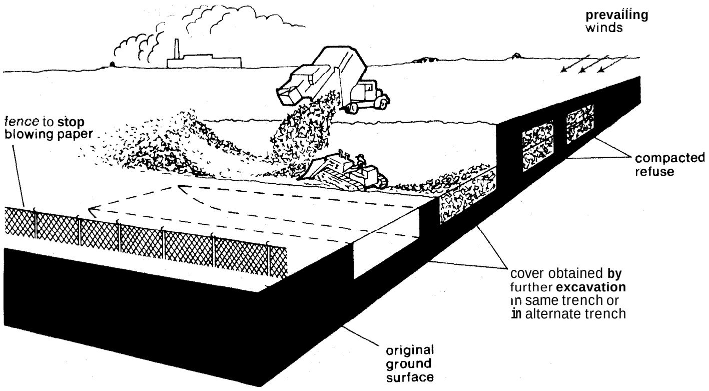
Disposal into sewerage system. Garbage ground into minute particles can be discharged into the community sewerage system. Little difficulty is encountered in a properly designed sewer system as a result of this, and household water consumption increases only 1 to 2 percent with the installation of a grinder. Increases in solids-handling capacities are needed at the sewage treatment plants in cities where large numbers of grinders are in service, however, and for this reason some communities with overloaded treatment plants have forbidden their use. They have, however, found increasing application in commercial and institutional food-handling establishments.

Ocean disposal. The practice by coastal cities of barging garbage and rubbish to sea has come under critical pressure as the lighter material finds its way back to the beaches. New York City, a major offender, was prohibited from this practice in 1933 by the U.S. Supreme Court. Nevertheless, by the late 1960s an estimated 50,000,000 tons of solid wastes, including dredging spoils, ship refuse, waste oil, industrial chemicals and sludges, and sewage sludge were annually disposed of at sea by the United States. Of that total 30,000,000 tons were dredging spoil. Some resort communities along the Atlantic coast pump sewage into the ocean. Similar practices on a lesser scale are worldwide; guidelines for regulation of ocean disposal have become a matter for United Nations action.

The sanitary landfill method

Disposal on land. Indiscriminate dumping of refuse on land is an age-old practice; in the late 1960s more than 90 percent of land disposal sites in the United States were classified as "dumps." The unsightliness and unavoidable presence of flies, rats, smoke, and odours at dump grounds led to the development of the sanitary landfill method (see illustration). A sanitary landfill is defined by the American Society of Civil Engineers as a method of disposing of refuse on land without creating nuisances or hazards to public health or safety, engineered to confine the refuse to the smallest practicable area and volume and to cover it with a layer of earth on at least a daily basis. It requires the same careful preliminary and operational planning and design that any engineering construction job must have to be successful. Site selection includes zoning arrangements, accessibility, haul distance from collection routes, availability of cover material, a study of geological formations to assess water pollution hazards, and the ultimate land use plan for the completed site.

By courtesy of the U S Environmental Protection Agency


prevailing winds
fence to stop blowing paper
compacted refuse
cover obtained by further excavation in same trench or in alternate trench
original ground surface

Trench method of sanitary landfill used to dispose of refuse in a flat land area.

The most commonly used equipment on sanitary landfills is the bulldozer or rubber-tired tractor. It performs the spreading, compacting, covering, and trenching operations; and in many instances it hauls the cover material. Other equipment includes scrapers, graders, compactors, draglines, and water sprinklers for dust control. Sanitary landfills serving up to 50,000 people, or handling up to 115 tons of solid wastes per day, may operate with one piece of equipment. Suitable land in urban areas has become increasingly scarce. Space requirements almost doubled between 1950 and 1970 (0.7 to 1.25 acre-feet [863 to 1542 cubic metres] per thousand persons). Intensive work has been under way to investigate the feasibility of long-distance haul of refuse by rail to remote disposal sites, such as abandoned strip mines. Reported sanitary landfilling costs range from $1.25 to $5.00 per ton of refuse disposed of depending on the size of the operation, land and site preparation costs, operational requirements, and the ultimate land use construction requirements.

Completed sanitary landfill sites are most commonly used for recreational areas, such as parks, playgrounds, and golf courses. An arboretum has been constructed on a Los Angeles County landfill site. Heavy construction on completed sanitary landfill sites, however, has generally been avoided.

Shredding automobiles into scrap

Abandoned automobiles. The problem of the derelict automobile has developed into a serious one in the United States and other countries. Large shredders that process automobile hulks into scrap were developed in the late 1960s, but despite the processing of more than 8,000,000 U.S. vehicles annually, it has been estimated that the number of unprocessed abandoned vehicles increased from about 6,500,000 in 1965 to as many as 20,000,000 by the late 1970s. The state of Maryland was a pioneer in government action to encourage the salvaging of automobiles when, in 1969, it passed legislation that provided a bountry for automobile hulks brought to regional processing centres.

Collection and disposal in rural areas. The collection and disposal of refuse in rural areas presents a special problem because sparse population and small communities result in high unit costs for transportation and disposal. In industrialized countries rural per capita generation of refuse is nearly as great as that in urban areas; the failure to provide services often results in indiscriminate dumping and littering.

RECLAMATION AND RECYCLING

The removal of salvageable items such as metals and paper from solid wastes–for direct reuse or as raw materials for industrial reuse is termed reclamation and recycling. The amount and kind of refuse salvaged from the solid-waste stream has traditionally followed economic patterns. Before World War II a number of cities operated garbage reduction plants to reclaim grease and produce tankage that was sold to reduce the cost of collection and disposal. Picking belts were operated to remove items such as metals and cardboard for resale. Rising labour costs, difficulty in obtaining labour for the separation process, and lack of markets for grease contributed to the closing of these plants.

Composting

Composting is the biochemical alteration of organic refuse from a noxious conglomerate to an innocuous and usable soil humus. A number of anaerobic and aerobic processes have been tried on varying scales in the United States. None has been successfully employed in U.S. communities on a continuing basis, primarily for economic reasons. Composting has been successfully undertaken in Europe, particularly in France, where more favourable economics prevail, but even in The Netherlands, where reclamation provides much of the arable land, the practice of cornposting refuse has been diminishing as other more economic methods of supplying needed humus to the soil are adopted. While limited markets and related economics have discouraged the salvage of the quality materials contained in household refuse, large quantities of industrial and commercial solid waste are recycled.

A continuing sharp rise in per capita generation of solid wastes has occurred as a result of increased materials production. For example, beverage containers produced for "consumption" by the U.S. populace doubled from 30,000,000 in 1965 to 63,000,000 in 1976. The productive energies of both industry and government have largely been centred on the goal of increasing the recycle of this waste material through better technology and the development of economic incentives to encourage re-use of the materials. Research has included utilization of fibrous wastes as sources of nutrients, laser-mediated lignin solid-waste fermentation, and the design of a water-disposable packaging container. Major governmental attention has been given to finding means of recovering materials and energy from solid waste. In addition, several states have enacted laws to reduce beverage-container refuse by placing a cash deposit on all containers, which is refunded when the containers are returned.

STREET CLEANING

Sweeping. Until recent times city streets were depositories of garbage and all forms of refuse, a situation that contributed to widespread epidemics. The paved streets of the 20th century, together with the development of mobile, powered cleaning equipment, have led to the establishment of street cleaning as one of the major tax-supported efforts of municipal government. Effective street-cleaning programs not only enhance the appearance of the community, but they protect the public health by reducing disease and injury from dirt entering the eyes, ears, nose, and throat; and they promote safety by eliminating causes of skidding, fire hazards, and sources of water pollution in storm runoff. Flushing and hand sweeping have long been supplemented by mechanical sweeping; by the 1970s vacuum techniques had been introduced in many European cities.

Vacuum sweeping

Snow and ice removal. Except in far northern regions, most cities treat snow and ice storms as emergencies, marshalling all forces available until traffic is moving freely, at least along main arteries. Many factors affect the impact of a storm on a community, including the amount and dryness or wetness of the snow, timing, rate, and duration of the storm, wind conditions, and the temperature. The emergency nature of snow and ice control work requires careful planning, organization, and operation; reliable weather forecasting service is essential. The four general categories of snow removal efforts used to combat snow storms in urban areas are: salting and abrasive spreading; plowing; physical removal (or lifting) of snow from business and commercial areas; and special removal operations in such areas as crosswalks and sidewalks, parking areas, fireplugs, and bus stops. Spreading chemicals (most commonly rock salt) or sand, or both, is the customary first line of offense against snow and ice. Salt has no effect, however, when the temperature drops below −6° F (—21" C), the temperature at which a concentrated solution of saltwater will freeze. Many cities start plowing operations before a depth of 1½ inches (3.8 centimetres) of snow has fallen, particularly if a severe storm has been forecast. Plows are mounted on various types of vehicles, including dump trucks, wheeled tractors, refuse collection trucks, and road graders. Where snow must be removed completely, mechanical loading equipment, such as elevating conveyors and front-end loaders, is used. Vehicles equipped with rotary plow blades, called snow blowers, are widely used to load hauling vehicles or to remove large accumulations of snow from highways or airport runways. Disposal of snow that is loaded and hauled away is accomplished by dumping it into large bodies of water or large sewers. Unlimited amounts of snow ordinarily can be dumped into large lakes, rivers, or in the ocean without undesirable side effects. Sometimes snow is dumped on vacant land. Snow melting is expensive but is used in some cities where the quantity of snow, haul costs, and other alternatives make it economically attractive. Montreal undertakes to clear all snow within 72 hours of a snowfall and not only maintains a large snow "tip" (dump) but operates several batteries of snow melters capable of melting 560 tons per hour. The method reportedly costs twice that of dumping and four times that of disposing of it into convenient sewers.

Snow blowers

SPECIAL PROBLEMS

Solid wastes include many types of refuse that constitute special and unique problems. The wastes may be hazardous, for example, radioactive wastes, toxic chemicals, and pathogenic wastes from hospitals or research laboratories, or they may be otherwise unique in nature, such as demolition wastes from urban renewal projects and ash by-products from the energy industry. All of these materials require special handling.

Concern over the control of radioactive wastes is worldwide, but no uniform policy for their handling and disposal has been formulated. Since the late 1970s, the U.S. Department of Energy has maintained control of the containment and disposal of nuclear wastes, a task originally performed by the former Atomic Energy Commission. Although the federal government has begun to delegate some authority to certain state agencies, it has not considered this to be a responsibility of local governments.

Solid wastes originate in all operations of the nuclear energy industry and include such items as contaminated paper, laboratory glassware and equipment, as well as end products, such as chemical slurries and sludges, evaporation solids, and ion-exchange resins. For a number of years radioactive wastes have been temporarily stored in above-ground repositories. Extensive planning and research have been undertaken to develop methods for permanent disposal of such wastes. One method under consideration involves the placement of radioactive wastes in an underground cavity, such as a mine, vault, or drill hole, in a geologically stable rock formation. A pilot project has been planned in which high-level wastes are to be buried deep in a bedded salt formation. Other proposals have included burial in the seabed and burial in the Antarctic ice sheet. Numerous technological problems must be solved before these methods can be implemented.

Demolition and construction refuse consists of lumber, pipes, brick masonry, and other materials from buildings and other structures. Some of this material is salvaged for resale (*e.g.*, old bricks, lead pipe, and copper), but most is hauled away by the contractor for disposal at landfill sites. Explosives and inflammable materials, highly alkaline or acidic sludges, magnesium, manganese, and cyanides are not handled as part of the regular community collection system because of potential injury to workers and special disposal requirements. Other special wastes include large dead animals and quantities of condemned food. Each is subject to special handling procedures before burial or incineration.

BIBLIOGRAPHY. AMERICAN PUBLIC WORKS ASSOCIATION, COMMITTEE ON SOLID WASTE DISPOSAL, ***Municipal Refuse Disposal***, 3rd ed. (1970), a comprehensive text on municipal disposal methods, system selection, and management, and ***Solid Waste Collection Practice***, 4th ed. (1975), a complete text on municipal collection systems and their management; F. FLINTOFF and R. MILLARD, ***Public Cleansing*** (1968), a modern text treating refuse collection and disposal methods in the United Kingdom; C.G. GOLUEKE, ***Comprehensive Studies of Solid Wastes Management: Abstracts and Excerpts from the Literature***, 4 vols. (1968–72), an extensive collection of annotated references from 1939 to 1972; AMERICAN PUBLIC WORKS ASSOCIATION, STREET SANITATION COMMITTEE, ***Street Cleaning Practice***, 3rd ed. (1978), a comprehensive text on street cleaning, snow and ice control methods, and equipment in use in the United States; R.D. LIPSCHUTZ, ***Radioactive Waste: Politics, Technology and Risk*** (1980), a work that provides both technical and nontechnical information on the nature, production, and management of radioactive wastes.

(L.We.)

Regeneration, Biological

Organisms differ markedly in their ability to replace lost or amputated body parts. Some grow a new structure on the stump of the old one. By such regeneration organisms may dramatically replace substantial portions of themselves when they have been cut in two, or may grow organs or appendages that have been lost. Not all living things regenerate parts in this manner, however. The stump of an amputated structure may simply heal over without replacement. This wound healing is itself a kind of regeneration at the tissue level of organization: a cut surface heals over, a bone fracture knits, and cells replace themselves as the need arises.

SIGNIFICANCE

Regeneration, as one aspect of the general process of growth, is a primary attribute of all living systems. Without it there could be no life, for the very maintenance of an organism depends upon the incessant turnover by which all tissues and organs constantly renew themselves. In some cases rather substantial quantities of tissues are replaced from time to time, as in the successive production of follicles in the ovary or the molting and replacement of hairs and feathers. More commonly, the turnover is expressed at the cellular level. In mammalian skin, the epidermal cells produced in the basal layer may take several weeks to reach the outer surface and be sloughed off. In the lining of the intestines, the life span of an individual epithelial cell may be only a few days.

The motile, hairlike cilia and flagella of single-celled organisms are capable of regenerating themselves within an hour or two after amputation. Even in nerve cells, which cannot divide, there is a continuous flow of cytoplasm from the cell body out into the nerve fibres themselves. New molecules are continuously being generated and degraded with turnover times measured in minutes or hours, in the case of some enzymes, or several weeks, as in the case of muscle proteins. (Evidently, the only molecule exempt from this turnover is deoxyribonucleic acid [DNA], which ultimately governs all life processes.)

There is a close correlation between regeneration and generation. The methods by which organisms reproduce themselves have much in common with regenerative processes. Vegetative reproduction, which occurs commonly in plants and occasionally in lower animals, is a process by which whole new organisms may be produced from fractions of parent organisms; *e.g.*, when a new plant develops from a cut portion of another plant, or when certain worms reproduce by splitting in two, each half then growing what it has lost. More commonly, reproduction is achieved sexually by the union of an egg and sperm. This is a case in which an entire organism develops from a single cell, the fertilized egg, or zygote. This remarkable event, which occurs in all organisms that reproduce sexually, testifies to the universality of regenerative processes. During the course of evolution the regenerative potential has not changed; only the levels of organization at which it is expressed have altered. If regeneration is an adaptive trait, it would be expected to occur more commonly among organisms that have the greatest need of such a capability, either because the hazard of injury is great or the benefit to be gained is great. The actual distribution of regeneration among living things, however, seems at first glance to be a rather fortuitous one. It is difficult indeed to understand why some flatworms are able to regenerate heads and tails from any level of amputation, while other species can regenerate in only one direction or are unable to regenerate at all. Leeches fail to regenerate, while their close relatives, the earthworms, replace lost parts with ease. Certain species of insects regularly grow back missing legs, but many others are totally lacking in this capacity. Virtually all modern bony fishes can regenerate amputated fins, but the cartilaginous fishes (including sharks and rays) are unable to do so. Among the amphibians, salamanders regularly regenerate their legs, which are not very useful for movement in their aquatic environment, while frogs and toads, which are so much more dependent on their legs, are unable to replace them. If natural selection operates on the principle of efficiency, it is difficult to explain these many inconsistencies.

Whether regeneration is adaptive

Some species are so clearly adaptive that they have evolved not only mechanisms for regeneration, but mechanisms for self-amputation, as if to exploit the regenerative capability. The process of losing a body part spontaneously is called autotomy. The division of a protozoan into two cells and the splitting of a worm into two halves may be regarded as cases of autotomy. Some colonial marine animals called hydroids shed their upper portions periodically. Many insects and crustaceans will spontaneously drop a leg or claw if it is pinched or injured. Lizards are famous for their ability to release their tails. Even the shedding of antlers by deer may be classified as an example of autotomy. In all these cases autotomy occurs at a predetermined point of breakage. It appears that wherever nature contrives to lose a part voluntarily, it provides the capacity for replacement.

Sometimes, when part of a given tissue or organ is removed, no attempt is made to regenerate the lost structures. Instead, that which remains behind grows larger. Like regeneration, this phenomenon—known as compensatory hypertrophy—can take place only if some portion of the original structure is left to react to the loss. If three-quarters of the human liver is removed, for example, the remaining fraction enlarges to a mass equivalent to the original organ. The missing lobes of the liver are not themselves replaced, but the residual ones grow as large as necessary in order to restore the original function of the organ. Other mammalian organs exhibit similar reactions. The kidney, pancreas, thyroid, adrenal glands, gonads, and lungs compensate in varying degrees for reductions in mass by enlarging the remaining parts.

It is not invariably necessary for the regenerating tissue to be derived from a remnant of the original tissue. Through a process called metaplasia, one tissue can be converted to another. In the case of lens regeneration in certain amphibians, in response to the loss of the original lens from the eye, a new lens develops from the tissues at the edge of the iris on the upper margin of the pupil. These cells of the iris, which normally contain pigment granules, lose their colour, proliferate rapidly, and collect into a spherical mass which differentiates into a new lens.

MODES OF REGENERATION

Basic patterns. Not all organisms regenerate in the same way. In plants and in coelenterates such as the hydra and jellyfishes, missing parts are replaced by reorganization of preexisting ones. The wound is healed, and the neighbouring tissues reorganize themselves into whatever parts may have been cut off. This process of reorganization, called morphallaxis, is the most efficient way for simple organisms to regenerate. Higher animals, with more complex bodies, regenerate parts differently, usually by the production of a specialized bud, or blastema, at the site of amputation. The blastema, made up of cells that look very much alike despite their often diverse origins, made its first appearance evolutionarily in flatworms and is encountered in the regenerative processes of all higher animals. It provides the tissue that will form the regenerated part.

Atypical regeneration. Sometimes the part that grows back is not the same as that which was lost, and, occasionally, regeneration may be induced without having lost anything at all. It is not uncommon for a regenerated part to be incomplete. Earthworms, for example, usually regenerate only five segments in the anterior direction even if more than that number have been amputated. Many insects regenerate abnormally small legs from which some segments may be missing. Tadpole tails when amputated grow back to about only half their original length. These and other cases testify to the fact that a little regeneration is often good enough—that it is not necessary in every case to reproduce a flawless copy of the original.

Variations from the originals in regenerated parts

Sometimes that which is regenerated is very different from the original. Among the arthropods there are cases in which the stump of an antenna grows a leg, while a cut eyestalk regenerates an antenna. More commonly, the regenerated part may be a reasonable facsimile of the original but will differ in details. A regenerated lizard tail contains an unsegmented cartilaginous tube instead of a series of vertebrae as did the original tail. The spinal cord lacks segmented ganglia, and the scales in the skin differ in character from the original ones. A regenerated tail, therefore, is easily distinguished from one that has never been amputated, yet it is apparently sufficient to serve the purpose. Another interesting case is that of jaw regeneration in salamanders. If the lower jaw is amputated a new one will grow back, but it is often smaller than the original. It contains teeth and a mandible, but lacks a new tongue. Furthermore, the mandible that is produced is a cartilaginous model of the original, and is not known to convert into bone.

Sometimes more of a part grows back than has been removed by amputation. A limb stump, for example, can occasionally give rise to hands with extra digits. Lobsters have been known to regenerate double structures, in which case the new parts are mirror images of each other.

THE REGENERATION PROCESS

Origin of regeneration material. Following amputation, an appendage capable of regeneration develops a blastema from tissues in the stump just behind the level of amputation (see Figure 1). These tissues undergo drastic changes. Their cells, once specialized as muscle, bone, or cartilage, lose the characteristics by which they are normally identified (dedifferentiation); they then begin to migrate toward, and accumulate beneath, the wound epidermis, forming a rounded bud (blastema) that bulges out from the stump. Cells nearest the tip of the bud continue to multiply, while those situated closest to the old tissues of the stump differentiate into muscle or cartilage, depending upon their location. Development continues until the final structures at the tip of the regenerated appendage are differentiated, and all the proliferating cells are used up in the process.

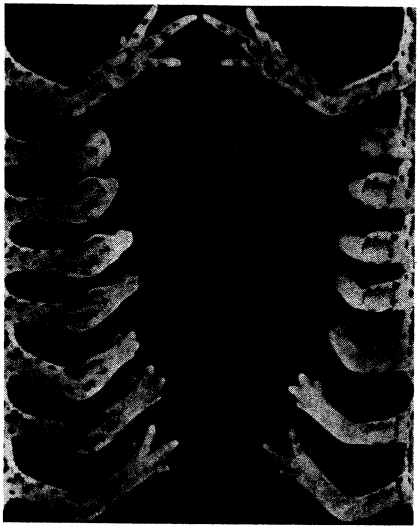


Figure 1: Successive stages in the regeneration of opposite limbs in a newt following amputation through (left) lower and (right) upper arms. At the top are the original limbs. From top to bottom are the successive stages of regeneration at 7, 21, 25, 28, 32, 42, and 70 days after amputation. The (right) more proximal stump elongates faster but differentiates more slowlv than does the (left) more distal stump.

R.J. Goss, Principles of *Regeneration* (1969), Academic Press. New York

The blastema cells seem to differentiate into the same kind of cells they were before, or into closely related types. Cells may perhaps change their roles under certain conditions, but apparently rarely do so. If a limb blastema is transplanted to the back of the same animal, it may continue its development into a limb. Similarly, a tail blastema transplanted elsewhere on the body will become a tail. Thus, the cells of a blastema seem to bear the indelible stamp of the appendage from which they were produced and into which they are destined to develop. If a tail blastema is transplanted to the stump of a limb, however, the structure that regenerates will be a composite of the two appendages.

Polarity and gradient theory. Each living thing exhibits polarity, one example of which is the differentiation of an organism into a head, or forward part, and a tail, or hind part. Regenerating parts are no exception; they exhibit polarity by always growing in a distal direction (away from the main part of the body). Among the lower invertebrates, however, the distinction between proximal (near, or toward the body) and distal is not always clear cut. It is not difficult, for example, to reverse the polarity of "stems" in colonial hydroids. Normally a piece of the stem will grow a head end, or hydranth, at its free, or distal, end; if that is tied off, however, it regenerates a hydranth at the end that was originally proximal. The polarity in this system is apparently determined by an activity gradient in such a way that a hydranth regenerates wherever the metabolic rate is highest. Once a hydranth has begun to develop, it inhibits the production of others proximal to it by the diffusion of an inhibitory substance downward along the stem.

Differences between whole organisms and appendages

When planarian flatworms are cut in half, each piece grows back the end that is missing. Cells in essentially identical regions of the body where the cut was made form blastemas, which, in one case gives rise to a head and in the other becomes a tail. What each blastema regenerates depends entirely on whether it is on a front piece or a hind piece of flatworm: the real difference between the two pieces may be established by metabolic differentials. If a transverse piece of a flatworm is cut very thin—too narrow for an effective metabolic gradient to be set up—it may regenerate two heads, one at either end. If the metabolic activity at the anterior end of a flatworm is artificially reduced by exposure to certain drugs, then the former posterior end of the worm may develop a head.

Appendage regeneration poses a different problem from that of whole organisms. The fin of a fish and the limb of a salamander have proximal and distal ends. By various manipulations, it is possible to make them regenerate in a proximal direction, however. If a square hole is cut in the fin of a fish, regeneration takes place as expected from the inner margin, but may also occur from the distal edge. In the latter case, the regenerating fin is actually a distal structure except that it happens to be growing in a proximal direction.

Amphibian limbs react in a similar manner. It is possible to graft the hand of a newt to the nearby body wall, and once a sufficient blood flow has been established, to sever the arm between the shoulder and elbow. This creates two stumps, a short one consisting of part of the upper arm, and a longer one made up of the rest of the arm protruding in the wrong direction from the side of the animal. Both stumps regenerate the same thing, namely, everything normally lying distal to the level of amputation, regardless of which way the stump was facing. The reversed arm therefore regenerates a mirror image of itself.

Clearly, when a structure regenerates it can only produce parts that normally lie distal to the level of amputation. The participating cells contain information needed to develop everything "downstream," but can never become more proximal structures. Regeneration, like embryonic development, occurs in a definite sequence.

Regulation of regeneration. There are certain prerequisites without which regeneration cannot occur. First and foremost, there must be a wound, although the original appendage need not have been lost in the process. Second, there must be a source of blastema cells derived from remnants of the original structure or an associated one. Finally, regeneration must be stimulated by some external force. The stimuli often involve the nervous system. An adequate nerve supply is required for the regeneration of fish fins, taste barbels, and amphibian limbs. In the case of many tail regenerations, the spinal cord provides the necessary stimulus. Lens regeneration in salamander eyes depends upon the presence of a retina. Arthropod appendages regenerate in the presence of molting hormones. Protozoan regeneration requires the presence of a nucleus. In case after case, regeneration depends on more than a healed wound and a source of blastema cells. It is often triggered by some physiological stimulus originating elsewhere in the body, a stimulus invariably associated with the very function of the structure to be regenerated. The conclusion is inescapable that regeneration is primarily the recovery of deficient functions rather than simply the replacement of lost structures.

The imperative of need is of further importance in suppressing excess regeneration. To be able to regenerate is to run the risk of regenerating too much or too often. If regeneration did not depend upon a physiological stimulus, such as those mediated by nerves or hormones, there would be no reason why simple wounds should not sprout whole new appendages.

Failure of regenerative processes

It is not known why regeneration fails to occur in many cases, as in the legs of frogs or the limbs and tails of mammals. The nerve supply might be inadequate, for when the number of nerves is artificially increased, regeneration is sometimes induced. This cannot be the whole answer, however, because not all appendages depend on nerves for their regeneration; newt jaws, salamander gills, and deer antlers do not require nerves to regenerate.

Possibly the failure to regenerate relates to the ways in which wounds heal. In higher vertebrates there is a tendency to form thick scar tissue in healing wounds, which may act as a barrier between the epidermis and the underlying tissues of the stump. In the absence of direct contact between these two tissues, the stump may not be able to give rise to the blastema cells required for regeneration.

THE RANGE OF REGENERATIVE CAPABILITY

Virtually no group of organisms lacks the ability to regenerate something. This process, however, is developed to a remarkable degree in lower organisms, such as protists and plants, and even in many invertebrate animals such as earthworms and starfishes. Regeneration is much more restricted in higher organisms such as mammals, in which it is probably incompatible with the evolution of other body features of greater survival value to these complex animals.

Protists and plants. ***Algae.*** One of the most outstanding feats of regeneration occurs in the single-celled green alga ***Acetabularia.*** This plant-like protist of shallow tropical water consists of a group of short rootlike appendages; a long thin "stem," up to several centimetres in length; and an umbrella-like cap at the top. The entire organism is one cell, with its single nucleus situated at the base in one of the "roots." If the cap is cut off, a new one regenerates from the healed over stump of the amputated stem. The nucleus is necessary for this kind of regeneration, presumably because it provides the information needed to direct the development of the new cap. Once this information has been produced by the nucleus, however, the nucleus can be removed and regeneration continues unabated.

If the nucleus from one species of ***Acetabularia*** is added to a cell-body of another species, and the cap of the recipient cell is amputated, the new cap that regenerates will be a hybrid because each nucleus exerts its own morphogenetic influences. On the other hand, if the nucleus from one species is substituted for that in another, regeneration reflects the properties of the new nucleus.

Protozoans. Most single-celled, animal-like protists regenerate very well. If part of the cell fluid, or cytoplasm, is removed from ***Amoeba,*** it is readily replaced. A similar process occurs in other protozoans, such as flagellates and ciliates. In each case, however, regeneration occurs only from that fragment of the cell containing the nucleus. Amputated parts that lack a nucleus cannot survive. In some ciliates, such as ***Blepharisma*** or ***Stentor***, the nucleus may be elongated or shaped like a string of beads. If either of these organisms is cut in two so that each fragment retains part of the elongated nucleus, each half proceeds to grow back what it lacks, giving rise to a complete organism in less than six hours. The way in which such a bisected protozoan regenerates is almost identical with the way it reproduces by ordinary division. Even a very tiny fragment of the whole organism can regenerate itself, provided it contains some nuclear material to determine what is supposed to be regenerated.

Green plants. The mechanisms by which vascular plants grow has much in common with regeneration. Their roots and shoots elongate by virtue of the cells in their meristems, the conical growth buds at the tip of each branch. These meristems are capable of indefinite growth, especially in perennial plants. If they are amputated they are not replaced, but other meristems along the stem, normally held in abeyance, begin to sprout into new branches that more than compensate for the loss of the original one. Such a process is called restitution.

Plants are also capable of producing callus tissue wherever they may be injured. This callus is proliferated from cambial cells, which lie beneath the surface of branches and are responsible for their increase in width. When a callus forms, some of its cells may organize into growing points, some of which in turn give rise to roots while others produce stems and leaves.

Invertebrates. ***Coelenterates.*** The vast majority of research on coelenterates has been focussed on hydras and some of the colonial hydroids. If a hydra is cut in half, the head end reconstitutes a new foot, while the basal portion regenerates a new hydranth with mouth and tentacles. This seemingly straightforward process is deceptively simple. From tiny fragments of the organism whole animals can be reconstituted. Even if a hydra is minced and the pieces scrambled, the fragments grow together and reorganize themselves into a complete whole. The indestructibility of the hydra may well be attributed to the fact that even the intact animal is constantly regenerating itself. Just below the mouth is a growth zone from which cells migrate into the tentacles and to the foot where they eventually die. Hence, the hydra is in a ceaseless state of turnover, with the loss of cells at the foot and at the tips of the tentacles being balanced by the production of new ones in the growth zone. If such an animal is X-rayed, the proliferation of new cells is inhibited and the hydra gradually shrinks and eventually dies owing to the inexorable demise of cells and the inability to replace them.

In colonial hydroids, such as ***Tubularia,*** there is a series of branching stems, each of which bears a hydranth on its end. If these hydranths are amputated they grow back within a few days. In fact, the organism normally sheds its hydranths from time to time and regenerates new ones naturally.

Planarian regeneration

Flatworms. Planarian flatworms are well-known for their ability to regenerate heads and tails from cut ends (see Figure ***2).*** In the case of head regeneration, some blastema cells become brain tissues, others develop into the eyes, and still others differentiate as muscle or intestine. In a week or so, the new head functions almost as well as the original.

After (top) T.H. Morgan in E. Korschelt, *Regeneration und Transplantation*, from (bottom: V. Hamburger, *Manual of Experimental Embryology* (© 1960), University of Chicago Press

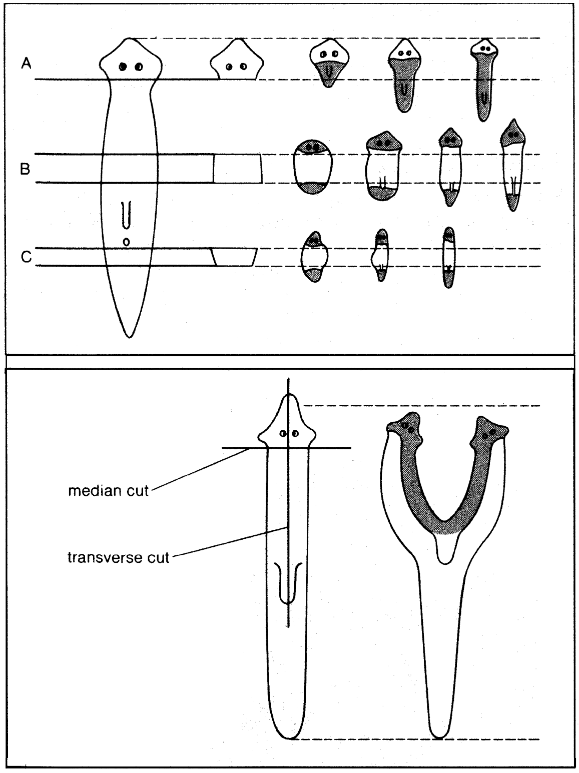


Figure 2: (Top) Regeneration in the planarian flatworm Dugesia. The three rows show regeneration from (A) anterior section, (B) midsection, and (C) posterior section of the animal at the left. (Bottom) Regeneration of a double head in a planarian. Regenerated tissue is shaded.

The blastema that normally gives rise to a single head is, under certain circumstances, even capable of becoming two heads if the stump of a decapitated flatworm is divided in two by a longitudinal cut. Each of the two halves then gives rise to a complete head. Thus, each blastema develops into an entire structure regardless of its size or position in relation to the rest of the animal.

In the case of flatworms there is still considerable disagreement concerning the origins of the blastema. Some investigators contend that it is derived from neoblasts, undifferentiated reserve cells scattered throughout the body. Others claim that there are no such reserve cells and that the blastema develops from formerly specialized cells near the wound that dedifferentiate to give rise to the blastema cells. Whatever their source, the cells of the blastema are capable of becoming many different things depending upon their location.

Regeneration in flatworms occurs in a stepwise fashion. The first tissue to differentiate is the brain, which induces the development of eyes. Once the head has formed, it in turn stimulates the production of the pharynx. The latter then induces the development of reproductive organs farther back. Thus, each part is necessary for the successful development of those to come after it; conversely, each part inhibits the production of more of itself. If decapitated flatworms are exposed to extracts of heads, the regeneration of their own heads is prevented. Such a complex interplay of stimulators and inhibitors is responsible for the successful regeneration of an integrated morphological structure.

Annelids. The segmented worms exhibit variable degrees of regeneration. The leeches, as already noted, are wholly lacking in the ability to replace lost segments, whereas the earthworms and various marine annelids (polychaetes) can often regenerate forward and backward. The expression of such regenerative capacities depends very much on the level of amputation. Anteriorly directed regeneration usually occurs best from cuts made through the front end of the worm, with little or no growth taking place from progressively more posterior bisections. Posteriorly directed regeneration is generally more common and extensive. Some species of worms replace the same number of segments as were lost. Hypomeric regeneration, in which fewer segments are produced than were removed, is more common, however.

Anterior regeneration depends upon the presence of the central nerve cord. If this is cut or deflected from the wound surface, little or no forward regeneration may take place. Posterior regeneration requires the presence of the intestine, removal of which precludes the formation of hind segments. Thus, it would seem that no head will regenerate without a central nervous system, nor a tail without an opening.

Arthropods. Many insects and crustaceans regenerate legs, claws, or antennas with apparent ease. When insect legs regenerate, the new growth is not visible externally because it develops within the next proximal segment in the stump. Not until the following molt is it released from its confinement to unfold as a fully developed leg only slightly smaller than the original. In the case of crabs, regenerating legs bulge outward from the amputation stump. They are curled up within a cuticular sheath, not to be extended until the sheath is molted. Lobsters and crayfish regenerate claws and legs in a straightforward manner as direct outgrowths from the stumps. As in other crustaceans, however, these regenerates lie immobile within an enveloping cuticle and do not become functional until their sheath is shed at the next molt.

Molting

In all arthropods regeneration is associated with molting, and therefore takes place only during larval or young stages. Most insects do not initiate leg regeneration unless there remains ample time prior to the next scheduled molt for the new leg to complete its development. If amputation is performed too late in the intermolt period, the onset of regeneration is delayed until after shedding; the regenerate then does not appear until the second molt. Metamorphosis into the adult stage marks the end of molting in insects, and adults accordingly do not regenerate amputated appendages.

Crustaceans often tend to molt and grow throughout life. They therefore never lose the ability to grow back missing appendages. When a leg is lost, a new outgrowth appears even if the animal is not destined to molt for many months. Following a period of basal growth, during which a diminutive limb is produced, the regenerated part eventually ceases to elongate. Not until a few weeks before the next molt does it resume growth and complete its development, triggered by the hormones that induce molting.

Vertebrates. Fishes. Many different parts of the fish's body will grow back. Plucked scales are promptly replaced by new ones, and amputated gill filaments can regenerate easily. The "whiskers," or taste barbels, of the catfish grow back as perfect replicas of the originals. The most conspicuous regenerating structures in fishes, however, are the fins. When any of these are amputated, new fins grow out from the stumps and soon restore everything that was missing. Even the coloured stripes or spots that adorn some fins are reconstituted by new pigment cells that repopulate the regenerated part. Fin regeneration depends on an adequate nerve supply. If the nerves are cut leading into the fin, regeneration of neither the amputated fin nor excised pieces of the bony fin rays can take place.

Amphibians. Salamanders are remarkable for their ability to regenerate limbs. Larval frogs, or tadpoles, also possess this ability, but usually lose it when they become frogs. It is not known why frog legs do not regenerate, and under appropriate stimuli they can be induced to do so.

Tadpoles and salamanders can replace amputated tails. Tadpole tails have a stiff rod called the notochord for support, whereas salamanders possess a backbone, composed of vertebrae. Both tails contain a spinal cord.When the salamander regenerates its tail, the spinal cord grows back and segmental nerve-cell clusters (ganglia) differentiate. Tadpoles also regenerate their spinal cords, but not the associated ganglia. If the spinal cord is removed or destroyed in the salamander, no tail regeneration occurs; if it is removed from the tadpole tail, however, regeneration can proceed without it.

Reptiles. Lizards also regenerate their tails, especially in those species that have evolved a mechanism for breaking-off the original tail when it is grasped by an enemy. When the lizard tail regenerates, however, it does not replace the segmented vertebrae. Instead, there develops a long tapering cartilaginous tube within which the spinal cord is located and outside of which are segmented muscles. The spinal cord of the lizard tail is necessary for regeneration, but the regenerated tail does not reproduce the ganglia that are normally associated with it. Occasionally, a side tail may be produced if the original tail is broken but not lost.

Birds. Regeneration of amputated appendages in birds is not known to occur; however, they do replace their feathers as a matter of course. While most species shed and regenerate feathers one at a time so as not to be grounded, flightless birds, such as penguins, may molt them all at once. Male puffins cast off their colorful beaks after the mating season, but grow new ones the following year. In like manner, the dorsal keel on the upper beaks of male pelicans is shed and replaced annually.

R.J. Goss, *Clinical Orthopaedics* and Related Research, no. 69, p. 228 (1970), J.B. Lippincott Company

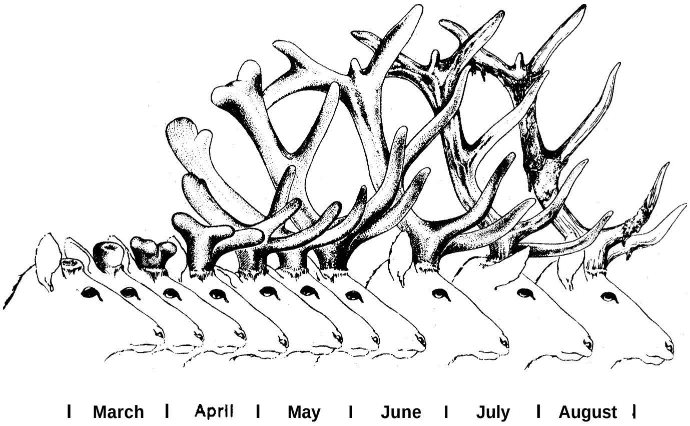


Figure 3: Regeneration of antlers in the elk (see text).

Mammals. Although mammals are incapable of regenerating limbs and tails, there are a few exceptional cases in which lost tissues are in fact regenerated. Not the least of these cases is the annual replacement of antlers in deer. These remarkable structures, which normally grow on the heads of male deer, consist of an inner core of bone enveloped by a layer of skin and nourished by a copious blood supply. During the growing season the antlers elongate by the proliferation of tissues at their growing tips. The rate of growth in some of the larger species may surpass one centimetre per day; the maximum rate of growth recorded for the elk is 2.75 cm per day. When the antlers have reached their full extent, the blood supply is constricted, and the skin, or velvet, peels off, thus revealing the hard, dead, bony antlers produced by the male deer in time for the autumn mating season. The regeneration of elk antlers spans about seven months (see Figure 3). The following spring, the old antlers are shed and new ones grow to replace them.

Still another example of mammalian regeneration occurs in the case of the rabbit's ear. When a hole is punched through the external ear of the rabbit, tissue grows in from around the edges until the original opening is reduced or obliterated altogether. This regeneration is achieved by the production of new skin and cartilage from the margins of the original hole. A similar phenomenon occurs in the case of the bat's wing membrane.

BIBLIOGRAPHY. R.J. GOSS, ***Principles of Regeneratzon*** (1969), a comprehensive analysis of comparative aspects of regeneration; E.D. HAY, ***Regeneration*** (1966), a short and incisive account of regeneration in invertebrates and vertebrates; V. KIORTSIS and H.A.L. TRAMPUSCH (eds.), ***Regeneration in Animals and Related Problems*** (1965), numerous reports by outstanding authorities on all kinds of regeneration; R.M H. MCMINN, ***Tissue Repair*** (1969), a thorough and up-to-date review of how various tissues of mammalian bodies heal injuries; A.E. NEEDHAM, ***Regeneration and Wound-Healing*** (1952), a detailed review, with numerous references; A.J. SCHMIDT, ***Cellular Biology of Vertebrate Regeneration and Repair*** (1968), on the cellular and molecular aspects of amphibian limb regeneration.

(R.J.G.)

Reinhardt, Max

A man of few words and little inclination or ability to develop or expound a dramatic theory, Max Reinhardt was a pragmatist whose instinctual feelings for the rightness of things transformed theatrical production in the 20th century. Before him, the idea of the director as a creative artist in his own right had been barely embryonic. With his work, the director emerged as the dynamic formative mind behind the production of a dramatic work.

By courtesy of the Theatre Collection, The New York Public Library at Lincoln Center, Astor, Lenox and Tilden Foundations

Reinhardt.

Discovery of the theatre. Reinhardt was born on Sept. 9, 1873, in Baden, near Vienna, the oldest of the seven children born to Wilhelm and Rose Goldmann, an orthodox Jewish couple. With his equally introverted only brother, Edmund, young Max played long hours with puppets and from their balcony watched the real puppets in the streets.

Though his parents were remote from theatrical life, they were sympathetic to his fascination with the actors of the Vienna Burgtheater, and at the urging of one of these, they allowed their son to exchange his boredom as a bank clerk for the excitement of drama school. Although he proved to be an inhibited actor, needing a beard and heavy makeup to release his talents, Reinhardt won local fame and friends in Salzburg. In 1894 he succumbed to an invitation from Otto Brahm, who had brought the drama of Henrik Ibsen to Germany, to join

With
Brahmin
Berlin

his Deutsches Theater in Berlin. He had assumed the stage name Reinhardt some time prior to moving to Berlin.

Reinhardt learned much from Brahm but was never wholeheartedly committed to the naturalism of his productions. He tired of "sticking a beard . . . and eating noodles and sauerkraut on stage every night," which latter activity was required by Brahm's notion of realism, in which nothing was to be simulated. This was not to be his direction in theatre. Quick to make friends despite his shyness, he met other young artists in cafes, gossiping, and loving. From their gatherings there emerged a lighthearted revue, ***Schall und Rauch (Sound and Smoke),*** to which Reinhardt contributed sketches. Playing before invited audiences, it was so successful that it was transformed into a serious work and settled into the Kleines Theater, a converted hall, in 1902. He planned a full season and directed his first play, Oscar Wilde's ***Salome.***

Career in full flower. Reinhardt evidenced his ability to make the right contact at the right time when he produced 14,000 marks to placate Brahm, who was furious over his breach of contract. He took over the Neues Theater in 1903, and his career moved ahead rapidly. By the end of 1904, he had directed 42 plays, but his early landmark of genius was the production in 1905 of Shakespeare's ***Midsummer Night's Dream.*** Reinhardt's staging was swift, light, and joyous, capturing for audiences the theatrical brilliance that had been buried for so long beneath productions devoted to a ponderous, reverent delivery of Shakespeare's words.

Direction of the Deutsches Theater

The young director became famous over night. Offered the artistic directorship of the Deutsches Theater, he would settle for nothing less than ownership. He purchased it for 1,000,000 marks, and at age 32 he had reached the pinnacle of his profession. He completely rebuilt the theatre, introducing the latest technological innovations in scenic design, and started a school. Purchasing a tavern next door, he remodelled it into a small theatre for plays that needed intimacy with the audience, summarizing the new concept in his word ***Kammerspiele,*** "chamber plays."

In his success, Reinhardt remained close to his family. He brought his brother, Edmund, who suffered from depressions, to Berlin and acted almost as his psychiatrist, setting him to work in the theatre to regain his confidence. Beginning in 1907, the Deutsches Theater toured throughout Europe and the United States. The production of ***The Miracle,*** which premiered in 1911 in London and played subsequently in New York and European cities, was Reinhardt's most spectacular work and, at the same time, probably the most characteristic. Reinhardt was fascinated by the sensuous quality of Roman Catholic rites and Gregorian chants. ***The Miracle,*** a work involving over 2,000 actors, musicians, dancers, and other personnel and without dramatic dialogue, was a modem-day reunification of drama and ritual. It was pure theatre in the most archetypal sense.

If in ***The Miracle*** he re-created an ancient unity, Reinhardt was equally important in giving new life to many of the great dramas from the theatre's past. His staging of Sophocles' ***Oedipus Rex*** in 1910 initiated the first large-scale revival of classical Greek drama in over 2,000 years. During the 1913–14 season he mounted new productions of 10 of 22 Shakespearean plays he had directed, using few, or no, settings and creating a major Shakespearean revival. In 1911 he brought a modem point of view to opera with his direction of the premiere of Richard Strauss's ***Rosenkavalier,*** with a libretto by Hugo von Hofmannsthal. After many years he succeeded in establishing the Salzburg Festival in 1920, staging Hofmannsthal's ***Jedermann (Everyman)*** in the city's cathedral square. It became an annual event, bringing about, before its termination in 1934, a new interest in the dramas of the Middle Ages from which ***Jedermann*** was adapted.

Return home and exile. Reinhardt had continued his work throughout World War I with no lessened sense of duty toward his art and his audience. In 1920, save for occasional engagements, he gave up direction of the Deutsches Theater. Retiring to a castle he had purchased in Austria, he attempted to find in his native country the regard he had been accorded abroad. His home was a meeting place for international celebrities, but enemies prevented him from feeling at home in his home town. He commuted in a circuit of Berlin, Vienna, and Salzburg. When the Nazis assumed power in Germany in 1933, Reinhardt was luckily abroad. In a letter to the Nazi government that was a typical blend of conceit, irony, rejection of politics, and prophetic perception, he left his theatrical empire to the German people. The era of private management of such institutions as the theatre is past, he wrote, and he foresaw that in the future it would be impossible to manage cultural undertakings without state backing.

Legacy to the German people

After further work in Europe, Reinhardt moved to the United States in 1938. He opened a workshop in Hollywood, where he had made a film of A ***Midsummer Night's Dream*** in 1934–35. His staging of ***Everyman*** in modem dress was followed by an unrealized plan for an all-Negro production of it. The final years of his life were filled with lesser fortunes and poor health, and he died speechless in New York City on Oct. 31, 1943.

Like the plots of the tragedies he so loved, Reinhardt's life was a rise to the heights of success and a fall to a life of uprooted exile. With his first wife, Else Heims, a beautiful and sensual actress, he had two sons. His second wife, Helene Thimig, was also a beautiful actress but, like Reinhardt, a shy person moved by an immense inner force and alive with conflicting appearances. He was an introvert capable of extreme extroversion and Falstaffian laughter. He disliked sentimentality in others yet was himself filled with romantic sentiments, a combination of Viennese sensitivity and German discipline with cosmopolitanism. His work summed up all theatre before him and opened new vistas for the theatre that followed.

BIBLIOGRAPHY. Books in English: HUNTLY CARTER, ***The Theatre*** of ***Max Reinhardt*** (1914), an attempt to define the nature of Reinhardt's work; O.M. SAYLER (ed.), ***Max Reinhardt And His Theatre*** (Eng. trans. 1924), a monumental collection of personal accounts, profusely illustrated. The main source is in German: GUSTI ADLER, ***Max Reinhardt*** (1964), a biography by Reinhardt's secretary for 20 years; see also ***Max Reinhardt sein Theater in Bildern,*** introduced by SIEGFRIED MELCHINGER (1968), a pictorial documentation of his work.

(Ho.I.P.)

Relativity

Relativity is concerned with measurements made by different observers moving relative to one another. In classical physics it was assumed that all observers anywhere in the universe, whether moving or not, obtained identical measurements of space and time intervals. According to relativity theory, this is not so, but their results depend on their relative motions.

There are really two distinct theories of relativity known in physics, one called the special theory of relativity, the other the general theory of relativity. Albert Einstein (*q.v.*) proposed the first in 1905, the second in 1916. Whereas the special theory of relativity is concerned primarily with electric and magnetic phenomena and with their propagation in space and time, the general theory of relativity was developed primarily in order to deal with gravitation. Both theories centre on new approaches to space and time, approaches that differ profoundly from those useful in everyday life; but relativistic notions of space and time are inextricably woven into any contemporary interpretation of physical phenomena ranging from the atom to the universe as a whole.

This article will set forth the principal ideas comprising both special and general relativity. It will also deal with some implications and applications of these theories.

THE SPECIAL THEORY OF RELATIVITY

Historical background. Classical physics owes its definite formulation to the British scientist Sir Isaac Newton (1642–1727). According to Newton, when one physical body influences another body, this influence results in a change of that body's state of motion, its velocity; that is to say, the force exerted by one particle on another results in the latter changing the direction of its motion, the magnitude of its speed, or both. Conversely, in the absence of such external influences, a particle will continue to move in one unchanging direction and at a constant rate of speed. This statement, Newton's first law of motion, is known as the law of inertia.

As motion of a particle can be described only in relation to some agreed frame of reference, Newton's law of inertia may also be stated as the assertion that there exist frames of reference (so-called inertial frames of reference) with respect to which particles not subject to external forces mave at constant speed in an unvarying direction. Ordinarily, all laws of classical mechanics are understood to hold with respect to such inertial frames of reference. Each frame of reference may be thought of as realized by a grid of surveyor's rods permitting the spatial fixation of any event, along with a clock describing the time of its occurrence.

The idea of absolute time

According to Newton, any two inertial frames of reference are related to each other in that the two respective grids of rods move relative to each other only linearly and uniformly (constant direction and speed) and without rotation, whereas the respective clocks differ from each other at most by a constant amount (as do the clocks adjusted to two different time zones on Earth) but go at the same rate. Except for the arbitrary choice of such a constant time difference, the time appropriate to various inertial frames of reference then is the same: If a certain physical process takes, say, one hour as determined in one inertial frame of reference, it will take precisely one hour with respect to any other inertial frame; and if two events are observed to take place simultaneously by an observer attached to one inertial frame, they will appear simultaneous to all other inertial observers. This universality of time and time determinations is usually referred to as the absolute character of time. The idea that a universal time can be used indiscriminately by all, irrespective of their varying states of motion—that is, by a person at rest at his home, by the driver of an automobile, and by the passenger aboard an airplane—is so deeply ingrained in most people that they do not even conceive of alternatives. It was only at the turn of the 20th century that the absolute character of time was called into question, and this as the result of a number of ingenious experiments described below.

As long as the building blocks of the physical universe were thought to be particles and systems of particles that interacted with each other across empty space in accordance with the principles enunciated by Newton, there was no reason to doubt the validity of the space–time notions just sketched. This view of nature was first called into question in the 19th century by the discoveries of a Danish physicist, Hans Christian Ørsted, the English scientist Michael Faraday, and the theoretical work of the Scottish-born physicist James Clerk Maxwell, all concerned with electric and magnetic phenomena. Electrically charged bodies and magnets do not affect each other directly over large distances, but they do affect one another by way of the so-called electromagnetic field, a state of tension spreading throughout space at a rapid but finite rate, which amounts to a speed of propagation of approximately 186,000 miles (300,000 kilometres) per second. As this value is the same as the known speed of light in empty space, Maxwell hypothesized that light itself was a species of electromagnetic disturbance; his guess has been confirmed experimentally, first by the production of lightlike waves by entirely electric and magnetic means in the laboratory by a German physicist, Heinrich Hertz, in the last quarter of the 19th century.

Maxwell's ether hypothesis

Both Maxwell and Hertz were puzzled and profoundly disturbed by the question of what might be the carrier of the electric and magnetic fields in regions free of any known matter. Up to their time, the only fields and waves known to spread at a finite rate had been elastic waves, which appear to the senses as sound and which occur at low frequencies as the shocks of earthquakes. Maxwell called the mysterious carrier of electromagnetic waves the ether, thereby reviving notions going back to antiquity. He attempted to endow his ether with properties that would account for the known properties of electromagnetic waves but was never entirely successful. The ether hypothesis, however, led two U.S. scientists, Albert Abraham Michelson and Edward Williams Morley, to conceive of an experiment (1887) intended to measure the motion of the ether on the surface of the Earth in their laboratory. On the reasonable hypothesis that the Earth is not the pivot of the whole universe, they argued that the motion of the Earth relative to the ether should result in slight variations in the observed speed of light (relative to the Earth and to the instruments of a laboratory) travelling in different directions. The measurement of the speed of light requires but one clock, if, by use of a mirror, a pencil of light is made to travel back and forth so that its speed is measured by clocking the total time elapsed in a round trip at one site; such an arrangement obviates the need for synchronizing two clocks at the ends of a one-way trip. Finally, if one is concerned with variations in the speed of light, rather than with an absolute determination of that speed itself, then it suffices to compare with each other round-trip-travel times along two tracks at right angles to each other, and that is essentially what Michelson and Morley did. To avoid the use of a clock altogether, they compared travel times in terms of the numbers of wavelengths travelled, by making the beams travelling on the two distinct tracks interfere optically with each other. (If the waves meet at a point when both are in the same phase—*e.g.*, both at their peak—the result is visible as the sum of the two in amplitude; if the peak of one coincides with the trough of the other, they cancel each other and no light is visible. Since the wavelengths are known, the relative positions of the peaks give an exact measure of how far one wave has advanced over the other.) This high-precision experiment, repeated many times with ever-improved instrumental techniques, has consistently led to the result that the speed of light relative to the laboratory is the same in all directions, regardless of the time of the day, the time of the year, and the elevation of the laboratory above sea level.

The special theory of relativity resulted from the acceptance of this experimental finding. If an Earth-bound observer could not detect the motion of the Earth through the ether, then, it was felt, probably any observer, regardless of his state of motion, would find the speed of light the same in all directions.

Relativity of space and time. An Irish and a Dutch physicist, George Francis FitzGerald and Hendrik Antoon Lorentz, independently showed that the negative outcome of Michelson's and Morley's experiment could be reconciled with the notion that the Earth is travelling through the ether, if one hypothesizes that any body travelling through the ether is foreshortened in the direction of travel (though its dimensions at right angles to the motion remain undisturbed) by a ratio that increases with increasing speed. If v denotes the speed of the body relative to the ether, and c is the speed of light, that ratio equals the square root of the quantity 1 minus the fraction speed squared over speed of light squared: $(1 - v^2/c^2)^{1/2}$. At ordinary speeds, c is so much greater that the fraction, practically speaking, is zero, and the ratio becomes $\sqrt{1}$, which is 1; *i.e.*, the foreshortening is nil; as v approaches c, however, the fraction becomes significant. The travelling body would be flattened completely if its velocity through the ether should ever reach the speed of light.

Suppose, now, that the variations in the speed of light were to be determined not by interference but by means of an exceedingly accurate clock and assume further that in such a modified experiment (whose actual performance is precluded at present, because even the best atomic clocks available do not possess the requisite accuracy) the motion through the ether were still imperceptible, then, Lorentz showed, one would have to conclude that all clocks moving through the ether are slowed down compared to clocks at rest in the ether, again by the factor $(1 - v^2/c^2)^{1/2}$. Thus, all rods and all clocks would be modified systematically, regardless of materials and construction design, whenever they were moving relative to

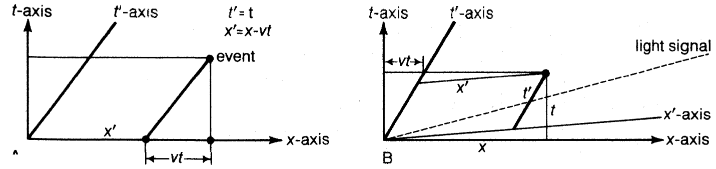
"Apparent" and "true" time and space

the ether. Accordingly, for theoretical analysis, one would have to distinguish between "apparent" and "true" space and time measurements, with the further proviso that "true" dimensions and "true" times could never be determined by any experimental procedure.

Conceptually, this was an unsatisfactory situation, which was resolved by Albert Einstein in *1905.* Einstein realized that the key concept, on which all comparisons between differently moving observers and frames of reference depended, is the notion of universal, or absolute, simultaneity; that is to say, the proposition that two events that appear simultaneous to any one observer will also be judged to take place at the same time by all other observers. This appears to be a straight-forward proposition, provided that knowledge of distant events can be obtained practically instantaneously. Actually, however, there is no known method of signalling faster than by means of light or radio waves or any other electromagnetic radiation, all of which travel at the same rate, c.

Suppose, now, that someone on Earth observes two events, say two supernovae (suddenly erupting very bright stars) appearing in different parts of the sky. Nothing can be said about whether these two supemovae emerged simultaneously or not from merely noting their appearance in the sky; it is necessary to know also their respective distances from the observer, which typically may amount to several hundred or several thousand light-years (one light-year, the distance light moves in one year, equals approximately 5.88×10^{12} miles, or 9.46×10^{12} kilometres). By the time one sees the eruption of a supernova, it has in actuality faded back into invisibility hundreds of years ago. Applying this simple idea to the observations and measurements made by different observers of the same events, Einstein demonstrated that if each observer applied the same method of analysis to his own data, then events that appeared simultaneous to one would appear to have taken place at different times to observers in different states of motion. Thus, it is necessary to speak of relativity of simultaneity.

Once this theoretical deduction is accepted, the findings of FitzGerald and Lorentz lend themselves to a new interpretation. Whenever two observers are associated with two distinct inertial frames of inference in relative motion to each other, their determinations of time intervals and of distances between events will disagree systematically, without one being "right" and the other "wrong." Nor can it be established that one of them is at rest relative to the ether, the other in motion. In fact, if they compare their respective clocks, each will find that his own clock will be faster than the others; if they compare their respective measuring rods (in the direction of mutual motion), each will find the other's rod foreshortened. The speed of light will be found to equal the same value, $c = 186{,}000$ miles per second, relative to every inertial frame of reference and in all directions. The status of Maxwell's ether is thereby cast in doubt, as its state of motion cannot be ascertained by any conceivable experiment. Consequently, the whole notion of an ether as the carrier of electromagnetic phenomena has been eliminated in contemporary physics.



From P. Bergmann, The Riddle *of Gravitation;* Charles Scribner's Sons

Figure 1: The Lorentz transformations.
(A) Nonrelativistic transformation; (B) relativistic transformation (see text).

The Lorentz transformations

The mathematical equations that relate space and time measurements of one observer to those of another observer moving are known as Lorentz transformations. If the relative motion is along the x-axis and if its magnitude is v, these expressions are (see Figure *1):*

$$x' = (1 - v^2/c^2)^{-1/2}\,(x - vt),\ y' = y,\ z' = z,$$
$$t' = (1 - v^2/c^2)^{-1/2}\,(t - vx/c^2).$$

Consequences. ***The limiting character of the speed of light.*** As the relative speed of one inertial frame of reference to another is increased, its rods appear increasingly foreshortened and its clocks more and more slowed down. As this relative speed approaches c, both of these effects increase indefinitely. The relative speed of the two frames cannot exceed c if light and other electromagnetic phenomena are to travel at the speed c in all directions when viewed from either frame of reference. Hence the special theory of relativity forecloses relative speeds of frames of reference greater than c. As an inertial frame of reference can be associated with any material object in uniform nonrotational motion, it follows that no material object can travel at a rate of speed exceeding c.

This conclusion is self-consistent only because under the Lorentz transformations the velocity of a body with respect to one inertial frame of reference is related to its velocity with respect to another frame not by the Newtonian rule that the difference in velocities equals the relative velocity between the two frames but by a more involved formula, which takes into account the changes in scale length, in clock time, and in simultaneity. If all velocities involved are in the same direction, then the velocity (see Figure **2)** in one frame, u, is related to the velocity in the other frame, u', by the expression stating that u' equals the sum of u and v divided by *1* plus the product of u and v divided by the square of c:

$$u' = (u + v)/(1 + uv/c^2).$$

As long as neither u nor v exceeds the speed of light, c, u' also will be less than c.

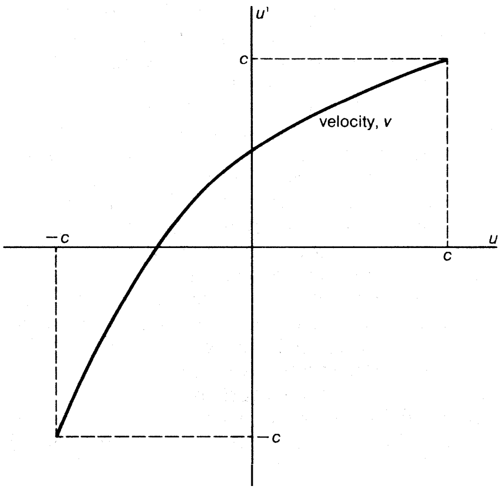


Figure 2: Velocities of the same body in two frames of reference (see text).

Variable mass. The mass of a material body is a measure of its resistance to a change in its state of motion caused by a given force. The larger the mass, the smaller the acceleration. If a material body is already moving at a speed approaching the speed of light, it must offer increasing resistance to any further acceleration so as not to cross the threshold of c. Hence the special theory of relativity leads to the conclusion that the mass of a moving body m is related to the mass that it would have if at rest, m_0, by a formula in which m equals m_0 divided by the square root of one minus the fraction v^2/c^2:

$$m = (1 - v^2/c^2)^{-1/2} m_0.$$

Relativistic mass

This changing value of the mass of the moving body, m, is called the relativistic mass. As v approaches c, the figure within the parentheses approaches zero and, ultimately, $m = m_0/0$, which would be an infinitely large number.

The relativistic mass formula may be interpreted as indicating that the relativistic mass of a body exceeds its rest mass m_0 by an amount that equals its kinetic energy E, divided by c^2: $m - m_0 = E/c^2$. Hence the hypothesis that generally the energy is c^2 times the mass, or $E = mc^2$, and that energy and mass are, in fact, equivalent physical concepts, differing only by the choice of their units. This hypothesis has been verified experimentally, in that all massive particles have been converted into forms of energy (for instance, gamma radiation) and conversely have been created out of pure energy. It was in part the recognition of this relationship that led to research out of which grew the technology of nuclear fission and fusion.

Invariant intervals. Data on pure time intervals obtained with respect to two relatively moving inertial frames of reference will differ and so will data on spatial distances. It is possible, however, to form from time intervals plus distances a single expression that will have the same value with respect to all inertial frames of reference. If the time interval between two distant events be denoted by T and their distance from each other by L, an expression, symbolized by τ, can be derived in which τ squared equals the square of the time interval minus the fraction of distance squared over speed of light squared: $\tau^2 = T^2 - L^2/c^2$. This will have the same value as $\bar{T}^2 - \bar{L}^2/c^2$, with $\bar{T}$ and $\bar{L}$ having been obtained in another inertial frame of reference. If τ^2 is positive, then τ is called the invariant (timelike) interval between the two events. If τ^2 is negative, then the expression λ, derived from the above as $\lambda^2 = L^2 - c^2T^2$, will be called the invariant (spacelike) interval.

The invariant interval between two instants in the history of one physical system equals the ordinary time lapse T measured by means of a clock at rest relative to that physical system, because, in such a comoving frame of reference, L vanishes. That is why such an invariant (timelike) interval is also referred to as the "proper time" elapsed between the two instants. Any clock will read its own proper time.

The "twin paradox." Given an inertial frame of reference and two similar material systems ("twins")—for instance, two atomic clocks of identical design—suppose now that one of these clocks remains permanently at rest in the given frame, whereas the other clock is moved at a high speed first in one direction away from the first clock and subsequently in the opposite direction until the two clocks are again close to each other. According to the Lorentz transformation, the second clock has been slower than the first throughout its journey, and hence it shows a smaller lapse of time than the clock that has remained at rest. By reading the clocks, one can then tell which clock has remained at rest, which one has moved. This difference in behaviour of the two clocks has been called the clock paradox, or the twin paradox if the clocks are replaced by space-travelling human twins.

The clock "paradox"

The "paradox" supposedly consists of a violation of the principle of relativity, according to which no asymmetric distinctions exist between different inertial frames of reference. The fallacy of this argument lies in the fact that no inertial frame of reference is associated with the second clock, as it cannot have moved free of acceleration throughout its journey: at least once its velocity (*i.e.*, the direction of its motion) must have been changed drastically, so as to enable it ever to return to its mate. Hence no violation of the principle of relativity; no paradox is involved. Various experiments on moving particles and atoms have indeed confirmed the predictions of the theory.

Four-diinensional space-time. The German mathematical physicist Hermann Minkowski pointed out that the invariant interval between two events has some of the properties of the distance in Euclidean geometry. Based on Euclidean geometry, the Cartesian coordinate system is designed to identify any point (event) in space by its reference to three mutually perpendicular lines or axes meeting at zero. The distance between two events, in accordance with Pythagoras' theorem, in any Cartesian (rectilinear) coordinate system is obtained by taking the square root of the sum of the squares of coordinate distances, $s^2 = x^2 + y^2 + z^2$, and its value is independent of the choice of coordinate system, though the values of x, y, and z are not. The invariant interval, similarly, is the square root of a sum and difference of squares of intervals of bcth space and time. Accordingly, Minkowski suggested that space and time should be thought of as comprising a single four-dimensional continuum, space–time, often also referred to as the Minkowski universe. Events, localized both as regards space and time, are the natural analogues of points in ordinary three-dimensional geometry; in the history of one particle, its proper time resembles the arc length of a curve in three-space.

In Minkowski's space–time the invariant interval may be either timelike or spacelike. If $L^2 - c^2T^2$ for two events happens to be zero, the invariant interval is neither, but null, or lightlike, as a light signal emanating from the earlier of the two events may just pass the second as the latter occurs. By contrast, in ordinary geometry the distance between two points, s, vanishes only if the two points coincide. To this extent the analogy between space–time and ordinary space is imperfect.

Minkowski's four-dimensional, geometric approach to relativity appears to add to the original physical concepts of relativity mostly a new terminology but not much else. Nevertheless, for the further conceptual development of relativity Minkowski's contribution has been of inestimable value.

THE GENERAL THEORY OF RELATIVITY

Physical origins. The general theory of relativity derives its origin from the need to extend the new space and time concepts of the special theory of relativity from the domain of electric and magnetic phenomena to all of physics and, particularly, to the theory of gravitation. As space and time relations underlie all physical phenomena, it is conceptually intolerable to have to use mutually contradictory notions of space and time in dealing with different kinds of interactions, particularly in view of the fact that the same particles may interact with each other in several different ways–electromagnetically, gravitationally, and by way of so-called nuclear forces.

Newton's theory of gravity

Newton's explanation of gravitational interactions must be considered one of the most successful physical theories of all time. It had accounted for the precise motions of all the constituents of the solar system with uncanny accuracy, permitting, for instance, the prediction of eclipses hundreds of years ahead. But Newton's theory visualizes the gravitational pull that the Sun exerts on the planets and the pull that the planets in turn exert on their moons and on each other as taking place instantaneously over the vast distances of interplanetary space, whereas according to relativistic notions of space and time any and all interactions cannot spread faster than the speed of light. The difference may be unimportant, for practical reasons, as all of the members of the solar system move at relative speeds far less than 1/1,000 of the speed of light; nevertheless, relativistic space–time and Newton's instantaneous action at a distance are fundamentally incompatible. Hence Einstein set out to develop a theory of gravitation that would be consistent with relativity.

Proceeding on the basis of the experience gained from Maxwell's theory of the electric field, Einstein postulated the existence of a gravitational field that propagates at the speed of light, c, and that will mediate an attraction as closely as possible equal to the attraction obtained from Newton's theory. From the outset it was clear that mathematically a field theory of gravitation would be more involved than that of electricity and magnetism. Whereas the sources of the electric field, the electric charges of particles, have values independent of the state of motion of the instruments by which these charges are measured, the source of the gravitational field, the mass of a particle, varies with the speed of the particle relative to the frame of reference in which it is determined and hence will have different values in different frames of reference. This complicating factor introduces into the task of constructing a relativistic theory of the gravitational field a measure of ambiguity, which Einstein resolved eventually by invoking the principle of equivalence.

The principle of equivalence. Everyday experience indicates that in a given field of gravity, such as the field caused by the Earth, the greater the mass of a body the greater the force acting on it. That is to say, the more massive a body the more effectively will it tend to fall toward the Earth; in fact, in order to determine the mass of a body one weighs it—that is to say, one really measures the force by which it is attracted to the Earth, whereas the mass is properly defined as the body's resistance to acceleration. Newton noted that the ratio of the attractive force to a body's mass in a given field is the same for all bodies, irrespective of their chemical constitution and other characteristics, and that they all undergo the same acceleration in free fall; this common rate of acceleration on the surface of the Earth amounts to an increase in speed by approximately 32 feet (about 9.8 metres) per second every second.

Weightlessness

This common rate of gravitationally caused acceleration is illustrated dramatically in space travel during periods of coasting. The vehicle, the astronauts, and all other objects within the space capsule undergo the same acceleration, hence no acceleration relative to each other. The result is apparent weightlessness: no force holds the astronaut to the floor of his cabin or a liquid in an open container. To this extent, the behaviour of objects within the freely coasting space capsule is indistinguishable from the condition that would be encountered if the space capsule were outside all gravitational fields in interstellar space and moved in accordance with the law of inertia. Conversely, if a space capsule were to be accelerated upward by its rocket engines in the absence of gravitation, all objects inside would behave exactly as if the capsule were at rest but in a gravitational field. The principle of equivalence states formally the equivalence, in terms of local experiments, of gravitational forces and reactions to an accelerated noninertial frame of reference (*e.g.*, the capsule while the rockets are being fired) and the equivalence between inertial frames of reference and local freely falling frames of reference. Of course, the principle of equivalence refers strictly to local effects: looking out of his window and performing navigational observations, the astronaut can tell how he is moving relative to the planets and moons of the solar system.

Einstein argued, however, that in the presence of gravitational fields there is no unambiguous way to separate gravitational pull from the effects occasioned by the noninertial character of one's chosen frame of reference; hence one cannot identify an inertial frame of reference with complete precision. Thus the principle of equivalence renders the gravitational field fundamentally different from all other force fields encountered in nature. The new theory of gravitation, the general theory of relativity, adopts this characteristic of the gravitational field as its foundation.

Curved space–time. *The* principles. In terms of Minkowski's space–time, inertial frames of reference are the analogues of rectilinear (straight-line) Cartesian coordinate systems in Euclidean geometry. In a plane these coordinate systems always exist, but they do not exist on the surface of a sphere: any attempt to cover a spherical surface with a grid of squares breaks down when the grid is extended over a significant fraction of the spherical surface. Thus a plane is a flat surface, whereas the surface of a sphere is curved. This distinction, based entirely on internal properties of the surface itself, classifies the surface of a cylinder as flat, as it can be rolled off on a plane and thus is capable of being covered by a grid of squares.

Einstein conjectured that the presence of a gravitational field causes space–time to be curved (whereas in the absence of gravitation it is flat), and that this is the reason that inertial frames cannot be constructed. The curved trajectory of a particle in space and time resulting from the effects of gravitation would then represent not a straight line (which exists only in flat spaces and space–times) but the straightest curve possible in a curved space–time, a geodesic. Geodesics on the surface of a sphere (such as the Earth) are the great circles. (The plane of any great circle goes through the centre of the Earth; all lines of longitude are great circles, but the Equator is the only line of latitude that is a great circle.) They are the least curved lines one can construct on the surface of a sphere, and they are the shortest curves connecting any two points. The geodesics of space–time connect two events (or two instants in the history of one particle) with the greatest lapse of proper time, as was indicated in the earlier discussion of the iwin paradox.

Riemannian space

If the presence of a gravitational field amounts to a curvature of space–time, then the description of the gravitational field in turn hinges on a mathematical elucidation of the curvature of four-dimensional space–time. Before Einstein, the German mathematician Bemhard Riemann (1826–66) had developed methods related directly to the failure of any attempt to construct square grids. Considering any small piece of (two-dimensional) surface and constructing within it a quadrilateral whose sides are geodesics, if the surface were flat, the sum of the angles at the four corners would be 360". If the surface is not flat, the sum of the angles is not 360". The deviation of the actual sum of the angles from 360" will be proportional to the area of the quadrilateral; the amount of deviation per unit of surface will be a measure of the curvature of that surface. If the surface is imbedded in a higher dimensional continuum, then one can consider similarly unavoidable angles between vectors constructed as parallel as possible to each other at the four corners of the quadrilateral, and thus associate several distinct components of curvature with one surface. And, of course, there are several independent possible orientations of two-dimensional surfaces, for instance, six in a four-dimensional continuum, such as space–time. Altogether there are 20 distinct and independent components of curvature defined at each point of space–time; in mathematics these are referred to as the 20 components of Riemann's curvature tensor.

The mathematical expression. Einstein discovered that he could relate ten of these components in a natural way to the sources of the gravitational field, mass (or energy), density, momentum density, and stress, if he were to duplicate approximately Newton's equations of the gravitational field and, at the same time, formulate laws that would take the same form regardless of the choice of frame of reference. The remaining ten components (usually referred to together as Weyl's tensor) may be chosen arbitrarily at any one point but are related to each other by partial differential equations (Bianchi's identities) at neighbouring points. Einstein's field equation,

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R = 8\pi\kappa \mathrm{T}_{\mu\nu}$$

($R_{\mu\nu}$ stands for those ten components of the curvature tensor that are called Ricci's tensor, $g_{\mu\nu}$ is the metric tensor, R is a combination of Ricci's tensor components known as the Ricci scalar, κ is a universal constant proportional to Newton's constant of gravitation, and $\mathrm{T}_{\mu\nu}$ denotes the components of the energy–stress tensor), along with the rule that a freely falling body moves along a geodesic, together form the comprehensive theory of gravitation known as the general theory of relativity.

Confirmation of the theory. The general theory of relativity is constructed so that its results are approximately the same as those of Newton's theories as long as the velocities of all bodies interacting with each other gravitationally are small compared to the speed of light; *i.e.*, as long as the gravitational fields involved are weak. The latter requirement may be stated roughly in terms of the escape velocity. The escape velocity is defined as the minimal speed with which a projectile must be endowed at any given location to enable it to fly off to infinitely removed regions of the universe without the application of further force. On the surface of the Earth the escape velocity is approximately 7.5 kilometres (4.7 miles) per second. A gravitational field is considered strong if the escape velocity approaches the speed of light, weak if it is much smaller. All gravitational fields encountered in the solar system are weak in this sense.

The success of Newton's theory, incidentally, must be considered a confirmation of the general theory of relativity to the extent that that application of the theory remains confined to situations involving small relative speeds and weak fields. Obviously, any superiority of the new theory over the old one may be inferred only if their predictions disagree and if those of the general theory of relativity are confirmed by experiment and observation.

As the principle of equivalence forms the cornerstone of general relativity, experiments testing this principle are of considerable inieresi. High-precision experiments with

Tests of the principle of equivalence

this objective were first performed by a Hungarian physicist, R. von Eotvos, who confirmed the principle to an accuracy of one part in 10^8 and, in the 1960s, by a U.S. physicist, Robert Dicke, who improved Eotvos' accuracy by another factor cf 1,000, achieving an accuracy of one part in 10^{11}. Through this work the principle of equivalence has become one of the most precisely confirmed genera! principles of contemporary physics.

Some other new predictions of general relativity are explained below.

Advance of Mercury's perihelion. All ellipses forming the trajectories of the various planets about the Sun turn slowly within their planes, because of the interactions of the planets with each other, but it was discovered in the 19th century that interplanetary perturbations could not account fully for the turning rate of Mercury's orbit, leaving unexplained about 43" of arc per century. The general theory of relativity predicts this discrepancy as a relativistic effect. R. Dicke has proposed that about one-tenth of Mercury's perihelion advance may be caused by the Sun's oblatencss (slight flattening at the poles), which, if confirmed, would spoil the otherwise perfect agreement between theory and observation. This matter was still unresolved in the early 1970s.

Gravitational red shift. General relativity predicts that spectral lines emanating from sources within a gravitational field will be shifted toward longer wavelengths (toward the red end of the spectrum) by an amount proportional to the gravitational potential at the site of the source. This effect was searched for and found first in astronomical objects, particularly in stars called white dwarfs, on whose surfaces the gravitational potential is relatively large. The best quantitative confirmation of gravitational red shift was obtained in laboratory experiments performed in Great Britain and the U.S. In the 1960s an accuracy of one part in 100 was achieved using the minute difference in gravitational potential between two sites differing in altitude by a few metres.

Optical effects of gravitation. General relativity predicts that the curvature of space–time results in the apparent bending of light rays passing through gravitational fields and in an apparent reduction of their speeds of propagation. The bending was first observed, within a couple of years of Einstein's publication of the new theory, during a total eclipse, when stellar images near the occulted disk of the Sun appeared displaced by fractions of 1″ of arc from their usual locations in the sky. The associated delay in travel time was observed in the late 1960s, when ultraintense radar pulses were reflected off Mercury and Venus just as these planets were passing behind the Sun. These experiments are difficult to perform and difficult to evaluate as to their ultimate accuracy, but it seems conservative to estimate the accuracy of confirmation of the relativistic effect no worse than a few parts in 100.

Gravitational waves. General relativity predicts the occurrence of gravitational waves, whose properties should resemble in some respects those of electromagnetic waves; they should travel at the same speed, c, and they should be capable of being polarized. It appears that a U.S. physicist, Joseph Weber, detected gravitational waves in the late 1960s by means of vibrations excited in large aluminum cylinders, each weighing several tons, that had been insulated with great care from all other contacts that might give rise to similar vibrations (from mechanical and electromagnetic forces). Though the pulses detected by Weber can be observed only marginally in the background of noise caused by self-generated thermal motions of the cylinders, the sources of these pulses must possess almost unimaginable intensities. Weber has coniectured that the sources are collapsing or exploding stars near the core of the Galaxy, which cannot be observed because of intervening dust and gas clouds.

Sources of gravitational waves

Because of the awesome implications of Weber's discovery, along with the admitted difficulty of evaluating the observations themselves, both Weber and other workers in the early 1970s were making great efforts to improve the sensitivity of the instrumentation, in hope of achieving more definitive results.

Future astrophysical tests. The discovery of quasars (quasi-stellar objects; see below ***Relativistic cosmology)*** in the early 1960s and of pulsars (sources of rapidly pulsing, intensive radiation in the range of radio frequencies) in the latter part of the same decade has made it probable that there are astrophysical situations involving "strong" gravitational fizlds, in which general relativity will play a more than marginal role in interpreting observational data. New space techniques, such as the establishment of a laser gauge between the Earth and the Moon, the launching of deep-space probes, and the construction of orbiting space observatories, wili probably give rise to new and sensitive tests of general relativity; in the early 1970s such experiments had not progressed beyond the planning stage, however.

Conceptual implications of general relativity. The general theory of relativity represents a further modification of classical concepts of space and time that goes far beyond those implicit in the special theory. The special theory does away with the absolute character of time and with the absolute distance between two objects that are at rest relative to each other. The geometric concepts appropriate to the special theory are the four-dimensional space–time continuum, in which events that are fixed in space and in time are represented by points, often referred to as world points (to distinguish them from the points of ordinary three-dimensional space), and the histories of particles moving through space in the course of time by curves (world curves); the representations of particles that are not accelerated by forces are straight lines.

Minkowskian space–time

Minkowski's space–time is a rigidly flat continuum, as is the three-dimensional space of Euclid's geometry. Distances between world points are measured by the invariant intervals, whose magnitudes do not depend on the particular coordinate system, or frame of reference, used. The Minkowski universe is homogeneous: that is to say, geometric figures constructed at any site may be transferred to another site without distortion. Finally, among all the possible frames of reference there is a special set, the inertial frames of reference, just as in ordinary space the rectilinear coordinate systems are distinguished by their simplicity among all conceivable coordinate systems. Space–time serves as the immutable backdrop of all physical processes, without being affected by them.

In general relativity, space–time also is a four-dimensional continuum, with invariant intervals being defined at least locally between events taking place close to each other. But only small regions of space–time resemble the continuum envisagec! by Minkowski, just as small bits of a spherical surface appear nearly plane. In the broad sense, according to general relativity, space–time is curved, and this curvature is equivalent ta the presence of a gravitational field. Far from being rigid and homogeneous, the general-relativistic space–time continuum has geometric properties that vary from point to point and that are affected by local physical processes. Space–time ceases to be a stage, or scaffolding, for the dynamics of nature; it becomes an integral part of the dynamic process. General relativity, it has been said, makes physics part of geometry. With at least equal justification it may be claimed that general relativity makes geometry part of physics; thai is to say, of a natural science. Not only are the properties of space and time subject to scientific investigation, to a study by means of experiments, but specific properties, such as the amount of curvature in a particular location at a specified time, are to be measured with the help of physical instruments.

Though the general theory of relativity is universally thought to be the most nearly perfect theory of gravitation now known and its new approach to space and time accepted, most theories of the atom and of elementary particles current in the early 1970s are based on the space–time concepts of the special, not the general theory of relativity. This is in part because currently available mathematical techniques are adapted to treating these problems against the flat background of the original Minkowski space–time. Perhaps more important is the fact that by the early 1970s the fusion of general relativity and of quantum theory (*i.e.*, the theory that energy does not exist as a continuous phenomenon but as "particles" called quanta, the interactions between particles of matter and quanta of energy being explained in terms of quantum mechanics and wave mechanics), which certainly cannot be circumvented in dealing with atomic and subatomic phenomena, had not been completed successfully.

A number of scientists, foremost among them Einstein himself, attempted, during the first half of the 20th century, to increase the flexibility of the geometric structure of space–time further, so as to incorporate into the geometric structure all known physical forces and perhaps the quantum phenomena as well. This program is known as that of unitary field theory. None of the attempts at unitary field theory has been sufficiently successful to find favour with the community of physicists, and these endeavours have received relatively little support until now.

Unitary field theory

Schwarzschild's solution of the field equations. Immediately on publication of Einstein's paper on general relativity, the German astronomer Karl Schwarzschild found a mathematical solution to the new field equations, which corresponds to the gravitational field of a compact massive body. such as a star or planet, and which is now referred to as Schwarzschild's field. If the mass that serves as the source of the field is fairly diffuse, so that the gravitational field on the surface of the astronomical body is fairly weak, Schwarzschild's field will exhibit physical properties very similar to those described by Newton. Gross deviations will be found if the mass is so highly concentrated that the field on the surface is strong. At the time of Schwarzschild's work, 1916, this appeared to be a purely theoretical speculation; but with the discovery of pulsars and their interpretetion as probable neutron stars composed of matter that has the same density as atomic nuclei (so-called nuclear matter), there exists the distinct possibility that strong fields may soon be available to astronomical observation.

The most conspicuous feature of the Schwarzschild field is that if the total mass is thought of as concentrated at the very centre, then at a finite distance from that centre, the Schwarzschild radius, the geometry of space–time changes drastically from that to which we are accustomed. Particles and even light rays cannot penetrate from inside the Schwarzschild radius to the outside and be detected. Conversely, to an outside observer any objects approaching the Schwarzschild radius appear to take an infinite time to penetrate toward the inside. There cannot be any effective communication between the inside and the outside.

The exterior and the interior of the Schwarzschild radius are not cut off from each other entirely, however. Suppose an observer were to attach himself to a particle that is falling freely straight toward the centre and that this observer is equipped with a clock that reads its own proper time. This observer would penetrate the Schwarzschild radius within a finite proper time (see Figure 3); moreover, he would find no abnormalities in his environment as he did so. The reason is that his clock would deviate from one permanently kept outside and at a constant distance from the centre, so grossly that the same event that seen from the outside takes forever occurs within a finite time to the free-falling observer.

From P. Bergmann, *The Riddle of Gravitation*; Charles Scribner's Sons

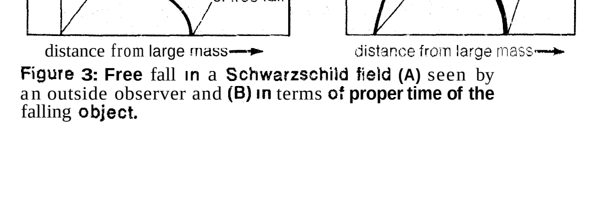


Figure 3: Free fall in a Schwarzschild field (A) seen by an outside observer and (B) in terms of proper time of the falling object.

These peculiarities of the Schwarzschild field may well have practical applications in astronomy. The U.S. physicist J. Kobert Oppenheimer had found that a star whose mass exceeds the mass of the Sun bv an appreciable factor is bound to contract and, eventually, to collapse under the Influence of its own gravitational pull, no matter how resistant its constituent matter. As a good many stars are believed to have such large masses, it is likely that there already exist some collapsed stars, so-called black holes. Though continuing to make its presence known by the gravitational attraction it exerts on other stars, a biack hole would absorb but not emit light, and thus be invisible, hence its name. In the early 1970s a number of astronomers had claimed the discovery of black holes, but these claims have not been confirmed.

APPLICATIONS OF RELATIVISTIC PRINCIPLES

Particle accelerators. Modem particle accelerators speed up particles to energies at which they very nearly achieve the speed of light. At tnese energies and speeds the differences in behaviour predicted by classical physics and by the special theory of relativity are huge; the machine must be designed in accordance with relativistic principles, or it will not operate.

In the early 1970s the most powerful electron synchrotrons were operating at energies of several thousand million electron volts, which means that the relativistic mass of one electron orbiting at maximum energy was roughly 10,900 times its rest mass. Accordingly, the magnetic field required to maintain the electrons in orbit was 10,000 times as powerful as it would have to be if nonrelativistic physics held, at the same speed. On the other hand, at that given energy the speed of the electrons is in fact very nearly equal to the speed of light, the difference amounting to no more than one part in 100,000,000 (10^8). *At* the same energy, but by nonrelativistic mechanics, the speed of the electrons would have been about 100 times the speed of light. This difference has a very practical consequence: in those panicle accelerators designed for highly relativistic energies, the synchrotrons, particles are injected into a circular orbit already near the speed of light, and their velocities change only slightly as their energies are brought up to the highest design value. If the orbit diameter is kept nearly constant. particles at all energies will circulate at the same frequency, and only the magnetic field that keeps them in orbit needs to be increased to keep pace with the increasing mass. The accelerating voltage is applied at the constant frequency required so that the particles will always be accelerated forward.

Operations of the synchrotron

Relativisticparticle physics. The physics of elementary particles (that is to say, subatomic particles) depends on the principles of the special theory of relativity. These principles have their greatest application when particles are created, annihilated, or converted into different particles. In most particle transformations, large amounts of energy are involved; the total (rest) masses of the particles involved in the transformations will change, and this change will be related to the amounts of energy expended or gained by the rule that the change in mass *(Am,)* is balanced by a corresponding change in energy (ΔE), divided through by the square of the speed of light (c^2): $\Delta m_0 = -c^{-2}\Delta E$. This rule has been confirmed universally and, by now, is being taken for granted.

The units, or quanta, of electromagnetic energy, increasing with frequency and called photons, have long been regarded as a species of particle in which are combined the properties of zero rest mass with nonvanishing relativistic mass, because they travel at the speed of light. The relativistic mass equals its total energy E, according to an equation in which $\hbar$ denotes Planck's universal constant h divided by 211 and ω the angular frequency of the radiation: $E = \hbar\omega$, divided again by c^2. The relativistic mass of a photon can be checked experimentally if the photon is absorbed or deflected in its interactions with particles, when the change in its linear momentum (product of velocity and relativistic mass) results in a recoil by the other particles. If a high-frequency photon, a gamma photon, collides with a free electron, the result is called the Compton effect, which involves both an observable recoil on the part of the electron and an altered frequency of the deflected photon. Again, relativity is confirmed by experiment.

It has been conjectured that gravitational waves, also, are composed of zero-rest-mass quanta travelling at the speed of light. These hypothetical particles have been dubbed gravitons. As the quantum theory of the gravitational field has not been definitely established and as the detection of individual gravitons may remain beyond experimental capabilities for years to come, the existence of gravitons could not be considered assured in the early 1970s.

Neutrinos

There is another species of zero-rest-mass particles, produced in radioactive decay involving the ejection of electrons or positrons from atomic nuclei (so-called beta decay). These particles, known as neutrinos, have no electric charge and travel at the speed of light. Four distinct species of neutrinos are now known, each of which is produced in a different kind of beta decay. Neutrinos interact with any type of particle extremely weakly. As a result, they are capable of traversing large amounts of matter with but a slight chance of being deflected or absorbed in the process. Though their existence has been confirmed beyond a doubt, their detection and detailed examination remained a challenging problem in the early 1970s.

Relativistic cosmology. Theories concerning the structure and history of the whole universe have assumed an increasingly empirical aspect in the course of the 20th century. In the 1960s, particularly, a combination of new observation techniques, new discoveries, and applications of the special and general theories of relativity has resulted in considerable progress.

The most important techniques added to those of observations by means of visible light were: radio astronomy; infrared, ultraviolet, X-ray, and gamma-ray astronomy from extraterrestrial platforms; cosmic-ray investigations; neutrino astronomy; and examination of the Moon and other astronomical bodies by unmanned and manned space exploration. Discoveries with cosmological impact included: the expanding universe, quasars, and the "primeval fireball." Each is explained briefly below.

Edwin Powell Hubble, a U.S. astronomer, had discovered that the more distant astronomical objects exhibited a shift of spectral lines toward the red (long wavelength) end of the spectrum, the extent of the shift increasing the greater their distance from Earth. This cosmological red shift has been generally interpreted as evidence of rapid recession of these distant objects in an expanding universe. The present rate of expansion is expressed as the amount of recession per unit distance and is known as the Hubble constant. It amounts to about a mile per second recessional velocity for a distance of 10^5 light-years. Equivalently, if the rate of expansion has been constant, it must have started about 2×10^{10} years ago.

Quasars, also called quasi-stellar objects (QSO's), appear to be structures that combine extreme luminosity (100 times that of a bright galaxy) with great compactness, taking up less space than the distance between the Sun and its nearest neighbour star. Wherever a spectral analysis of a quasar's emitted light has been possible, the spectral lines have been considerably red shifted, increasing their wavelengths from a few percent to several hundred percent. If these red shifts are interpreted as cosmological (and this interpretation has not been accepted by all astronomers), some quasars are more distant from the Galaxy than any other known objects. As such they may provide indications of the large-scale structure of the universe, which could not be obtained from investigations confined to "close" surroundings. The term close is to be understood in relation to distance in which Hubble's red shift becomes large ("cosmological distances"), distances amounting to thousands of millions of light years.

The notion of the primeval fireball

Finally, the term primeval fireball refers to the discovery of an all-pervasive background of radiation whose frequencies lie in the border region between microwave radio frequencies and infrared, corresponding to wavelengths of the order of millimetres and centimetres. In the early 1970s this radiation was interpreted as a remnant of the original intensive radiation that must have been associated with the early history of the universe, when matter was both extremely dense and extremely hot; hence its name. Its spectral composition, which at the time of writing was the object of intensive investigation, might provide some clues to the early history of the universe.

General relativity contributes to a theoretical discussion of cosmology the idea that the universe as a whole need not be flat even on the average and that it probably is not. Even if one were to assume that on a very large (cosmological) scale the universe is homogeneous and isotropic (*i.e.*, having the same properties in all directions), which appears a reasonable working hypothesis in the absence of any evidence to the contrary, there are a number of different possibilities. The universe might be spatially open (as a flat one surely is), or it might be closed, somewhat as the surface of a sphere is closed, without boundaries. Likewise, in the time direction the universe might be either open or closed; it is a little difficult to visualize a time-wise closed universe, which appears to be inconsistent with ordinary notions of cause and effect. But as these notions are distilled out of normal experience, which does not extend too far beyond one human lifetime, they might be inapplicable on the scale of billions of years. In brief, many different cosmological models have been proposed and investigated theoretically, but present observational information does not seem to favour one particular type. The information at hand appears to favour types that involve an expansion from an early stage involving fireball conditions.

MODIFICATIONS OF GENERAL RELATIVITY

The steady-state theory of the universe

Aside from the attempts at unitary field theories, which have been mentioned above, two modifications have been proposed in the past three decades. One is the steady-state cosmological theory developed in England by the astronomers Hermann Bondi, Thomas Gold, and Fred Hoyle. These authors have suggested that the exotic early stage of the universe and its progressive thinning out could be obviated if it were assumed that matter is created continuously at a very low rate, just sufficient to maintain a constant density of matter in the universe, in spite of the observed expansion. On the cosmological scale their model of the universe would persist forever without changing its over-all characteristics. The steady-state theory evoked much interest for some years, but in the early 1970s discovery of the universal background radiation was considered an indication that there was an early dense state of the universe very different from the universe known today. Nevertheless, the discussion cannot be considered definitely closed.

The other modification to be mentioned is the scalar-tensor theory proposed by a German physicist, Pascual Jordan, and, independently, by a U.S. physicist, R. Dicke. This theory proposes to make the universal constants of nature, such as the ratio of electric charge to mass of the electron, variable and dependent on the expansion of the universe. The proponents of this theory have adduced evidence gathered from the history of the Earth and the structure of the Sun, and they have designed experiments that would make it possible to make a clear-cut choice between the scalar-tensor theory and conventional general relativity. By the early 1970s the evidence in hand appeared to favour Einstein's theory, but not decisively.

BIBLIOGRAPHY

Expositions for general readers: ALBERT EINSTEIN, ***Relativity, the Special and General Theory: A Popular Exposition***, 3rd ed. (1921), a popularization for the lay reader of a classic work (published originally in German in 1917), written by one of the greatest scientists of all times; BERTRAND RUSSELL, ***The ABC of Relativity***, rev. ed. ***by*** F. PIRANI (1958); ALBERT EINSTEIN and LEOPOLD INFELD, ***The Evolution of Physics*** **(1938);** LEOPOLD INFELD, ***Albert Einstein: His Work and Its Influence on Our World***, rev. ed. (1950), two books that cover the whole of physics, with special em-

phasis on relativity (Infeld was one of Einstein's chief collaborators in the 1930s); P.G. BERGMANN, ***The Riddle of Gravitation*** (1968), a work that emphasizes the general theory of relativity and includes a discussion of current astronomical and cosmological research.

presentations for readers with technical training: H.A. LORENTZ ***et al., The Principle of Relativity*** (1952), *a* collection of fundamental research papers, all in English; ALBERT EINSTEIN, ***The Meaning of Relativity,*** 5th ed. (1955), based on lectures by Einstein delivered in 1921, with two appendixes containing Einstein's views on cosmology through 1945, and his work on the "nonsymmetric" unitary field theory to the time of his death in 1955; DAVID BOHM, ***The Special Theory of Relativity*** (1965), a thoroughgoing treatment of the special theory combined with a discussion of the philosophical foundations of physics; A.P. FRENCH, ***Special Relativity*** (1968), *an* introduction at the undergraduate level; H. BONDI, ***Cosmology,*** 2nd ed. (1961), a survey of cosmology at a technical level, including observational data through the late 1950s; P.G. BERGMANN, ***Introduction to the Theory of Relativity*** (1942); C. MOLLER, ***The Theory of Relativity*** (1952); and J.L. SYNGE, ***Relativity: The Special Theory*** (1956) and ***Relativity: The General Theory*** (1960), technical texts, on the graduate level, that represent three distinct approaches to the subject by active research workers.

(P.G.Be.)

Relaxation Phenomena

The term **relaxation** is used by chemists and physicists to describe the interval or time lag between the application of an external stress to a system—that is, to an aggregation of matter—and its response. The relaxation effect may be caused by a redistribution of energy among the nuclear, electronic, vibrational, and rotational energy states of the atoms and molecules that comprise the system, or it may result from a shift in the ratio of the number of product molecules to the number of reactant molecules (those initially taking part) in a chemical reaction. The measurement of relaxation times can provide many insights into atomic and molecular structures and into the rates and mechanisms of chemical reactions.

Historical survey. The word relaxation was originally applied to a molecular process by the English physicist James Clerk Maxwell. In a paper "On the Dynamical Theory of Gases," which he presented in 1866, Maxwell referred to the time required for the elastic force produced when fluids are distorted to diminish or decay to l/e (*e* is the base of the natural logarithm system) times its initial value as the "time of relaxation" of the elastic force. The earliest suggestion of a chemical relaxation effect is contained in a dissertation (Berlin, 1910) based on research directed by the German physical chemist Walther Nernst. Measurements of sound propagation through the gas nitrogen tetroxide, which breaks up, or dissociates, into nitrogen dioxide, led Nernst to suggest that experiments at frequencies at which the dissociation reaction could not keep pace with the temperature and pressure variations that occur within a sound wave, would permit evaluation of the dissociation rate. Ten years later, at a meeting of the Prussian Academy of Sciences, Albert Einstein presented a paper in which he described the various theoretical aspects of this relaxation effect.

Dielectric relaxation

The detection of the chemical relaxation effect predicted by Nernst and Einstein did not become technically feasible until the last half of the 20th century. In the first half of the century physicists and chemists in studying relaxation concentrated on physical relaxation processes. Peter Debye referred to the time required for dipolar molecules (ones whose charges are unevenly distributed) to orient themselves in an alternating electric field as dielectric relaxation. Sound absorption by gases was used to investigate energy transfer from translational, or displacement in space, to rotational (spinning and tumbling) and vibrational (oscillations within the molecule) degrees of freedom, the three independent forms of motion for a molecule. The former requires only a few molecular collisions, but the transfer of energy between translational and vibrational modes may require thousands of collisions. Because the processes are not instantaneous but time dependent, relaxation effects are observed. Their measurement provides information about molecular bonding and structure. Chemical relaxation was rediscovered by the German physical chemist Manfred Eigen in 1954. Since then, technological advances have permitted the development of techniques for the measurement of relaxation times covering the entire range of molecular processes and chemical reactivity.

The great variety of relaxation phenomena and of the techniques developed for their study precludes a comprehensive survey. To facilitate a general discussion, the relaxing system, its initial and final states, the nature of the disturbance, and the system's response are considered separately. Examples are cited that emphasize the important features of relaxation phenomena and illustrate the variety of information that can be obtained from their study. A moderately detailed description of one relaxation technique, the temperature-jump method, is used to summarize the discussion.

Dimers and monomers

Relaxation. *General* considerations. The chemical relaxation of nitrogen tetroxide is easy to visualize, and it illustrates principles common to all relaxation phenomena. Nitrogen tetroxide (formula N_2O_4) actually is a dimer (a molecule consisting of two similar molecules called monomers) that dissociates into two monomers, or molecules of nitrogen dioxide (formula NO.,). The monomer and dimer are easily distinguishable: the former is a brown gas; the latter is a colourless gas. The product and reactants exist in equilibrium, represented by the reversible reaction,

$$N_2O_4 \rightleftarrows 2NO_2.$$

1 molecule 2 molecules

At ambient (room) temperature and atmospheric pressure, approximately 80 percent of the molecules in the mixture are dimers and the remaining molecules are monomers. The distribution of molecules between the two forms remains unchanged as long as the temperature and pressure are held constant. If the external conditions are altered, then the ratio of monomers to dimers will adjust to a new value. The dependence of the equilibrium on pressure is intuitively understandable as follows: to a good approximation the volume that a gas occupies at a given pressure and temperature depends directly on the number of gas molecules. The dissociation of one molecule of nitrogen tetroxide into two moiecules of nitrogen dioxide entails an expansion of the gas, a doubling of molecules, which is opposed by the external pressure. If the external pressure is increased, the system acts to relieve the stress by reducing its volume; *i.e.*, by combining monomers to form dimers and thus reducing the number of molecules. The equilibrium shifts in favour of dimers under increased pressure, and in favour of monomers under reduced pressure. At any steady pressure, the ratio of the two forms eventually becomes constant.

Chemical relaxation results from the inability of systems at equilibria to respond instantaneously to changes in external conditions. The rate of re-establishment of equilibrium, or re-equilibration, is limited by the concentrations of the reactants and their reactivities. At any specified temperature and pressure, there is a definite probability per unit time that a nitrogen tetroxide molecule will dissociate into two nitrogen dioxide molecules and that the latter will recombine to form a dimer. The average lifetime of a nitrogen tetroxide molecule at ambient temperature and atmospheric pressure, for example, is about one-third of a microsecond (one-millionth of a second). The product of the reciprocal of the average lifetime times the concentration of nitrogen tetroxide molecules gives the rate at which they dissociate. At equilibrium there is no net change in the number of nitrogen tetroxide molecules, because their dissociation rate is exactly balanced by the rate at which they are being re-formed through association of nitrogen dioxide molecules. If the external conditions are altered, the reactivities of the monomer and dimer change instantaneously but their concentrations change at a finite rate until the balance between the association and dissociation rates is re-established.

Sound propagating through a gas can be pictured as a

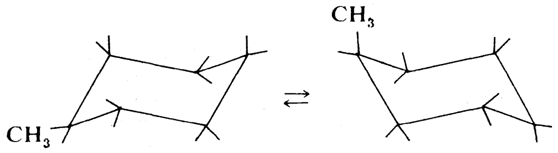
Sound waves in gases

pressure wave whose alternating increase and falling off of pressure, called a sinusoidal variation of pressure, with time at any point in the medium is accompanied by a corresponding fluctuation in the temperature. The effect of the varying temperature and pressure of a sound wave moving through nitrogen tetroxide gas on the dissociation of nitrogen tetroxide depends on the frequency of that sound wave. When the pressure oscillates slowly enough, the dissociation reaction will remain at equilibrium with the oscillation; that is, the extremes in the monomer–dimer ratio will coincide with the extremes of pressure and temperature. If, on the other hand, the pressure fluctuates too rapidly for the reaction to follow, the ratio of monomers to dimers will remain constant at the equilibrium value for the ambient temperature and pressure; but at intermediate frequencies, a relaxation effect may be observed, and a readjustment of the chemical equilibrium will lag behind the pressure variation within the gas.

The relaxing chemical equilibrium results both in the absorption of sound by the gas and in dispersion of, or changes in, the sound velocity. Measurement of either of these effects permits evaluation of the relaxation time. The maximum absorption of sound occurs, for example, when the angular frequency (two pi times cycles per second) of the sound wave equals the reciprocal of the relaxation time. The relaxation time can then in turn be related to the mechanism of the chemical reaction and to the reactivities of the reactants.

The relaxing system. Relaxation may occur between any two allowed energy states of nuclei, atoms, or molecules in the solid, liquid, or gas phase. A distinction has already been made between chemical relaxation, which involves a transformation between two chemically distinguishable molecules such as the dissociation of nitrogen tetroxide, and physical processes such as the transfer of energy between translational and vibrational states of a molecule displayed by sound absorption in a homogeneous gas. Although it is useful to classify relaxation processes as chemical or molecular, the distinction between them depends on the height of the energy barrier separating the chemical species, and it becomes blurred when structural isomerizations are considered. Liquid methylcyclohexane, for example, absorbs sound of ultrahigh frequency. The relaxation effect is attributed to an isomerization (change in structure) between two forms of the molecule called the axial and equatorial chair forms, as shown below:

Isomerizations



equatorial form of methylcyclohexane

axial form of rnethylcyclohexane

In the axial form, the methyl group lies perpendicular to the principal axis of the carbon ring, whereas in the equatorial form the methyl group lies in the plane of the ring. Whether or not the interconversion is considered a chemical or a molecular relaxation process is largely a matter of definition.

Atomic nuclei may exhibit relaxation effects. Some nuclei spin mechanically. Because nuclei are charged, there is a magnetic field associated with a spining nucleus: it behaves like a simple bar magnet with a north and south pole. The nucleus is said to have a magnetic moment that will experience a force when placed in an external magnetic field. A hydrogen nucleus in an external magnetic field, for example, may orient its nuclear magnetic moment either parallel or antiparallel to the external field. The latter is a higher energy orientation, called the upper spin state. The equilibrium distribution of many hydrogen nuclei between the two spin states (parallel and antiparallel) can be perturbed (*i.e.*, changed) by the absorption of electromagnetic radiation of appropriate frequency. The system will then relax to the equilibrium distribution by time-dependent radiationless transitions of the hydrogen nuclei from the upper to the lower spin state. This process of returning to the equilibrium distribution is called spin-lattice relaxation, because the excess energy of the upper spin state is transferred to molecules surrounding the relaxing hydrogen nuclei as increased translational, rotational, or vibrational energy.

As with nuclei, atoms and molecules can be excited to higher energy states by the absorption of electromagnetic radiation. A nonequilibrium distribution of atoms or molecules in excited states is frequently accomplished by a technique, called flash photolysis, in which the system of atoms or molecules is subjected to an intense flash of visible or ultraviolet light. The excited species may undergo many fates, but if they decay to the equilibrium distribution between the ground, or lowest, states and the excited states of the original atoms or molecules, the system is said to have relaxed.

The word relaxation is sometimes used to describe the radiation of energy by individual molecules, atoms, or nuclei, rather than by a large number. A hydrogen nucleus, for example, may decay from the upper to the lower spin state by transferring radiant energy to a nearby hydrogen nucleus in the lower spin state. This exchange of spins is called spin-spin relaxation. It shortens the lifetime of an individual excited nucleus, but it does not restore the equilibrium distribution of parallel and antiparallel spins. Although it is convenient to think of an individual excited nucleus as relaxing, only the response of an excited population of many nuclei can be measured. This usage of the term relaxation obscures the most useful experimental feature of relaxation processes.

The initial and final states. In virtually all relaxation experiments a thermodynamic equilibrium state is disturbed, and the time required for re-equilibration is measured. The practical advantage of starting with a system at equilibrium is most apparent in the study of chemical reactions in solution. Nearly all of the elementary steps in chemical reactions, such as transfers of protons and electrons from one molecule to another, occur in less than a millisecond, and yet as late as the 1960s solution reactions with half-times (time in which the reaction is half completed) shorter than a millisecond could not be studied. This limit was imposed by the hydrodynamic problem of mixing two solutions. Reaction rates had been studied by mixing the reactants and monitoring the rate at which products appeared. The most elaborate mechanical mixing devices that have been built so far require a millisecond to initiate a solution reaction. Manfred Eigen was the first person to clearly perceive that mixihg could be avoided by perturbing an equilibrium and watching it relax. His enormous contribution to the study of fast chemical reactions was recognized by the award of a Nobel Prize in 1967.

Perturbation

Instead of disturbing an equilibrium system, a stationary state may be perturbed. Many enzyme-catalyzed reactions, for example, are experimentally irreversible. Nevertheless, for much of the time course of the reaction, the chemical intermediates are present in a stationary state; that is, their concentrations do not change. The stationary state can be disturbed and the rate of its re-establishment may be used to deduce the lifetimes of the chemical intermediates. Combined rapid mixing and relaxation techniques have been used successfully in a study of catalysis by the enzyme ribonuclease.

The nature of the disturbance. Eigen has divided the methods used to disturb systems into indirect, or competition, methods and direct, or perturbation, methods. In the first method, the relaxing system is continuously disturbed. The competition between the disturbance and the relaxation process results in the establishment of a stationary state, from which information about the relaxation process must be inferred. Ultrasonic absorption is an example of a competition method. The competition between the temperature and pressure variations in the sound wave and the dissociation of nitrogen tetroxide sets up a stationary state in which re-equilibration of the chemical reaction lags behind the pressure fluctuations in the sound wave. The reactivities of the monomer and dimer are derived indirectly from measurements of sound absorption. Flash photolysis is an example of the second method, in which the system is momentarily perturbed. The molecules are electronically excited from the ground, or lowest and normal, energy state to higher energy states by the flash. Their return, or decay, to the ground state can be followed directly by monitoring the reemission of the absorbed light.

A chemical equilibrium can be disturbed by changing the pressure or temperature or by applying an electric field. If a volume change accompanies a chemical reaction, the ratio of products to reactants at equilibrium will depend on the pressure. The point at which equilibrium sets in will depend on temperature, if heat is absorbed or released in the reaction. The ratio will also depend on electric field strength, if the polarizabilities (change in orientation or position of electric charges) of the reactants and products are different. Nuclear and electronic states can be excited by the absorption of electromagnetic radiation, and the latter can also be excited thermally. Perturbation forces, when expressed mathematically in terms of strength and time, are called forcing functions. In principle, a forcing function may assume any form, but in practice, it must be easy to generate experimentally and to analyze mathematically. Examples of forcing functions are the sinusoidal temperature and pressure variations in a sound wave (charting the variations produces a curve called a sine curve—which varies from positive to negative values) and sinusoidally alternating electric fields, which are used in dielectric relaxation measurements. Other convenient forcing functions are step, or incremental, perturbations and rectangular pulses (pulses of which the strength rises nearly instantaneously, remains at the higher value for a period of time, and then rapidly returns to its initial value).

Step perturbations of the temperature and pressure can be produced in shock tubes. A gas at high pressure is separated by a membrane from the gas being studied at low pressure. When the membrane is burst, a plane pressure wave caused by the high-pressure driving gas moves through the low-pressure gas under study. Temperature increases of several thousand degrees may accompany moderate pressure shocks. The shock front travels through the gas with a velocity comparable to the mean molecular velocity, so that the width of the shock front is only a few mean free paths (average distances travelled by the molecules between collisions). As the shock passes, the translational energy of the molecules in the shock front is increased. The system relaxes as the excess energy is distributed by collisions to rotational and vibrational degrees of freedom.

Rectangular temperature perturbations (plotted on a graph these show up as a curve that periodically rises suddenly, stays constant for an interval, then drops suddenly to original value) can be produced in aqueous solutions of reacting systems by using microwaves to heat the solution. Water molecules can absorb energy of rotation at 10^{10} hertz (cycles per second). By concentrating the microwave energy in a small volume, an increase of several degrees in temperature can be obtained in one microsecond using pulses of radar. Since the radar generator can be repeatedly pulsed, coupling it with a continuous flow system improves the experimental accuracy by averaging over the period of the experiment.

Detection

Response of the system. Any of the techniques for disturbing an equilibrium can be combined with a variety of detection systems. Depending on the nature of the relaxation effect, it can be monitored by absorption or emission spectroscopy, by fluorometry, or by polarimetry. Conductance changes can be measured. Crystals are used to detect ultrasonic waves and to measure absorption effects.

While a priori there is no restriction on the magnitude of the displacement from equilibrium, in practice small disturbances are used to permit the application of a linear rare equation (terms denoting changes with time are to the first power). The rate of disappearance, for instance, of a small displacement from equilibrium is approximately proportional to the magnitude of the displacement. This relationship is given by the differential equation

$$-\frac{d}{dt}(\mathrm{AX}) = \frac{\Delta X}{\tau}.$$

Here, the displacement (AX) is the difference between the instantaneous and the equilibrium values of the relaxing property, which might be the kinetic energy of molecules behind a shock front or the concentration of a chemical reactant. The reciprocal of the constant of proportionality has units of time and is called the relaxation time (τ, tau). Since the equilibrium values may be time dependent, the solution of the rate equation depends on the form of the forcing function. Propagation of a sound wave through nitrogen tetroxide gas, for instance, causes a sinusoidal variation of the equilibrium concentrations of monomers and dimers with time. A great advantage of relaxation methods is that the response to small disturbances can be approximated by a first order differential equation.

The relaxation time for a chemical process can be related to the reactivities of the reactants if the reaction mechanism is known. Conversely, it may be possible to deduce the reaction mechanism from the dependence of the relaxation time on reactant concentrations. If several chemical reactions are coupled, or if more than one vibrational state is excited, a spectrum of relaxation times may be observed. The relaxation times for the individual relaxation processes can be determined from the measured relaxation times, which are the normal modes for the coupled system.

Temperature-jump experiment. To summarize and to clarify this discussion, a temperature-jump relaxation experiment will be described. The name temperature jump is reserved for the relaxation technique in which a stepwise temperature perturbation is achieved by passing a large electric current through the solution under study and thus heating it almost instantaneously. Instrumentally, it is one of the simplest relaxation techniques. It is also the most generally useful method for the study of fast chemical reactions in solution.

A typical temperature-jump instrument produces a temperature rise of approximately **8° *C*** within five microseconds. The principles of this instrument are briefly explained as follows. A 0.05-microfarad capacitor is charged to between 30 and 40 kilovolts. The electrical energy stored on the capacitor is proportional to its capacitance and to the voltage squared. It is discharged through the reaction cell at time zero by closing a variable spark gap. The time required for dissipation of roughly 80 percent of the stored energy is given by the product of the capacitance times the cell resistance. The energy is dissipated through collisions between the ions, which conduct the discharge current through the solution and the solvent molecules. The rapid temperature increase causes a shift in the concentrations of reactive molecules in the solution to new equilibrium values. If this shift is accompanied by a colour change, the reaction rate can be monitored spectrophotometrically (*i.e.*, the change in the intensity of light of a selected wavelength with time is measured). The results are recorded on a storage oscilloscope for later display. Provided thar the rise time of the temperature pulse is much shorter, and that the thermal re-equilibration time is much longer, than the response time of the chemical reaction being studied, the temperature jump can be apprcximated as a step perturbation. At times greater than zero, the equilibrium concentrations of the reactants remain constant at the values corresponding to the higher temperature. Consequently, the differential equation for the disappearance of the displacement of reactant X from equilibrium can be integrated to show that this value decays exponentially.

In the introduction to an article on the ***Molecular Basis of*** Visual *Excitation* the Nobel laureate George Wald wrote,

> I have often had cause to feel that my hands are cleverer than my head. That is a crude way of characterizing the dialectics of experimentation. When it is going well, it is Like

a quiet conversation with Nature. One asks a question and gets an answer; then one asks the next question, and gets the next answer. An experiment is a device to make Nature speak intelligibly. After that one has only to listen.

Relaxation phenomena afford a unique method for making nature speak intelligibly about rapid energy transfers and chemical reactions. They have only begun to be exploited, especially to probe the elementary steps in complex biochemical reactions.

*BIBLIOGRAPHY. **Kolloid-Zeitschrift,*** vol. 134 (1953), a symposium on the relaxation properties of matter, includes Meixner's theoretical treatment of relaxation phenomena based on irreversible thermodynamics; ***Discussions of the Faraday Society,*** no. 17 (1954), a colloquium on the study of fast reactions including a discussion of chemical relaxation by Eigen; ***Zeitschrift für Elektrochemie,*** vol. 64 (1960), an international colloquium on fast reactions in solution; ***Molecular Relaxation Processes*** (1965), Chemical Society symposium emphasizing the use of relaxation to determine molecular structures; M. EIGEN and L. DEMAEYER, "Relaxation Methods," in S.L. FRIESS ***et al.*** (eds.), ***Technique of Organic Chemistry,*** vol. 8, pt. 2, pp. 793–798, 895–1054 (1963), an elegant and comprehensive discussion of chemical relaxation.

(L.D.F.)

Religion, Philosophy of

In addition to treating what is commonly called the philosophy of religion, this article considers a wide spectrum of situations, experiences, and issues recognized as "religious" and endeavours to appraise the characteristic approaches and attitudes not only of the adherents of particular religions but also of those who stand outside any particular religion, whether as sympathizers or caustic critics. Outside the scope of this article, however, are questions relating to the study of religions and its methodology or questions relating to the types of argument by which one interpretation of a religious claim is preferred to another (see also RELIGION, STUDY OF).

This article is divided into the following sections:

I. Nature and characteristics: preliminary approaches and common or typical views

RELIGION AS A FACT IN HUMAN EXPERIENCE, CULTURE, AND HISTORY

Evidences of religious attitudes and loyalties exist in every sector of human life—in human experience in general; in "culture," the complex interweaving of attitudes, concerns, and views; and in history, the record of social and personal behaviour.

Characteristic feelings and emotions of religion

The findings of psychology. Religion incorporates certain characteristic feelings and emotions such as wonder, awe, and reverence. The religious person tends to show a concern for values, moral and aesthetic, and to seek appropriate action to embody these values. He is likely to characterize behaviour not only as good or evil but also as holy or unholy and people as not only virtuous or unvirtuous but also as godly or ungodly.

As a feature of human existence, religious life can be studied, for example, in terms of psychology, sociology, and history. Among the first books in the psychology of religion were two by Jonathan Edwards, an 18th-century American theologian: ***A Faithful Narrative of the Surprising Work of God*** (1737) and ***A Treatise Concerning Religious affections***(1746). About a century later, during a period of religious "revivals," interest developed concerning the age at which conversions most often took place—the period of adolescence. Reflections on such facts, and in this sense the psychology of religion, only came, however, with the works of two American psychologists: Edwin Diller Starbuck's ***Psychology of Religion*** (1889) and the classical treatment by William James's ***Varieties of Religious Experience (1902).*** Generally, the psychology of religion has shown that though religion for some is a crisis experience, for others it is a natural growth.

As psychology became more analytical it became more interested in the abnormal, in neuroses and dreams, in the techniques of hypnosis, and in the kinds of experience induced by drugs. When Freud spoke of religion as an illusion, he maintained that it is a fantasy structure from which a man must be set free if he is to grow to maturity; and in his treatment of the unconscious he moved toward atheism. The study of the unconscious by the Swiss psychiatrist Jung, however, suggested that dominant archetypes (implying innate tendencies to form symbolic images) are supplied by a racial unconscious, thus providing a psychological approach to belief in God.

Classifications of individuals

In classifying individuals into different types, psychology has distinguished between religious people who are: "extrovert" or "introvert" (Jung), "healthy minded" or "sick" (William James), and "objective" or "subjective" worshippers (J.B. Pratt). There is always the danger, however, that psychological distinctions may beg too many philosophical questions.

One of the most widely accepted studies of religious experience in regard to feelings was written by the modern German Protestant theologian Rudolf Otto. In his ***Idea of the Holy,*** Otto analyzed what is distinctively religious in terms of the unique concept of the "numinous"; *i.e.*, something both awesome and appealing, both fearful and attractive.

Psychology, however, is concerned not only with individuals but also with what is known about group behaviour, which can also be of importance in any study of the Christian Church or other religious institutions regarded as communities of religious people. The authority of a religious leader, like that of all leaders, is derived from his symbolic character and the extent to which the leader and his followers share a common ideal.

The findings of sociology. The ideas and images of a religion are much influenced by the social culture in which it emerges. Some of the oldest social institutions and practices, such as those concerning birth and death, marriage and the family, and art and music, have developed in a religious context. Religion has often been a driving force in the reform of social abuses, but also it has been associated with reaction and oppression. More recently, the sociology of religion—influenced by contemporary sociology—has been concerned with making use of sociological criteria and of demographical and statistical studies in planning the church's mission and appraising its significance.

The findings of the history of religions. Conclusions in the history of religions have been largely determined by the particular ideas of man or history with which the study was approached. Some scholars have supposed that at the dawn of human existence there was a belief in a single god and that only later there occurred a development into a belief in many gods as well as animism (a belief in souls or spirits in man and other aspects of nature). Other scholars have supposed an evolutionary development of religion, which only reached monotheism—considered to be the highest form of religious belief—after a long period of purification. The two approaches sponsor, respectively, two contrasting myths about primitive man. According to the one, there was once a golden age of innocence and harmony; according to the other, the life of the earliest man was hasty, brutish, and short.

Granted the ubiquity of religion and its diversity, historians have found no universal essence expressible in terms of common beliefs. What is probably common to all religions is nothing more than the claim that reality is not restricted solely to what is yielded by sense experience itself.

The ambiguity of religion's influence on culture

The role of religion in culture. Religion has had a strong but ambiguous cultural influence. The thought that a man depended for his life and existence on a power not his own has encouraged some persons to be lazy, as it has inspired others to greater effort. A conviction about another world has led some religious people to disvalue human life; it has led others to view human life as having the significance of a state of probation. It has been plausibly argued by some (*e.g.*, the German sociologist Max Weber) that Protestantism provided a seedbed for modem capitalism; Catholicism, according to others, easily accommodates a Socialist point of view.

Because a religious view is generally associated with a conviction about the inadequacy of "things seen and temporal," religion as a cultural influence usually shows itself dissatisfied with things as they are. Often, however, when confronted with novelty, religion has tended to be conservative. Thus, religion has alternately opposed or fostered social and cultural development.

VIEWS WITH TRANSCENDENT REFERENCES

A situation is regarded as religious when through its spatiotemporal features what can be termed depth or another dimension can be disclosed objectively. In this sense, there cannot be such a thing as a religion that is nontranscendent. On the subjective side, there will be a matching self-disclosure, a "coming to one's self" that occurs as a response to a vision of the eternal in and through the temporal.

Relation to an ultimate power, or being, to values, or to ideals. Different religious approaches can be distinguished by the different interpretations they give of what is objectively disclosed, of what in this sense is the transcendent. In primitive religion, for example, the transcendent is always interpreted in terms of an ultimate power or activity expressing itself, whether singly (monism) or with multiplicity (pluralism), through the objects and events of the world.

Animism, totemism, and interpretations of cosmic power

Animism views the world as having life, power, and feeling as do men. A monistic view of the universe is conceptually akin to the view according to which people or objects exert the peculiar influence they do and have the strange significance they possess because of mana—a power or force somewhat similar to the scientific concept of energy—that they embody. Animism becomes more diversified and pluralistic when it becomes spiritism, which locates the cosmic life, power, and feeling in particular objects. Totemism involves a highly complex system of beliefs and practices whereby an animal or plant becomes a totem, or a focal symbol for the life and well-being of a tribe. Just as tribal communities are sustained by a power that the totem symbolizes and expresses, so the patterns of tribal behaviour are maintained by taboos. Persons, things, and behaviour are taboo, or are prohibited to members of a society, when they are judged to be so highly charged uith sacred power that ordinary "profane" persons must keep their distance. (See also the articles PRIMITIVE RELIGION; ANIMISM; TOTEMISM.)

These primitive viewpoints have a certain conceptual kinship with what the more sophisticated religious viewpoints have labelled with such terms as theism, polytheism, pluralism, and Idealism. Theism interprets the one cosmic, life-giving power in personal terms-different versions varying in their views of the adequacy of those personal terms. Polytheism posits a multiplicity of cosmic personal powers on whose activity (whether in cooperation or conflict) the universe depends. Pluralism views cosmic power as mediated and expressed through a multiplicity of ultimates (*e.g.*, finite persons) or otherwise views the universe as best understood in terms of ultimate atomic units, with no claim made for the absolute supremacy of any one of them. In this way, pluralism—even of a personal kind--differs from theism, which holds that God is a Supreme Person.

Absolute Idealism maintains that activity is an ultimate category but makes no claim, as does theism, for this activity to be personal. Instead, it takes a biological organism as its dominant model. Theism, like deism, has sometimes posited an ultimate personal power or being beyond, above, and certainly separated from the changing scenes of life, whereas absolute Idealism posits an ultimate power or being that is considered to be the whole, of which the changing scenes of life are but a part (see also THEISM; POLYTHEISM; DEISM; IDEALISM).

Interpretations of salvation

Seeking salvation in a life beyond. Religion is not merely a matter of being aware of a transcendent dimension nor is it merely a claim for a broader and more comprehensive view of the reality. Fundamental to religion is the conviction that through a right relation with a cosmic power or powers, man will find his salvation. Various views of such salvation have been held. Salvation has been regarded as something attainable only after this life. Other views, however, tend to posit a salvation for man through escape lather than fulfillment. Alternatively, salvation may be viewed as something anticipated in the present but fulfilled perfectly after this life. Salvation also has been interpreted in terms of fellowship with God or as a state of bliss needing no God (as by the early 20th-century British philosopher J.M.E. McTaggart), as a state of ultimate peace that arises when man sees his peculiar and rightful place in the whole universe (as by Spinoza), and as a state of bliss in which man cannot properly speak of himself as a self-conscious individual as in the Buddhist state of Nirvāṇa.

VIEWS WITH ANTHROPIC REFERENCES

Inner attitudes and dispositions. A religious view of the universe contends that a new dimension and depth can be disclosed within the person who responds. Though religious faith has its characteristic inner attitudes and dispositions, they must be of a transcendently self-involving kind, and there must be a depth to any attitude or disposition before it can be called religious. Thus, the attitude of awe is related to the feeling of fear. For fear to become awe, however, it must be characterized by a particular depth and self-involvement that come from responding to the presence and activity of God, or of the sacred or holy that call it forth.

The concept of conversion

Religion relates to the whole of a man's personality and because of this totality of human response, people speak of "conversion" in relation to religious attitudes. Generally, a person who becomes religious or ceases to be religious undergoes a profound transformation. Persons who have become converted to religion speak of the world as having taken on a fuller and richer dimension; those for whom the religious vision has disappeared speak of a world as having become flat, dead, and bleak.

Behavioral discipline with prescribed practices. Many religions bind their adherents to specific practices and particular moral codes. Thus, conversion has often shown itself in radical changes of behaviour; *e.g.*, an alcoholic becoming a total abstainer. Such behaviour as murder, lying, breaking promises, stealing, and committing adultery have been condemned by the world religions. So strong is the ethical element in Confucianism that some regard it more as an ethical system than a religion. Yet, ethical (and ceremonial) codes can be transformed imperceptibly into no more than current social conventions and mere customs. Whether such codes have changed or not, their range and detail vary widely. Pork is eaten by Christians but is considered to be unclean by Jews and Muslims. Muslims and Buddhists abstain totally from alcohol; Christians and Jews need not. A Sikh will not shave his beard; but Hindus, Christians, and Muslims are free to do so if they wish. In contrast with Christians, Buddhists will not kill animals, and Muslims may practice polygamy.

Participation in a social institution. Whatever the diversities, religious faith is not only self-involving, but it has a social dimension as well. Hermits apart, religion brings people together as children of one family having a common father. For Christians, the significance of the universal religious community, the church, has been variously interpreted. Some, with a Protestant emphasis, have viewed the church as a voluntary institution created ad hoc for the convenience of its members to enable them to gather together to worship, to sing hymns, and to share common interests and beliefs. The Catholic view is that the church is a social institution that is derived from God and whose structure expresses the givenness of God himself.

To be of religious significance, however, social practices and moral codes, like inner experiences must have depth, a transcendent dimension, or they become superficial and dangerous parodies of religion, all the more dangerous for being in their outward features so similar.

II. The meaning of religion for persons within the particular religions

THE VIEW FROM WITHIN AS PRIVILEGED

The assertion that the view of religion from within is privileged needs careful analysis.

First, religious faith is logically privileged insofar as it is characterized by a self-involvement, commitment to which partial commitments can only point. A temporary loyalty, however intensive, is at best a distant pointer to a conversion. Further, because religious faith is grounded in a disclosure, there is something logically privileged about it in the same sense that some are "privileged" to understand a joke when others do not. Yet, even though religion has a disclosure basis, it is still true that just as there are techniques for jokes so also are there techniques for meditation, whether in Christianity or in other religions. By virtue of such techniques men can have a reasonable expectation of a view of religion from within. In another sense, the view that religion from within is privileged may merely mean that if a man believes something and is committed, he is more involved than a man who does not believe.

One aspect of the logically privileged position of religion might be called its semantic privilege; *i.e.*, the fact that a religious vision cannot adequately be expressed. One fundamental problem for religious language, according to linguistic analysts, is to discover more reliable rather than less reliable ways of talking. One need not presuppose, however, so fundamental a distinction between the sacred and the secular that men become committed to total silence on religious matters. When St. Paul, for example, wrote of being "caught up" (in II Cor.) he "heard things that cannot be told, which man may not utter." If this statement of Paul's were generally true of religion, however, religious people would be so privileged that they would be living in a segregated silence.

Some scholars have argued that the privileged character of religion makes it unsuitable as a proper study for the philosopher, who must in principle be detached, not committed, and have an openness to all truth. The lack of finality in philosophical thought is contrasted with religious commitment and the final claims sometimes made for religious doctrine. Nevertheless, insofar as anyone has a coherent world view, there will be some degree of commitment.

Religion is not, however, altogether beyond argument, and those who are outside a religion can still have some inkling of what is being discussed within a religion and the manner in which it is being discussed, especially when the social, cultural, historical, and psychological embodiments of the religion are described. For this reason Western Christians and Jews, for example, are able to know something about the primitive religion of an African or Indonesian tribe.

Thus, much about religion can be known by those outside it, however, views about the nature of religion and definitions of religion have a systematic inadequacy about them. Like everything of the spirit, religion cannot be described so as to make clear to the detached observer the characteristic quality and depth of religious awareness and commitment.

THE DIMENSION OF RELIGION FOR INSIDERS

The essence or core of religion. For the insider, the essence of religion is given in a moment of vision and disclosure. Friedrich Schleiermacher, a German philosopher of the 18th and 19th centuries, described the basic religious experience in terms of a kiss or an embrace. Attempts to understand such a unity can only be made in terms of the particulars into which the unity subsequently breaks, and such particulars then fall broadly into subjective and objective compartments.

Faith, prayer, and ritual

The subjective and objective aspects of religion. Faith describes a subjective state that accepts what a disclosure discloses and is akin to personal trustfulness, to a conation or striving that, according to Spinoza, all living things display. Prayer is the utterance of words (rite) with or without some dramatic context (ceremonial) designed either to carry one into the presence of what is worshipped or to express appropriate sentiments in the presence of what is worshipped. Most prayers incorporate words that function in both ways. Ritual is especially concerned with events in human life that have disclosure possibilities and in which mystery is at its highest.

Mystery in the context of religion refers to situations, such as birth, reproduction, death, and suffering, in which there are numerous possibilities for new insights and yet further insights. Public worship must constantly renew and realize in the liturgy the possibilities of the past disclosures. If the outward expressions and forms come to dominate, ritual can become an empty shell, and religious practices can become devoid of religious significance.

Effects of religious beliefs and practices. One of the effects of religious beliefs and practices is sacralization, a process in which certain persons, days, or objects become regarded as sacred. If such objects are granted more than the status of symbols, they may become objects of idolatry or superstition.

Belief in salvation, which often accompanies religious commitment, can have various practical results. If salvation is viewed as something that inspires progress and may be accomplished in the realm of time, such doctrines of salvation encourage social reforms and projects that envision an abundant life for humanity. If, on the other hand, salvation is viewed as something that is beyond the realm of time and set entirely apart from this world—something for which at best this world is a probation and at worst a sink of misery and iniquity from which the sooner man is released the better—such doctrines of salvation can be excessively individualistic and may even encourage oppression and tyranny.

Conclusions about nature and history

Religious belief has sometimes led men to detailed conclusions about nature and history. Good harvests have been interpreted as due rewards for appropriate worship or good behaviour, or both; calamities have been viewed as the results of sin, either ceremonial or moral. If God is believed to be in control of history, a nation that does what is right and follows his guidance, as expressed through its prophets and other religious personages, is expected to experience national prosperity and success. In previous periods, when this did not occur, the ensuing calamities were attributed to the backslidings of earlier generations.

Many observers of religion claim that in the modern world few would suppose that God intervenes in this direct and predictable way. According to this view, God's activity in the world, apart from being expressed in its constant creativity and conservation, is effective through man's own intellectual and physical activity. Insofar as man's own creativity is exercised, however, within the framework of the order that the world displays, and in no way violates it, one cannot exclude a similar creativity on the part of God. Admittedly, the fact that man expresses his activity through an intermediate organism (his body) indicates that there is no exact parallel between God and man; nevertheless, because God's activity terminates with the universe, the analogy with God's activity might very well be expressed in the number of ways in which man can effect creative development in his own body.

INTERNAL CRITICISMS OF RELIGION

Internal criticisms of religion have their basis in the imbalance that occurs when one aspect or one understanding of religion is allowed to dominate the rest. Heresies have arisen when one way of understanding has been developed without balancing it with another. In the development of doctrines concerning the nature and person of Christ within Christianity, for example, heresies arose when a particular model (*e.g.*, that of fatherhood and sonship) was believed to be capable of infinite development. The model of the Father–Son relationship was pressed too far, and the Son was subordinated to the Father in a way inconsistent with Christian orthodoxy, thus leading to what became heresy. Sectarianism develops when religious insights are associated exclusively with one particular doctrinal or theological phrase, such as justification by faith, or with one particular theological view regarding religious practices; *e.g.*, baptism. Because religion is at once infinite and mysterious, it is important that religious belief does full justice to a wide variety of approaches.

Another criticism of religion has been that it has tended to be overintellectual; and when this trait has been combined with moral laxity and factional rivalries, it has led to protests about the arrogance of intellectualized religion, often leading to the opposite error of supposing that belief does not matter as long as common sentiments are shared. Religious believers have not always recognized that for the most part their belief explicates metaphors, images, and symbols. Though ways of religious reasoning are appropriately informal and variegated, having their origins in a multitude of images and symbols, it nevertheless is considered a religious duty to produce the most reliable overall discourse based on the various images and models.

Balancing intelligibility and mystery

The fundamental difficulty of all religious understanding, however, is to balance intelligibility and mystery. If the intelligibility is neglected, religious belief can become dishonest and religious men can lose integrity; if mystery is neglected, there may be splendid controversy and exercises in logical appraisal, but the heart of religion will have disappeared.

The basic difficulty of all religions and of historical religions in particular is to effect a constant rebirth of symbols in changing cultures. In the course of time some of the most powerful images and symbols lose their fertility in promoting ideas that inform a religious community. This might be said of the image of sacrifice in the Christian religion. Religious practices and institutions, though they may have social merits, can all become stereotyped routine, as happens when they fail to preserve a sense of reverence and fail to disclose the givenness of the sacred or holy. Because religious belief is so important and influences all aspects of a society, there is a tendency for religious institutions to become authoritarian and oppressive. If a religious institution becomes interwoven with political views it can become tyrannical. Religion's only compulsion, according to some scholars, must be the compelling power of a vision, as the modern English–American philosopher Alfred North Whitehead expressed it: "The power of God is the worship He inspires." The authority of any religion is the authority of a vision, the authority of that which, in being disclosed, inspires men and leads them to fulfillment in their lives. For a Christian, the final authority is the love of God in Christ, and love is not love if its power is anything but inspiration. For other religions there is the compelling inspiration of that to which—Nirvāṇa or the Qur'ān, the Buddha or Muhammad—point.

Internal criticisms of religion usually focus on such themes as narrowness, sectarianism, traditionalism, conventionalism, materialism, and immorality. Some criticism is also reserved for religiosity, which, though granting a dimension of faith, treats faith in an altogether superficial and often unbalanced way. Religiosity represents an excessive preoccupation with religion that is depicted in an incoherent and oversimplified relating of religious faith to intellectual views and social and personal practices.

III. Attitudes toward religion by persons outside particular religions

THE REJECTION OF RELIGION OR RELIGIOUSNESS

Because religious commitment is so all-embracing and tends to influence thought, feelings, and behaviour, it is not surprising that there are many reasons why religious claims have been rejected.

Rejections because of alleged incoherence. Religious claims have been rejected because of their alleged logical or moral incoherence.

Alleged logical incoherence. Logical incoherence may arise internally or externally and in relation to different issues. In regard to internal coherence, critics have maintained that man should be able to expect that God would see to it that there could be no possibility of ignoring his existence or of making mistakes about religious beliefs and behaviour, if religious convictions are so important. They have also claimed that it is altogether too naïve, though inevitable, to think of God as made in the image of man. Some have rejected theistic belief because of the incoherence of the idea of God, which must—they claim—combine so many incompatible predicates; *e.g.*, God is eternal, yet acts in time, or he is loving and yet incapable of suffering or feeling.

External incoherence, in conflict with other views of world

Religious beliefs have been alleged to be externally, as well as internally, incoherent because of their conflict with other views about the universe, especially scientific views. The doctrines of heaven and hell, in particular, which have given great personal and social significance to religious belief, have been rejected by many critics when these doctrines were viewed literally. Yet it has been the supposed actuality of heaven and hell that has given religious persons their hope and their terror respectively. Absolute Idealism, it has sometimes been alleged, is incoherent insofar as it states that time is not "real" and that evil does not really exist. This is not to say, however, that there is no temporal succession or nothing evil, claims that would be obviously incoherent. What is being claimed is that within a particular interpretation of the universe, time and evil are not left as ultimate categories but are in some sense derivative from other categories.

It has been argued that by far the greatest problem of external incoherence that belief in God has to face is that of the evil and suffering that characterize the world. Critics have stated that if God cannot rid the world of evil and suffering, he is not all-powerful; if he could, but he won't, then he isn't all-good; if he is powerful and good but not all-wise, then, even though he is trying his best, there are bound to be disasters. The most serious classical expression of this problem was given by David Hume, in his ***Dialogues Concerning Natural Religion (1779).*** With such considerations in mind, some philosophers, such as John Stuart Mill, have been willing to argue for a limited God—*i.e.*, the great fellow-sufferer who understands and has compassionate sympathy.

Alleged moral incoherence. Though religious conviction shows itself in moral behaviour, it has been argued that religious people have not shown outstanding moral qualities. An 18th-century English philosopher and churchman, Bishop George Berkeley, when presented with this objection, remarked that nothing evil can be attributed as such to the Christian religion and that the only legitimate comparison is that between a person who is a Christian and what the same person would have been otherwise. The distinctiveness of the Christian faith, however, has sometimes been supported by arguing a stark contrast with morality. The 19th-century Danish philosophical theologian Søren Kierkegaard, for example, by a too literal misreading of the biblical story of Abraham and Isaac (Gen. 22), supposed that religious obedience must be in radical opposition to moral duty. However that may be, religious men often may be only too well aware of their moral lapses, their sins, and for this very reason they seek the grace and power of God. The good that they would do they do not do, and the evil that they despise they continually do, as St. Paul noted in his letter to the Romans. In this moral predicament, those with a Christian commitment believe that the grace and power

The problem of sin and grace

of God comes to inspire and release them from the dominion of sin. This does not mean that the Christian never sins, but it does mean that he is assured of ultimate victory over sin. The Christian Church is viewed not as a society of saints but a school for sinners.

The exclusiveness of religious sects is regarded by those outside the sects as hardly to the sectarians' credit. For Christians, sectarian exclusiveness is viewed as a scandal to the gospel that they preach. On the other hand, the criticisms of Puritanism that hold it as inevitably negative and oppressive sometimes fail to see that it may be neither negative nor oppressive if it is grounded in a spiritual and religious vision.

The doctrine of grace (the view that God grants man abilities that man does not merit by his own efforts) has sometimes appeared to make God himself—interpreted as the spirit dwelling in a man—the actual agent of good behaviour. In this way, some interpretations of the doctrines of grace have compromised man's freedom and come close to denying man responsibility for his actions.

Outside Christianity, critics have pointed to the gap between religious profession and moral action, though within Christianity, with its strong emphasis on moral transformation, the gap has been very wide and the criticism most challenging. In Hinduism, for instance, Gandhian reformational and nonviolence ideals have not mixed well with social corruption or with the type of neutralism that allowed China to persecute Tibetan Buddhists. Again, the Buddhist who goes to a temple is not necessarily compassionate as his religion dictates, and the Muslim who attends services in a mosque may be less filled with an inner sense of justice and patience than with thoughts of a holy war. In Sri Lanka (Ceylon) and Vietnam nationalist loyalties have given rise to a violence untypical of Buddhism. In the last resort, however, each religion will appeal to its doctrine of salvation when presented with a gap between its moral ideals and the actual actions and behaviour of its adherents.

Other grounds for the rejection of religion. *Rejection of historical beliefs, practices, and institutions as spurious or irrelevant.* When a religion appeals to historical events, other grounds for its rejection arise. The Old Testament view of history appears to have been exceedingly selective in order to emphasize a particular point about God and his activity. The miracles of Jesus—both those relating to his own person (his birth and Resurrection) and those that he himself performed (especially nature miracles)—conflict with what men experience in the normal course of their natural lives and experience. Prayers requesting favourable weather, plentiful crops, or safety in a journey are characterized by many as spurious and irrelevant. Ideas of God intervening in the universe, according to such critics, satisfy neither science nor religion. From a scientific point of view, "laws" of nature are no longer viewed as divine prescriptions; and the word law becomes, in fact, misleading. Furthermore, in order to allow for miraculous intervention of this kind, God's providential care is viewed as a compromise. He thus becomes the absentee landlord who absented himself from the world, which must take care of itself except for some spectacular visitation. According to this view, the only coherent way to speak of an intervention of God is to interpret it in the context of personal intervention.

Religious institutions have been criticized on the grounds that they conflict with the ideas of the founder and are supported by claims that cannot be historically verified. These claims, according to critics, depend on taking certain historical events on which the religion is founded, and reinterpreting them by theological speculation or a very full imagination, to produce, for example, a doctrine of papal supremacy according to which Christ is believed to have given explicitly to the successors of St. Peter final jurisdiction over the church.

Rejection of religious sentiments or dispositions as valueless. According to some views, anyone who prizes "another world" must despise this world and be uncertain in his attitude toward the world around him. In this way, it is said, religion dries up the sources of its activity and attacks such happiness as this world can provide—though promising happiness hereafter, which has been called "pie in the sky" or "opiate of the people" by critics of religion. A humanist concern to liberalize and relax laws (*e.g.*, on abortion and divorce), to abolish capital punishment, and to encourage birth control has always been opposed, according to humanists, by Christian orthodoxy, which they interpret as having a negative and conservative attitude that has proceeded from a nervous fear of a decline in moral standards. At the same time, humanists would continue, moral standards have hardly been upheld by sectarian strife and persecutions. Further, they point out, too often the church, in its desire to indicate what abides, has confused what is abiding with current social and political institutions and traditions inherited from the past, generally resulting in an illiberal obscurantism and a reactionary outlook.

Humanist criticisms of religion

Some critics of religion have contended that almost all scientific progress has been hindeied by religious beliefs and attitudes. Biology, physics, and geology, they have claimed, only made the rapid progress that they did when they were freed from a context of religious belief by the 17th-century philosopher René Descartes, who devised a metaphysical myth of the separation of mind and body.

Naturalistic or skeptical views of the origin and development of religion. In the matter of the origins and development of religion, many (*e.g.*, the psychologist James Henry Leuba in his *Psychology of Religious Mysticism* [1925]) have argued that there is a close connection between mysticism and hallucination, between hysteria and ecstatic institutionalized inspiration as, for example, in Pentecostal churches. Religious people, according to such views, often have personality weaknesses and are psychologically disturbed. Freud, the founder of psychoanalysis, maintained that inner conflicts--often the result of repression, particularly in relation to sex—become expressed in peculiarities of behaviour and mood, especially in the vivid imagery of dreams that erupt from the unconscious area of one's personality. By comparing the symbolism of dreams and mythology, Freud held that belief in God—in particular, the father image—merely perpetuates in fantasy what the individual must in actual fact overcome as part of his growth to maturity, thus giving religious belief a treatment that not only made belief in God unnecessary but positively unhelpful.

Carl Jung, a former disciple of Freud, gave a different account of the psychology of the unconscious. Each person displays a libido, a fundamental striving that is creative and purposive and of which there is evidence in the symbolic language of dreams. Behind all such symbolic language are archetypes (innate tendencies to form symbolic images), which all humanity shares and which inspire a person to move toward a balanced integration to which the energy of the libido would creatively move, if given proper freedom and encouragement. Thus, Jung posits a racial or impersonal unconscious in which, at the deepest level, all individual human beings share. Jung's archetypes raise the metaphysical question of whether they are symbols of an existent God or gods—a question that psychology leaves open. For many psychologists it is a question of little interest, because for them the archetypes themselves suffice in practice.

Problem of the genetic and reductionist fallacies

In addition to such naturalistic or skeptical views about the origin and development of religion are other claims that religion is merely an infantile reaction to fear, a more or less harmful sublimation of sex, a projection of wishful thinking, or a social device for use in the class struggle. On the other side, however, it is likely to be pointed out that one must be careful not to indulge in the genetic fallacy: no account of the origin and development of anything, of religion in particular, is necessarily a reliable analysis of what that particular phenomenon is now; a single explanation of the origin and development of a phenomenon as complex and variegated as religion is difficult to describe and maintain. It is also necessary to beware of the "really only," or reductionist, fallacy. To say "x is really only *y*" is, in effect, denying the significance of *y* language despite the fact that y-talk as well as x-talk already occurs; *e.g.*, persons are really only "machines," or worship is really only a social occasion. Oversimplification streamlines discourse at the cost of adequacy and truth.

Some have thought of religion as no more than a body of stories designed to encourage a noble attitude toward life and humanity. If, however, one asks why or how these attitudes encourage and why a particular attitude is valued, what begins as a simple account of religion becomes, in the end, as complicated as any. Another criticism of religion, arguing for its redundancy, claims that the progress of man in society can and should be determined by scientific considerations. This contention, however, goes beyond the particular conclusions of the individual sciences; it is to make a philosophy out of science. On the one hand, such a scientific view of man and society would be open to philosophical criticism, not the least if it were suggested that man's subjectivity—that which makes him the unique person he is—has to be analyzed in terms of the objects of science. On the other hand, if science becomes a philosophy, it might be said to have assumed a religious dimension itself.

Ambiguities in attitudes of persons in Christian societies

In the realm of religion in the latter part of the 20th century, in what might still be called the Christian societies of the West, the attitude of very many people lies in an intermediate zone between religious belief and atheism, but the content appears rather to be given to agnosticism. Such persons believe in God but dislike formal worship, pray only on exceptional occasions, and find it difficult to have a sense of sin but admire saintliness. They are critical of the need for a Christian ministry except insofar as a priest or pastor can show sympathy and social concern. They are distrustful of dogma and critical of Christian sectarianism. They may be uncertain of Christ's divinity, but the words and example of Jesus are viewed as a guide to the good life. This outlook has many affinities with the "natural religion" of the 18th century in which the ethical example and teachings of Jesus were emphasized. Though, as in the 18th century, there may be an intent to reject revelation, persons holding such an outlook may rather be rejecting certain stylings of Christian revelation.

Examples of occurrence of such a "natural piety" can also be found in religions other than Christianity, though significantly not in Islām—unless the Baha'i movement be taken as an approximation of this outlook. This attitude, for example, has provided the basic cohesion for the State of Israel in the latter half of the 20th century. Further, the spread of technology has gradually been alienating many Hindus and Buddhists from their traditional beliefs, but the Hindu has continued to treasure his spiritual ideology, which may well give to technological development its needed direction and wider setting. Buddhism in Japan, and perhaps elsewhere in the East, is still valued in the 20th century insofar as it supplies a local religious dimension to a society whose public and industrial life has been increasingly Westernized. Thus, an attitude has arisen that is sympathetic to the broad claims of religion, but has been critical, if not disdainful, of theological dogma and rivalries.

THE ACKNOWLEDGMENT OF I OR RELIGIOUSNESS AS VALID

Traditional **iustifications.** *Religion* as *pointing* to an *ultimate* power, *being,* or value. More generally, persons who are outside the particular religions and who have nevertheless acknowledged religion as significant often seem to base their views on a fundamental feeling of absolute dependence. The grandeur of the universe, the character of the moral struggle, reflections on human nature, and an awareness of moral values inspiring men to reform society have all pointed men to an ultimate power or being—a "power, not ourselves, which makes for righteousness," according to the 19th-century English poet Matthew Arnold.

The fundamental difference in the latter part of the 20th century between the secularist and the religious person most likely has been between someone who takes a narrower and someone who takes a wider view of humanity. That there is an acknowledged need in modern times to give a moral direction to technology seems to many to bring with it the need for a religious view of the universe, even though they may not themselves be adherents of a particular religion.

Religion as producing wholesome spiritual or moral effects. Others point to examples of the wholesome moral and spiritual effects that religion has had. They mention that society has ceased to practice child exposure and there has been a notable development in the status of women in society. Religion, where it is not parodied, misrepresented, or misunderstood, broadens rather than narrows vision. Insofar as human nature is inadequately understood, if no place is granted to the spirit of man, human nature, it is argued, will never find satisfaction except through the self-realization and self-fulfillment that come from responding to the inspiring ideal.

Alternatives to traditional beliefs, practices, and institutions. The quest for authentic existence. In the 20th century various alternatives to traditional religious beliefs, practices, and institutions have become apparent. Chief among these is the quest for authentic existence. This has been encouraged and portrayed by various Existentialists (those who view man in terms of his existing thoughts and actions rather than in terms of his "essence"), who have been concerned in one way or another with emphasizing the significance of certarn situations. In this way, they have given their own versions of salvation —that situation in which a person finds his true significance. For some, such as the German philosopher Martin Heidegger, a sense of authentic existence is given to each person when he realizes his true subjectivity, which his life in the world and his social transactions so often conceal. Authentic existence is often contrasted with cosmic **anxiety**—*i.e.,* anxiety of a deep and far-reaching kind to which the antidote is to find oneself and one's freedom in a total commitment to what is called the ground of Being.

Existentialists of an atheistic persuasion, such as the philosopher and Nobel laureate Jean-Paul Sartre, regard human existence as absurd and other people as hell, because, though one needs other people, they can never be other than "other people^{w}—their subjecthood, their freedom is inaccessible. Love is, thus, doomed to permanent frustration. The need to know others like oneself is matched by its impossibility. According to Sartre, this condition only reflects the absurdity of man's own existence, which is always attempting to overcome a radical estrangement between man as the object of scientific study and man himself (en soi) and the subjectivity man knows in consciousness (pour soi). Suicide is the final absurdity, for in getting rid of en soi, what man is, pour soi disappears at the same time.

Religious Existentialism as a God-given fellowship

This pessimistic estimate of human life and its apparent absurdity, however, has been converted into a religious view by other Existentialists, such as Gabriel Marcel, another French philosopher, who point to a participation —a mysterious self-involvement that persons can have intersubjectively with each other—in a kind of fellowship that is viewed as God-given. According to this view, man needs to open himself to the presence and grace of God for a dynamic transformation in which the mysterious transcends the purely problematic. Common to all Existentialists, however, is the view that the authentic man is not merely satisfied with playing a role, with being a cog in industrial society. He breaks free and realizes himself —for better or worse. One way or another, the quest for authentic existence is to discover the means by which man can recapture and enjoy occasions of self-disclosure. So significant are these occasions that they have been viewed by some theologians to be the paradigm for the kind of situation that the Christian gospels recount.

Secular religion. Another feature of 20th-century development has been society's rediscovery of the significance of the secular. This change has led to an outlook and attitude that has been characterized as "religionless Christianity," a Christianity influenced by its residual social and political ideal, but bereft of its specifically religious practices, doctrines, or institutions. Such practices as traditional intercessory prayer are dismissed as empty approximations to magic; doctrine is condemned as outdated and expressed in terms of past cultures; institutions are criticized as oppressive and conservative.

Behind all this suspicion of structures and doctrinal schemes and practices, however, is a desire to get back to basic principles and origins, to learn again what is distinctive about the religious point of view. According to some proponents, such a goal might be attained by beginning with the secular, with activities in the secular world, not least with compassionate service, by seeing where the need arises for religious conviction and by ascertaining what contribution faith will make to secular endeavour. Though secular religion broadens out into a more sympathetic and a more positive attitude than agnosticism, it is never as explicit or particularized as orthodoxy.

Marxism. Marxism, which provides remarkable evidence of the power of dominant key ideas to inspire and direct man, is undoubtedly one of the greatest challenges to traditional religious belief. Based on the socio-economic philosophical thought of the 19th-century thinker Karl Marx, Marxism can be said to be a quasi-religion on two counts. First, Marxism had connections with the metaphysics of G.F.W. Hegel, an 18th–19th-century German philosopher who interpreted reality in terms of a spiritual Absolute. Furthermore, the thinking of Marx had religious overtones, whether from his own Jewish background or from a Christian atmosphere, not least in Britain where he lived from 1849 to 1883. Second, Marxism can be called a quasi-religion insofar as it calls from its followers a devotion and a commitment that in their empirical character greatly resemble the commitment and devotion that characterize religious people. Marxism has undoubtedly fired the spirit of man and given to revolutions, whether in Russia or China, a powerful direction that has maintained stability and avoided anarchy. Furthermore, like a religion, it has provided themes of fulfillment and hope—a revolution interpreted as the initiation of a Communist world society that would be a final consummation. There are many logical similarities between the doctrine of the Marxist millennium and the Christian doctrine of Christ's Second Coming. Marxism has also stressed the significance of cooperating with the immanent spirit of the times—something comparable to the providence of God—in economic and military struggles that are viewed as the travail by which society would be reborn. The main difference between Marxism and Christianity in the 19th and eariy 20th centuries, according to some scholars, was that for many the Christian vision encouraged men to endure tyranny, while the Marxist view inspired men to rebel. Yet, once it can be established that religion is not the servant of oppression, is not necessarily linked with an illiberal regime, and does not use concepts of "other worldliness" to make men content with tyranny and injustice, then religion may yet have a place in the Communist state. Such a religion would not have to concern itself with the kind of supernaturalism that Marxism now rejects; it would not have to appeal to an invisible world entirely other than the present world. It is not without significance that Marxism has its own form of public ceremonial and its own language of glorification. If it has to be granted that many religions have a ceremonial, a symbolism, and a moral code that has lost the vision they once had, Marxism is a social program, a doctrine, and a ceremonial searching for a vision that haunts it and that may at some time bring it to fruition. In this regard, Chinese Marxism is particularly significant insofar as Marxism in China cannot escape some interweaving with Chinese Buddhism. Chinese Buddhism brings with it a natural framework of absolute Idealism, which may yet supply Marxism with the spiritual dimension that for many critics appears to be Marxism's main inadequacy, something it lost when it shed its Hegelian metaphysics and became the anti-God Materialistic world-view of the U.S.S.R.

Marxism as a quasi-religion

IV. The professional philosophy of religion

HIST OF THE PHILOSOPHY OF RELIGION OR RELIGIOUS PHILOSOPHY

Most philosophies have incorporated religious views in the wide sense of being concerned with a reality beyond appearance, and in this sense they have provided a philosophy of religion.

Developments* in *the West. *Ancient and medieval concepts.* For the Greek philosophers Plato and Aristotle, wonder was the beginning of philosophy. From such wonder, according to Plato, emerged religious knowledge that was also mediated through Ideas, eternal entities or concepts in which the things of time participate. In performing every good act, man realizes his link with eternity and the Idea of the Good. For the moment, however, man, as in a cave, is chained by his earthly existence so that he cannot see the light outside; he can only see shadows on the wall, which are signs and tokens of the eternal light behind him. This was Plato's way of styling the relationship between time and eternity, between appearance and reality, and it is a styling that found a particular welcome in the Christian tradition and not least by Christian Platonists, whether of the 2nd or 17th centuries. Plato's philosophy also led to belief in God, and his *Timaeus* is a philosophical creation story.

Religious ideas of Plato and Aristotle

Aristotle, impressed with organic life in man and animals, took as his fundamental category growth and development. The nature of anything was thought of as a form by which its movement and development as an organism was to be understood. It was as if the form supplied the driving force. In this context, God was thought of as pure form, as final cause, and as prime mover. Aristotle provided for St. Thomas Aquinas, the great medieval philosopher of Western Christendom, the foundation on which he developed Scholasticism, which has been a distinctive feature of Christian philosophy of religion since the 13th century. Other medieval philosophers, such as Erigena, with his pantheism (God in all); Abelard, with his critical questions; Eckehart, with his mysticism; and Duns Scotus and Bonaventure, with a wider view of reason than could be contained in the Scholastic philosophy, all illustrate the variety and independence of Christian thinkers.

Modern concepts. Descartes, the "father of modern philosophy," is significant in terms of his reacting against external authority in matters of belief, seeking a fresh basis for certainty, and finding it in the existence of his own mind. He must think in order to doubt his existence, hence his famous statement, *Cogito ergo sum ("I* think, therefore, I am"). Henceforward, much significance was given to the individual mind, and the resulting myth of the body–mind separation enabled both physics and biology to develop without the risk of ecclesiastical interference. Only in recent years has the inadequacy of the Cartesian body–mind myth come under general criticism not only because of the metaphysical problems it poses but also because it fails to do justice to the unity of personality that recent developments in medicine, such as those pertaining to psychosomatic disorders, presuppose.

The myth of the body–mind separation

Many of Descartes's 17th- and 18th-century successors can be best understood by reference to him. Nicolas Malebranche, a French Cartesian philosopher, and the occasionalist philosophers, were more radical than Descartes; they dispensed with any unity whatever in man himself and linked together man's mind and body by means of the constant correlation effected by God himself, claiming that mental events were merely "occasions" for God effecting material change. For Spinoza, the whole universe had not only Descartes's two attributes of mentality and materiality but an infinite number of attributes, and it could be alternatively named God or Nature. Each existent in the world could be pictured as a particular whirlpool in an infinitely deep sea made up of endless layers of particular fluids of which man knows only two—mentality and materiality. Gottfried Leibniz viewed Descartes's minds as the only ultimate existents, so that even material things were colonies of souls. God was viewed as the supreme monad (the ultimate substance) that establishes coherence and harmony among all other monads. What appears to men as the external world is, so to speak, the result of blurred vision on the part of those groups of monads that are human beings.

After Descartes there appeared the British Empiricists:

John Locke, George Berkeley, Joseph Butler, and David Hume. Locke, though rejecting some of Descartes's characteristic doctrines, nevertheless took over Descartes's view of the human mind and then concerned himself with the philosophical psychology of how the mind comes to have the ideas it possesses. By the time of David Hume (died 1776), the mind was viewed as nothing more than a collection or bundle of ideas thought of as very similar to images, which means, as Hume frankly admitted, that it becomes impossible to do justice to the subjectivity that makes each person distinctively the person he is. The significance of Berkeley (died 1753) in this sequence is that he saw the need for an extended Empiricism that took the notion of personality seriously and that regarded activity as a key concept. Indeed, for Berkeley the fundamental unit for thought was "activity-directed-towards-and-terminating-in ideas," and it was the activity of God directed to those ideas, which make up the external world, that gave to this world its continuous independent existence. His contemporary Butler also argued for a broader Empiricism, which for him centred on the significance of man as a moral agent and on a reasonableness that need not always conform to a mathematical paradigm. In a matter of great consequence, a man's action can be reasonable even though there may be little supporting evidence for his decision and though, indeed, the evidence may be very much against it. It may, thus, often be a moral duty to act in such problematical circumstances. This led to Butler's famous doctrine of probability—"probability is the very guide of life"—a view that influenced the treatment of belief in ***The Grammar of Assent*** (1870), by the English theologian John Henry Newman.

The Kantian and Hegelian idea of unity

Irnmanuel Kant has been called the second founder of modern philosophy. With Kant, late 18th-century philosophy began to take an interest in human knowledge, its varieties, scope, and limits. In Kant's critical philosophy, which emerged in his old age, he showed how scientific knowledge left room for morality. Though he was inclined to interpret all religious assertions in terms of morality, belief in God was justified as the holding of a regulative idea that brings coherence into all of man's thinking. The foundation of this idea is to be found, in fact, in those experiences of unity to which moral ideals, beauty, and the notion of a purposive universe all point. This idea of unity, largely implicit in Kant, was developed by Hegel, who came to regard the universe and its cultural, social, and political progress as but manifestations in time of an unchanging absolute spirit. In this way, Hegelianism provided a spiritual interpretation of the universe, but it regarded particular religions as no more than visual aids toward understanding Hegelian truths. A century later, the British philosopher F.H. Bradley was able to use a Hegelian approach in a much more empirical and far less intellectual context. Whatever form Hegelianism took and though its spiritual insights seemed on first view to make it a friend to religion, it has proved to be a position in opposition to Christianity, whether by its minimizing the historical element or by the way in which it compromises belief in a personal God.

Since the absolute Idealists, there has perhaps been only one philosopher in the mainstream of tradition—Alfred North Whitehead—who, in taking becoming rather than being as the fundamental category, made "process philosophy" possible. This philosophical view maintains a metaphysics that not only provides an interpretative scheme linking God, man, and the world but one that incorporates scientific and historical thinking, though in taking growth and process as fundamental, Whitehead seems, to some, to have an evolutionary God.

There were two main reactions against Hegelianism. The first, initiated by Kierkegaard, viewed Hegelianism as altogether too detached and objective and its ways of reasoning entirely unsuited to the deepest experiences of human life, the tragic situations in which human beings find themselves. From Kierkegaard, the Existentialist movement began. Also, in reaction against Hegel, were the modern Empiricists, such as Bertrand Russell and G.E. Moore from England, whose watchword was clarification in their attempts to create a straightforward, unambiguous language. This movement passed easily into Logical Positivism (a philosophical position that accepts only scientific knowledge as factual and rejects metaphysics), which challenged not only the truth but the meaning of theological assertions. (See also JEWISH PHILOSOPHY; PHILOSOPHY, HISTORY OF WESTERN; CHRISTIAN PHILOSOPHY.)

The concept of Nirvana

Developments in the East. ***Buddhist concepts.*** Among the religious philosophies of the East, the conservative Theravāda (Way of the Elders and another term for Hinayiina) Buddhism regarded all existence as a succession of transitory states: what alone was permanent was Nirvana, a deathless realm the existence of which was revealed to the Buddha himself in the Enlightenment that came to him while he meditated beneath the bo tree (late 6th century BC). About Nirvāṇa, the wise will say little more except to affirm its existence and to express their conviction that the plurality of individual souls that man knows in this world cannot in the same way exist in that deathless realm where there is no rebirth. Such ideas find a natural home in the philosophical standpoint of absolute Idealism, and Nirvāṇa can be regarded as an alternative word for the Absolute. Broadly speaking, Buddhism is agnostic both about a personal creator and personal immortality, though Theraviida Buddhism explicitly rejects belief in a creator. Undoubtedly, the dominant theme of Buddhism is the quest for release from the changes and chances of this world, which will lead to the serenity and peace of Nirvāṇa. A Buddhist saint is someone who has indeed become the Absolute, which thus incorporates and transcends all human imperfections and struggles and all the imperfect ideas, ideals, and deities of popular religion and popular ways of thinking. The difference between the ***arhat*** of Theravāda, and the ***bodhisattva*** of the Mahāyāna is one between two different routes of realizing Nirvāṇa—the one through self-concentration; the other through self-sacrifice for the welfare of others. The difference is one between two "saintly" routes to the one saintliness—being possessed by and dwelling in the Absolute.

Thus, Buddhism, by embracing what is, in effect, a metaphysical concept of the absolute, not only could but did hold together a complex mythology within a unifying philosophical insight and was able, as in Japan and China, to incorporate a complex popular pantheon of the cult of ancestors. Furthermore, it could combine a popular devotion to a personal lord with a mystical contemplation that had encouraged the development of Buddhist monasteries. In sponsoring such a broad synthetic (all-embracing) view, the philosophical significance of Mahāyāna (Greater Vehicle) Buddhism emerged. Such developments began about 100 BC and lasted for several centuries; it was Mahāyāna Buddhism that spread to China and East Asia to influence and modify the religions native to those areas. In Mahāyāna, the humanitarian saviour notion of the ***bodhisattva*** has some echoes in Kenotic Christianity (*i.e.*, emptying oneself to become a suffering servant), and attitudes to the Buddhist scriptures have parallels with those of Christians toward the Bible. Common to both is the view that revelation can express itself in developing forms and that it is a mistake to concentrate on the texts themselves, sacred though they are, rather than on that which transcends them and of which they are symbols and to which they point. In this respect, one may contrast the open and exploratory attitudes of many Buddhists and Christians toward their sacred books with the closed and rigid attitudes of most Muslims.

Confucian, Taoist, and Japanese concepts. Prior to the introduction of Buddhism into China in the 1st century AD, the two main strands of religious thought in that country were represented by Confucianism and Taoism. Confucianism displays a reverential propriety that is expressed and developed in social relationships and fulfilled in Heaven. Taoism claims that the wise man will constantly seek harmony and rapport with Tao (the Way), which, at one and the same time, is the way for men to follow if they would reach blessedness and the principle that underlies and sustains the world. As a concept that is both moral and cosmological, Tao has a logical status similar to that of the Logos (or Word, the active principle of God in creation and revelation) in Christian philosophy. The Taoist thinks little of the ways of the world, including the decorum of the Confucianist; his outlook rather encourages a laissez-faire policy toward the world and even withdrawal from its affairs. 'The immediate mystical experience of Taoism or the inspired behaviour of Confucianism can easily blend with Buddhism, which sets both within a metaphysics of the Absolute.

In Japanese religion are found the same two themes that are found in most religions, though in their extreme forms they are mutually exclusive. On the one hand is mysticism—specifically, nature mysticism, of which the mysticism of the Zen practitioner belonging to an intuitive meditative form of Buddhism is a specific example. In Zen Buddhism, religion is scarcely distinguished from an aesthetic experience in which shrines, gardens, mountains, woods, and streams reveal a mysterious beauty and in which the exercise of the intellect is at a minimum. In contrast to mysticism, there is devotion to a supreme personal lord, at one time symbolized in the emperor as a descendant of Amaterasu, the Sun-Goddess. At other times Shintb ("Way of the gods") devotion focussed on particular shrines and particular deities, just as Zen Buddhism could concentrate on a particular image or on particular events. In both types of devotions, however, it could be argued that such particularity was fulfilled and transcended in the unity revealed to a mature mystical insight. These different philosophical positions have an interesting reflection in the Christian position in which the Christian claims to find evidence of God's presence and activity in particular places and situations (especially in the incarnation of Christ), though at the same time allowing for God to be omnipresent.

Doctrine of an eternal soul and rebirth

Hindu concepts. This mixture of a mystical contemplation, which sees the divine everywhere, and a personal devotion to a particularized divinity recur in Hinduism. The most characteristic feature of Hinduism, however, is the doctrine of an eternal soul and its rebirth. The universe is pictured as the arena in which the immortal soul engages in a succession of incarnations from which man seeks release, a release that true contemplation can give him, especially when approached through Yoga (a mental, physical, and spiritual meditation technique). At the same time, a sensitivity to the numinous (spiritual) has left open the possibility of and certainly encouraged personal devotion. The most famous of Indian scriptures, the ***Bhagavadgitd*** ("Song of God") has for its recurrent theme the majesty, glory, and terror of God and the devotion due to him, though as in Christianity these attributes are compatible with a loving God. In the matter of revelation and incarnation, it is an open question as to how far the Hindu conception of revelation is the same or similar to Christian or Muslim conceptions. The Hindu view of ***avatiira*** ("incarnation"), however, implies many incarnations and in a Christian context would demand many Christs; thus, the concept of ***avatiira,*** a salient feature of Vaiṣṇavism (centring on the veneration of Viṣṇu, the preserver), cannot be easily reconciled with the uniqueness attributed to Jesus.

Depending on the particular questions that determine a particular content of discussion, Hinduism can talk of a plurality of souls, when it would concentrate on the theme of reincarnation, but, especially when influenced by Buddhist (and also pre-Buddhist) ideas, it can also sponsor an absolutism, or a monism; yet, again, it can come very close to a traditional Western theism. On the whole, however, it might be said that Hinduism holds together in a creative tension both theism and monism, though often it appears that in conceptual foundations and philosophical discussion the theistic strand predominates. Even in its classical period (600 BC to 450 BC) Hinduism was characterized by an astonishing variety of doctrines and cultures. Indeed, it well illustrates a characteristic of Indian thought that is becoming more acceptable to Western ways of thinking—the notion that there are many different approaches to the truth, which matches the concept of a multiple theology. It was regarded, however, as a retrograde step when these varieties of culture, ritual, and mythology became hardened into social strata and castes.

Concepts of *ātman* and Brahman

In the medieval period, Sankara (*c.* 788-4320), the leading exponent of Advaita Vedanta, or nondualism, is the most significant Hindu figure in the philosophy of religion. Arguing in a way very reminiscent of absolute Ide-. alism, he claimed that the only existent was an absolute and that all else was an illusion. In this context he equated *ātman* (the individual soul) with Brahman (the universal or absolute soul). Both were viewed as one in a cosmic consciousness. For Sankara, only ignorance or lack of insight into the nature of being prevents a man from realizing his identity with Brahman and thus becoming here and now aware of the freedom that is his. Sankara also allows as permissible, without being accepted as the truth, talk of God as personal and as creator and of men as separate souls related to one another and to him. This, however, is only considered a way of talking—salvation in the Absolute transcends all such imperfect discourse. The same logical problems recur here in the concept of Nirvana in Buddhism. In Hinduism, the ***Upaniṣads,*** Hindu philosophical treatises, and the ***Bhagavadgitd*** use the imperfect language of finite man, who has not yet found release, and in this way they can only point beyond themselves to that which they cannot adequately express. Here again are ideas reminiscent of some of those in Western philosophy of religion in the modern world: the importance of theological reticence, the limitations of theological language, and, in another context, the significance of "existential situations."

Twentieth-century Hinduism has been chiefly characterized by attempts to purify and reform the doctrines of its medieval period, to deepen its spirituality, to reassert its moral dimension, and to inspire social reform. Mahatma Gandhi and Sri Aurobindo, the founder of a spiritual community and a Communist, were significant in such ventures. Aurobindo has been compared with the French Jesuit paleontologist and theologian Teilhard de Chardin insofar as both have a repeated experience of cosmic consciousness and a profound belief in evolution, both of which point to a divinization of man.

The strength and limitations of a total surrender

Islāmic concepts. At the heart of Islam is an experience of awe before the one, all-powerful, mysterious creator Allah. Thus, its dominant theme has been surrender, though it must not be forgotten that it has nurtured mystics to whom the mysterious and awesome God has revealed himself through created things. Allāh controls man's destiny, whether to salvation or damnation, which points to the ultimacy of God, to his majesty and power. The concept of heaven inspired warriors to fight to the death; the concept of hell encouraged loyalty by showing what terrible punishments awaited the disloyal. The Qur'ān (the Islamic sacred scriptures) is regarded as an infallible book—a transcript of a tablet that is eternal in the heavens. Islām shows pre-eminently the strength and limitations of a total surrender based on clear-cut beliefs, themselves arising from a basis in infallible texts, the whole being translated into vigorous political and social practices associated with a rigorous ritual and ceremonial discipline. Its mixture of both rigour in theology and vigour in politics in India and the Middle East from the Middle Ages to the 20th century can perhaps be compared with the same mixture as has been seen in the Protestant and Catholic communities in Ireland since the 17th century. However much the concept and practice of holy warfare is repugnant to many minds today, in the context of Islām it implies a sensitivity to evil and a conviction that evil has to be resisted and overcome in a total dedication. In this way the faith of Islām has shaped human history by obedience to a resolute and powerful God. Islam also illustrates the point that predestination need not bring with it a submissive fatalism. Furthermore, it has to be granted that Islām has allowed, within itself, for some allegorical interpretations of the scriptures–explicitly by the Ṣūfīs (mystics)—and it has also allowed for differences of piety and beliefs and even intellectual exploration on the part of particular disciples. Nevertheless, to other religions Islām has shown itself to be very conservative and with a distrust of compromise and a passionate desire to proselytize.

East and West: common ground. In reviewing the different philosophical understandings of religions in both East and West, two points clearly emerge. First, that however great the variety, there is almost universal agreement that "what there is" is not restricted to the facts and features of the world as they are given to or received by man's senses. Secondly, what has been for the philosophy of religion in the East almost a permanent problem is coming to be a crucial problem for the West, viz., how to preserve both the concept of absolute spirit and the significance of personal individuality or, alternatively, how far one can speak reliably of God as a person. The West is becoming aware of the problematic character of religious discourse. If, in such ways, Western philosophy of religion can benefit from some of the insights of the East, so also can the East—as a growing interest in the Empiricists of the West demonstrates—gain from the West. Not least, scientific developments have created Eastern interest in the English Empiricists, particularly John Locke; Eastern philosophers also have been impressed by the political liberalism of some modern Western Empiricists, such as Bertrand Russell. The Empirical philosophy of religion, as it has been recently developed in the West, may provide basic approaches and techniques for a closer mutual study of religions in East and West. (See also INDIAN PHILOSOPHY; CHINESE PHILOSOPHY; JAPANESE PHILOSOPHY; ISLAMIC THEOLOGY AND PHILOSOPHY.)

BASIC THEMES AND PROBLEMS IN THE PHILOSOPHY OF RELIGION

The kinds of proofs of God's existence

The problem of God, the Absolute, or the supreme value. The existence of God. The so-called proofs of God's existence are of two kinds: independent logical exercises or particular conclusions set within an overall metaphysics. Either way, the discourse of these independent proofs or metaphysical schemes is best viewed as speech designed to evoke a disclosure. A particular argument recommends, as a way of speaking about what the disclosure discloses, a particular brand of discourse offering an interpretation of the world and man and one that develops from a specific key idea grounded in the disclosure. The existence of an Absolute or a supreme value has never been concluded as a result of an isolated logical exercise but has always arisen in the context of a total metaphysics. Thus, a quasi-mathematical structure, for Spinoza; a dialectic method, for Hegel; and evolutionary considerations, for the modern French philosopher Henri Bergson, determined the discourse that these three philosophers used in order to evoke that situation to which God or Nature, the Absolute Spirit, or the life force became for them respectively key concepts of interpretation. Bradley similarly reached a belief in an Absolute Spirit by reflecting on the logical problems of relatedness.

The following are some traditional arguments for the existence of God restyled along the lines suggested above:

The ontological argument of Anselm of Canterbury (*c.* 1033–1109) takes a phrase "that than which nothing greater can be conceived" and uses it as a technique for disclosure, directing one without limit to an ever-increasing perspective, in the hope that at some point the light will dawn, whereupon the phrase "necessary being" will be used to develop talk of the God.

The cosmological argument uses as a technique for disclosure such questions as "Why is this thus?" or "Why is there anything at all?" In receiving replies to these questions in causal terms, the cosmological argument builds up an ever-increasing causal spread until a disclosure occurs, whereupon the phrase "first cause" specifies what is disclosed and advocates certain ways of talking.

The argument from design takes a story with acknowledged disclosure possibilities—*e.g.*, the inter-related parts of a watch—and uses this as a catalyst to evoke a disclosure around some ever-broadening purpose patterns of the universe, in relation to which one can speak of God in terms, for example, of eternal purpose.

What is, in different ways, implied by these arguments is that the word God is unique in its logic, that it works in discourse as no other word exactly works. Thus, one cannot say "God exists" but rather "God necessarily exists." This is sometimes expressed by remarking that the existence of God is not the existence of a physical object or even the existence of a person, though what can be said about persons is less misleading in speaking about God than in speaking about the logic of things. This point is sometimes made, albeit misleadingly by saying that God does not exist, but this is only a picturesque way of saying that he does not exist in the way that a table exists.

Qualifications for personal models

The *nature* of God. These reflections are of wider applicability in relation to the nature and attributes of God. Such attributes are spoken of in terms of personal models, such as wisdom, goodness, power, love, mercy, righteousness, and so on. These models, however, will always need qualification by words such as infinite, perfect, and all. What is quite clear is that grammar itself is no clue to the logic of phrases such as "infinitely wise." Although that phrase is similar in grammar to one such as "exceedingly wise"—a phrase that is entirely descriptive in its logic—it is logically quite different, because "infinitely wise" has both descriptive and what has been called performative force. In other words, it not only describes some matters of fact—some specimens of wisdom—because of the word wise, which works descriptively as a model, but it also generates something—the word infinite acting as an operator, continually directing persons to expand their understanding until a moment of vision emerges. Alternatively, the point that God is not a being has sometimes been made by saying that God is the ground of Being—"the ground of" functioning as a qualifier, operating as the model of beings, or things. The emphasis of such qualifiers is twofold. First, they remind one of the inadequacy of all language used to speak of God—language authorized by particular models that, arising in a moment of vision or disclosure, naturally originate speech about what the disclosure discloses. Secondly, qualifiers constantly point one back through developed discourse to that moment of vision in which the discourse originated and in which alone one knows what the discourse is speaking about. The logic of models and qualifiers is a way of combining the intelligibility and mystery that any philosophy of religion must preserve.

Language about God thus develops as a multiple discourse, having various strands of which each is authorized by a particular model and of which each must, somewhere along the line, be modified by the presence of the others. Thus, theological understanding is a complex interweaving of different strands, and not least is the task of the philosopher of religion to produce the most comprehensive, coherent, consistent, and simple discourse he can. When problems arise that seem to be problems about the nature of God—for example, the conflict between different attributes--these are most profitably translated into problems of language. They then become problems of how to create discourse of the kind that in the end produces the best understanding of a cosmic disclosure with a single individuation, in which all the pertinent discourse originates and about which all the different strands endeavour to speak.

Natural and revealed theology

The knowledge of God. Natural theology is the name given to the kind of discourse about God and the world that originates in natural moments of vision without reference to God's revelation of himself in an incarnation, and in this sense "natural theology" is distinguished from "revealed theology." Among some philosophers—*e.g.*, Locke—the distinction is one between general and special revelation. In natural theology are generally included the "proofs" of the existence of God, discussions about the immortality of the soul, and discussions about God's providential control of the world, which provides for man a state of moral probation.

Some have viewed religious experience as affording direct evidence for the existence of God. In any discussion of religious experience, however, it is important at the outset to distinguish religious experience in general—a sense of awe or reverence, or a sense of the numinous—

from mystical experience. The language of mystics is notoriously confusing to those not accustomed to the mystical idiom, and a leading question is how far mystical experience can establish the kind of objective reference it claims. Words such as immediate, direct, and intuitive refer rather to the way in which the experience occurs as a disclosure rather than justifying one in taking as guaranteed the interpretation that this disclosure appears to bring with it. If one already has an interpretative scheme, then mystical experience may provide an instance of such a scheme, but this has been rightly described as supporting belief in God "on the way back" rather than "on the way out." The concept of revelation is used by Christians to describe the way in which God's activity is uniquely disclosed in Christ, and faith relates to the human attitude and response that matches revelation subjectively. Revelation is sometimes contrasted with discovery, the former being said to relate to a passive subject, the latter to an active subject, but the distinction is largely one of emphasis. Philosophers of religion are now inclined to view revelation in terms of activity that waits to be interpreted rather than as a revelation of propositions. Revelation thus relates to events rather than to doctrine. According to this view, doctrine could never have the ultimacy and finality that necessarily belongs to the givenness of God in his incarnation or incarnations.

Special problems. Freedom. Among the classical problems in the philosophy of religion are those of free will, self-identity, immortality, evil, and suffering. The freedom of the will is a claim for the uniqueness of the subject, known in occasions of activity in which the subject "comes alive" and realizes his subjectivity as that which cannot be reduced to the behaviour patterns and facts—*i.e.*, the objects—of the natural and social sciences. Such freedom is realized in responding to a situation that has equally come alive objectively to inspire a person and call forth such response. Some claim the predictable character of human behaviour rules out man's freedom; others state that the extent to which human behaviour is unpredictable argues for freedom. This controversy, however, does not in any way solve the problem of freedom; it only makes evident what kind of problem the problem of freedom is, viz., how far human nature is capable of being analyzed into behavioral terms without any residue.

Arguments for personal immortality

Self-identity *and* immortality. When there has been a self-disclosure of transcendence, of what cannot be characterized in space and time, one cannot say that any self so disclosed entirely comes to an end. In this sense, there is an argument for personal immortality, though one can only talk sensibly about it by expressing immortality in terms of continuing personal life. In Christianity this becomes speech about the resurrection of the body, and in Hinduism it becomes speech about reincarnation in this world or in the universe at large. All detailed talk about a future life, whether in Christianity or other religions, is only a way of spelling out and pointing back to that experience of man's transcendence here and now, in terms of language that expresses the claim that such a transcendent element is not annihilated by death. To be articulate about immortality, emphasis is placed on features of life that, at first view, have high significance and point here and now to experiences in which man's self-disclosure is most often found—*e.g.*, inspiring music or the intimate and deep fellowship of a particularly significant meal. General claims for immortality in relation to an objective disclosure (whether it be spoken of in terms of God or moral ideals) have to be distinguished from, though they have evident similarities to, the Christian claim for eternal life. Eternal life is, according to the Christian view, a subjective self-disclosure alongside the objective disclosure of God's activity in Jesus Christ, and it is as unique as the uniqueness of God in Christ, a uniqueness that is, however, an inclusive, and not exclusive, uniqueness.

Evil and suffering. The problem of evil arises (1) from the loss of a sense of God's presence in the face of evil or suffering and (2) from an apparent conflict between the language used to describe God (*e.g.*, all powerful, all good, and all wise) and that to describe the world as being characterized by evil and suffering. The solution proffered by the Book of Job in the Old Testament is that of evoking such a sense of awe around the created universe that, discovering in this way a renewed sense of God's presence, one accepts both evil and good and contents himself verbally by acknowiedging a final incomprehensibility.

Other solutions relate good and evil to God and thus seek consistency by relating good and evil to God's primary and secondary will or to God's willing and permitting, respectively. In demanding some overall purpose to complete such a story, however, these solutions point to others that seek to resolve the conflict between good and evil within some reconciling model, which is then used to specify, with suitable qualification, a purpose or attribute of God. Thus, the conflict necessarily involved in the creation of a community of freely responsible persons is used as a model to illuminate a personal conflict exhibited, for example, by war. Also, the conflicts resulting from general rules imposed for the sake of training are used to provide a model to illuminate the disharmony exhibited in, for example, earthquakes or floods. These models are then developed and amplified in order to lead one to a renewed disclosure of God's presence. These solutions—by raising questions about God's character—perhaps point to another solution that attributes to God redeeming love—something that, as directed to evil, can be creative of personal maturity and fulfillment in a way not otherwise attainable. This attribute must then be appropriately qualified so as to lead to a renewed disclosure of God's presence, in this way enabling one both to face evil and to talk of it more coherently in relation to God.

In the matter of absolute Idealism, which is the kind of metaphysics implied in Eastern religions generally, evil and good are transcended in the Absolute Spirit that is beyond good and evil. Logically, this is akin to the solution of the Book of Job.

THE PRESENT SITUATION IN THE PHILOSOPHY OF RELIGION

Interest in religious language

In the latter part of the 20th century in western Europe and the United States there has been an Empirical philosophy of religion the interest of which has been in religious language and the kind of Empirical basis there can he for religious discourse. The definitive question has been concerned with what are the patterns of religious reasoning and what is the character of religious language if such discourse points back to and articulates situations of the particular kind that have been discussed above. The approach originated in what has been called Logical Positivism. According to the verification principle, which gave what the Positivists considered to be the touchstone of meaning, an zssertion had meaning if and only if it was verifiable at least in principle by sense experience. Logical Positivists were not at all daunted when this seemed to exclude the whole of theology and a good deal of ethics from meaningful discourse.

Since the 1950s, however, there has been a reaction against the Positivist's veto, and the works of the Austrian British philosopher Ludwig Wittgenstein are symptomatic of those who broadened Empiricism so that it has become interested in displaying and elucidating the variegation of language, in setting language in actual contexts, and in relating it to specific situations. Significantly, this mellowing of Empiricism has been accompanied by a growing interest in personality and the self. This newer emphasis of Empiricism unites with Existentialism in suggesting that personal situations may very well provide helpful parallels to religious situations. There has been introduced into the philosophy of religion a renewed sense of the significance of mystery and a new emphasis on theological reticence. With this has come a renewed awareness of the significance of metaphor, myth, and symbol, and there has also emerged a significant use of the concept of the model. The use of terms like myth and mythological, it is important to recognize, does not mean that the assertions so called are false. Myth includes stories that try to articulate what is objectively given in a certain religious situation. Myths also relate to historical events—though the myth may be selective in its choice of such events—when speech about these events is used to articulate a claim of a transcendent kind. In other words, myth, metaphor, symbol, and model are all ways of expressing in ordinary language an extraordinary point.

Questions of reference and criteria

The present stress on metaphors and models in religious language, however, inevitably raises two far-reaching questions: the question of reference and the question of criteria. The former concerns the possibility of the assurance that one is talking about anything at all. The latter concerns what the criteria are for better and worse ways of talking. The question of criteria has been answered in terms of the logical character and the empirical pattern of the multimodel discourse to which the different strands arising from the different metaphors or models give rise. That this discourse talks about something must in the end rest on the claim that, in a disclosure situation, a subject is relatively passive—*i.e.*, aware of an activity bearing on his own and thus aware of something other than himself about which he is talking.

In this context the significance of the Existentialist approach is to underline, as does recent Empiricism, the importance of a wider view of human experience than ordinary scientific experience might allow and to point one to highly significant personal situations that cannot be netted in scientific terms. The phenomenological approach, as developed by the Moravian philosopher Edmund Husserl, represents an attempt to be objective and scientific about experience, an endeavour to set out facts uncompromised and unprejudiced by metaphysical frameworks. As an endeavour to reach agreement on what is being talked about and as an attempt to seek the simplest and clearest interpretations, the phenomenological approach has been applauded by many philosophers of religion and theologians. There can be no question, however, about a purely scientific account of a religious situation-—that would be a contradiction in terms, and, though there can be a phenomenological approach to religious situations, there can be no phenomenological explanation of them that claims to be adequate. The main contribution of Phenomenology is that of encouraging scholars to describe situations with as much critical analysis as possible.

Logical Empiricism, it might be said, has absorbed something of the Phenomenologists' concern. It has certainly raised questions about and directed interest toward the way situations are talked about and interpreted, the possibility of there being different interpretations of the same situation, and so on. In this way it has provided the tools for and greatly stimulated contemporary interest in hermeneutics (critical interpretations)—a second order appraisal of interpretations together with an interest in their empirical bases.

The future of the philosophy of religion

As to the future of the philosophy of religion, a merging of the Empirical and Existential strands may well be expected. Metaphysical and religious views of the future most likely will combine conviction with tolerance and commitment with openness. The commitment and the conviction will probably come from moments of vision. Claims to finality, fanaticism, and bigotry will disappear, it is hoped, when it is realized that there are no self-guaranteed translations of what is disclosure-given and tolerance and openness will arise from the acknowledgment that all understanding of these moments of vision is a multiple exploration, an exploration yielding different strands of discourse. Solutions to contemporary problems, social and intellectual, demand a multiple consideration by scholars from many disciplines of all the issues involved in the problem, a consideration set within a framework of faith and morality in which man is interpreted as distinctively human, characteristically a person. Prom such interprofessional, interdisciplinary groups may emerge a new metaphysics and a new theology linked with, but by no means prescriptive of, assertions in other subjects. In this way there may be created a new culture—scientific, moral, religious, and technological at the same time. To be involved in such groups would seem to he the main task of the philosopher of religion, as of the metaphysician, today. If he is successful and if these interdisciplinary groups are creative, the modern period will then take its place among those that have marked crucial turning points in the history of mankind and its culture.

BIBLIOGRAPHY. General introductions include H.D. LEWIS, ***Philosophy of Reiigion*** ("Teach Yourself Book") **(1965);** and J. HICK, ***Philosophy of Religion*** **(1963).** Introductory books that range over a somewhat narrower field include J.L. GOODALL, ***An Ztztroducrion to the Philosophy of Religion*** **(1966);** I.T. RAMSEY, ***Religious Language*** **(1957** and **1963);** and THOMAS FAWCETT, ***The Symbolic Language of Religion*** (**1971).** NINIAN SMART, ***Philosophers and Religious Truth, 2nd*** ed. (1969), centres discussion of some salient issues around particular philosophers. NINIAN SMART, ***The Religious Experience of Mankind*** **(1969);** and EDWARD GEOFFREY PARRINDER, ***Comparative Religion*** (1962), provide general introductions to the comparative study of religions. In *World* ***Religions*** (1966), H.D. LEWIS and R.L. SLATER consider issues in the world religions that are highlighted by contemporary approaches in the philosophy of religion and the comparative study of religions respectively. In the psychology of religion, the classic work of WILLIAM jAmes, ***The Varieties of Religious Experience*** **(1902,** reprinted 1952), is probably still the best introduction that might then be followed by a comprehensive survey of the contemporary field, such as L.W. GRENSTED, ***The Psychology of Religion*** **(1952).** Most of these books contain excellent bibliographies for further reading. NINIAN sMART, ***Historical Selections in the Philosophy of Religion*** **(1962);** and I.T. RAMSEY, ***Words About God*** (1971), are useful sourcebooks for some classic discussions of topics in the philosophy of religion; an excellent survey of recent thought is given in JOHN MACQUARRIE, ***Twentieth-Century Religious Thought: The Frontiers of Philosophy and Theology, 1900–1960*** **(1963).** F.R. TENNANT, ***Philosophical Theology, 2*** **vol. (1928–30);** H.H. FARMER, ***The World and God: A Study of Prayer, Providence and Miracle in Christian Experience*** **(1935);** and H.D. LEWIS, ***Our Experience of God*** (1959), represent different general treatments of the subject.

Books dealing with specific problems of religious belief are: PETER R. BAELZ, ***Prayer and Providence*** **(1968);** J. HICK, ***Evil and the God of Love*** **(1966);** A. FARRER, ***Love Almighty and Ills Unlimited*** **(1966);** I.T. RAMSEY, ***The Problem of Evil*** **(1972)**; and W.T. STACE, ***Mysticism and Philosophy*** **(1960).**

Books concerned with the challenge of recent Empiricism to religious beliefs are: R. ROBINSON, ***An Atheist's Values*** **(1964)**; and R.W. HEPBURN, ***Christianity and Paradox*** **(1958);** FREDERICK FERRE surveys the challenge in ***Language, Logic, and God*** **(1961).** D.D. EVANS, ***The Logic of Self-Involvement*** (1963), develops ***J.L.*** Austin's doctrine of performatives in a religious direction. GEDDES MACGREGOR, ***God Beyond Doubt*** **(1966)**; H.P. OWEN, ***The Christian Knowledge of God*** **(1969)**; and JOHN MACQUARRIE, ***God-Talk: An Examination of the Language and Logic of Theology*** (1967), represent different reactions to the empirical challenge, as do the books by **I.T.** Ramsey above. NINIAN SMART, ***Reasons and Faiths*** **(1958);** and H.H. FARMER, ***Revelation and Religion: Studies in the Theological Interpretation of Religious Types*** (1954), are examples of a predominantly philosophical approach to the study of differentreligions.

Books concerned with particular issues in different religions are: R.C. ZAEHNER, ***Evolution in Religion*** (1971), ***Mysticism, Sacred and Profane*** (1961), and ***At Sundry Times: An Essay in the Comparison of Religions*** **(1958);** and EDWARD GEOFFREY PARRINDER, ***Avatar and Incarnation*** **(1970).**

In the psychology of religion, P.E. JOHNSON, ***The Psychology of Religion*** **(1945);** and G.W. ALLPORT, ***The Individual and His Religion*** (1950), give comprehensive treatments of the subject and discuss a wide range of material from different schools of psychology. Recent books of importance in the psychology of religion dealing with specific topics are: VICTOR WHITE, ***God and the Unconscious*** **(1952);** DAVID COX, ***Jung and Sr. Paul*** **(1959);** R. HOSTIE, ***Analytische Psychologie en Godsdienst*** **(1954;** Eng. trans., ***Religion and the Psychology of Jung,*** **1957);** and R.S. LEE, ***Freud and Christianity*** **(1949).** W.G. MACLAGAN, ***The Theological Frontier of Ethics*** (1961), raises difficulties, from the standpoint of moral philosophy, about certain stylings in Christian theology of the doctrines of grace and freedom. Further discussions of topics on the frontier of theology, ethics, and philosophy are to be found in ILLTYD TRBTHOWAN, ***Absolute Value: A Study in Christian Theism*** **(1970);** and KEITH WARD, ***Ethics and Christianity (1970).***

(I.T.R.)

Religion, Social Aspects of

Religion can be viewed from many perspectives: by philosophers, historians, aestheticians, and theologians. The subject of this article is the study of religion from the perspective of social scientists. Social scientists ask how religion is related to the structure and processes of human societies and how it both reflects and affects stratification systems in society, political and economic processes, levels of integration and of conflict, and the course of social change. In viewing the religious patterns themselves—in which the varieties of religious organizations and the types of leaders will be examined—this article will not be simply concerned with description but also with identifying the conditions under which these varieties appear. Varieties of individual religious experience will also be considered.

NATURE AND SIGNIFICANCE

To examine religion from the perspective of the social sciences is to require a rather drastic shift in the way religion is usually viewed. The sharply contrasting beliefs about the road to salvation, for example, are neither evaluated nor simply described. The social scientist is rather concerned with the conditions under which the various beliefs appear, the variations of beliefs among societies, groups, and individuals, and the consequences of the various conceptions for social interaction.

Concerns of social scientists

The social aspects of religion are thus studied in the conviction that major dimensions of religious belief and practice are only partly understood if they are not seen as part of a larger social order. The reverse of this is also true: many of the most critical problems in the social sciences involve a religious factor. A purely rational economic model of man, for example, whether it involves the market as a system or individual economic behaviour, must take account of the "distortions" produced by religious motivations and values. Prediction and explanation of political behaviour are improved if, in addition to knowledge of class, education, occupation, traditional political identity, and the like, there is a knowledge of religious group membership, belief, and behaviour.

Economics and political science, however, have not made extensive use of religion as a factor in their studies. Theory and research dealing with the social aspects of religion have been developed more fully by anthropology, sociology, and psychology. In the first two particularly, most of the major theorists have given a prominent place to the study of religion, and for some of them—despite wide variation in their perspectives—it has been the central concern.

Alongside such disciplinary studies of religion—which, however, are seldom narrowly constrained to one discipline—there has developed a more general science of religion. To some degree, this represents the pressure toward synthesis of the analytic studies. Few of the major problems are specifically social or psychological or even cultural. To understand them, the theorist must take each of these areas and their interactions into account, however much he may think of himself as a specialist. To see the Black Muslim movement in the latter half of 20th-century America, for example, simply as an expression of the needs of certain individuals or as a manifestation of a cultural system or as the product of Negro-white interactions is to receive only a partial picture. There is no universally accepted term for this anthropology psychology sociology of religion, though the terms science of religion or sociology of religion—since sociologists have perhaps been most active in developing a general theory of religion—have been most frequently used.

Theology, critical studies of sacred texts, and philosophy are also—somewhat indirectly—major stimulants to the emerging science of religion. They too are pressed toward a general theory of religious structures and behaviour as one of their concerns. Some of their tendencies toward a science of religion have been caught up, in recent years, in the study of comparative religion and the history of religion, although these disciplines are seldom limited to, or explicitly concerned with, the science of religion. From the Anglo-German philologist Max Miiller's "Comparative Mythology," written more than a century ago, to such recent works as *The Comparative Study of Religions*, by Joachim Wach, a German-American historian of religion, can be seen the interweaving of scientific, moral, and theological interests. This tends to be true also of German *Religionswissenschaft* (science of religion). Despite the multiplicity of interests, many comparative and historical works contribute valuable data and interpretations to the science of religion.

Tendencies toward a science of religion

MODERN THEORIES OF RELIGION

Systematic thought about religious institutions and behaviour goes back many thousand years, but this article will refer only to some of the major modem theories of religion and society.

Evolutionary theories. Reflecting in part the intellectual impact of evolutionary theory, late-19th- and early-20th-century students of the social aspects of religion were strongly interested in the origin of religion. An empirically based theory of religious origins is impossible since only remnants of archaeological evidence are preserved to give hints of what the earliest forms might have been like. Theories of origin, nevertheless, have proved to be valuable starting points for functional and other theories of religion and thus deserve attention. Some authors have emphasized what they saw as the individual elements in religious origins. In the 19th century the British anthropologist Edward B. Tylor, in his book *Primitive Culture,* interpreted religion in its earliest forms as an effort to deal with the many puzzling phenomena that the "savage mind" had insufficient knowledge to explain. Dreams, echoes, visions, and, above all, death were accounted for by the idea of a soul that could leave the body. This animistic (soul) view, which Tylor considered the basis of religion, represented what he called a "fairly consistent and rational primitive philosophy." On it, he claimed, were built all later and more complex forms of religion, first polytheism and then monotheism.

Tylor's theory was sharply criticized for its rather formal evolutionism (according to which religion disappears as scientific explanations improve), for its excessive individualism, and for its rationalism. The evolutionary perspective was developed more fully by the British folklorist Sir James G. Frazer in his great compendium of facts about religious and magical practices, *The Golden Bough.* He viewed human thought as developing from magic, to religion, to science. The oversimplification and ethnocentrism (upholding the superiority of one's race) of such social evolutionism, and scarcely less of its critics, nearly halted the exploration of systematic processes of change for half a century. In the latter half of the 20th century, evolutionism is reappearing, although in a more cautious and empirical, as well as a less ideological, form. New efforts are being made to discover whether social change, when studied comparatively, proceeds in a random, a cyclical, an evolutionary, or some other pattern. The American sociologist Robert Bellah, in investigating whether or not religion evolves along with science in a parallel fashion, has suggested five "ideal typical" stages through which religion tends to move: primitive, archaic, historic, early modem, and modem, the last of which is characterized by the possibility of a continuously self-revising religious system congruent with science. Magic also probably evolves similarly, taking new forms in a literate, mobile, and scientific era, though not disappearing, since the needs that it expresses have not disappeared. Thus, societies faced with sudden and dramatic challenges to a moral order or a physical security that has been taken for granted may seek out the "witches" in their midst—whether through witch trials (as was done in Salem, Massachusetts, in the late 17th century) or through public hearings concerning threatening ideologies (as was done in the United States in the 1950s under the leadership of Senator Joseph McCarthy)—in an effort to help the societies to redefine what is bad and to discover where the boundaries of the social order are.

Magic and religion

Psychological and sociological theories. The rationalistic quality of Tylor's theory, and of 19th-century evolutionism generally, was drastically modified by psychological and sociological theories that followed. Although he referred to Tylor favourably at several points and did not explicitly criticize his theory, Sigmund Freud developed a sharply contrasting theory of the origins of religion. According to Freud, a primeval slaying of a tyrannical father by his sons––primarily because the father had monopolized the females of the horde—led to guilt and then repression. Totemism (a belief system involving animal symbolism), which Freud' regarded in his book *Totem and Taboo* as the earliest form of religion, was an effort to deal with this guilt and repression.

> The totemic system was a kind of agreement with the father in which the latter granted everything that the child's phantasy could expect from him, protection, care and forbearance, in return for which the pledge was given to honor his life, that is to say, not to repeat the act against the totem through which the real father had perished. (From Sigmund Freud, *Totem and Taboo; A.A.* Brill [tr.], 1927.)

More generally, and moving from an origins theory to a functional theory, Freud viewed religion as a culturally furnished system of projection, derived from family experiences, particularly as characterized by the helpless child and the powerful father. Society could not exist, according to Freud, without the repressions of culture—despite their psychic costs—and man's natural frustrations. In this context Freud wrote:

> The gods retain their three-fold task: they must exorcise the terrors of nature, they must reconcile one to the cruelty of fate, particularly as shown in death, and they must make amends for the sufferings and privations that the communal life of culture has imposed on man. (From *The Future of an Illusion.* By permission of Liveright Publishers, New York.)

Sociological theories of the origin of religion view various human interactions, rather than individual needs, as the starting point. The German sociologist Ceorg Simmel developed the thesis that religion is the heightening and abstracting of various human relations found widely in social life from their particular content. Faith, he suggested, was first of all a relationship between individuals, an essential element in group life.

> In faith in a deity the highest development of faith has become incorporate, so to speak; has been relieved of its connection with its social counterpart. (From Georg Simmel, "A Contribution to the Sociology of Religion," *American Journal of Sociology,* November 1905.)

This same idea has been expressed by the American theologian Reinhold Niebuhr. In regard to the question as to why religious faith persists three centuries after the first major triumphs of modern science, his answer was a functional interpretation that contrasts sharply with Freud's theory. He stated that basic trust in human relations is born from the security given to a child by his parents.

> But life is full of ills and hazards, of natural and historical evils, so that this childlike trust will soon be dissipated if maturity cannot devise a method of transmuting the basic trust of childhood, based on obvious security, to a faith which transcends all the incoherences, incongruities, and ills of life. (From Foreword to *The Religious Situation: 1968,* Donald Cutler [ed.]; Beacon Press, 1968.)

Religious faith, Niebuhr affirmed, is such a transmutation.

The role of ritual in religion

The French sociologist Émile Durkheim developed a more thoroughly sociologistic view of the origin of religion. In *The Elementary Forms of the Religious Life,*he focussed attention on the rites, cult organization, and shared sacred beliefs of society, which itself was the object of religious veneration.

> So everything leads us back to this same idea: before all, rites are means by which the social group reaffirms itself periodically.

Although propounded as a theory of religious origins, Durkheim's theory, like most other origin theories, stands or falls on its ability to aid in the understanding of the persistent interconnections of religion and society, not of the historical origin of religion. Freed of its exaggerations, it furnishes a necessary complement to individualistic theories. Like Plato, Durkheim emphasized religion as a source of social integration at a time when social dissolution was feared and experienced.

Karl Marx also developed a social theory of religion in the 19th century, but it stressed the relationship of religion to social conftict rather than to integration. Since command–obey relationships are viewed as the central facts of society, religion, it follows, will reflect them. Religion was thus viewed by Marx as "the opiate of the people" administered by those in high positions to preserve their power. One of the ironies of history is that the partial truth of the Marxian theory of religion is well documented by communisni as a religion. The opposite truth—religion as a stimulant of the people—is also documented by communist movements, as well as by other sectarian protests.

RELIGION VIEWED SOCIOLOGICALLY

Many attempts have been made to describe the essence of religion and to indicate its basic dimensions. Rudolf Otto, a German theologian and philosopher, described the essence of religion as the experience of "the holy," which engenders a dialectic between two contrasting attitudes: awe and dread (*mysterium tremendum*) and wonder and attraction (*mysterium fascinans*). The Belgian-French ethnologist Claude Lévi-Strauss stated that the essence of religion is embodied in its symbol systems (*i.e.,* its myths), which express the basic qualities of social reality. Such formulations have stimulated the work of many scholars but are themselves too general to form the basis of an empirical science of religion. Joachim Wach noted that all religions, despite their differences, are characterized by systems of belief, of worship, and of organization. He referred to these as the theoretical, the practical, and the sociological expressions of religion. Although belief aspects seem most visible to modern man, rites and worship and the groups that sustain them are no less important and perhaps historically prior to elaborated systems of belief.

Religious systems

Another line of thought views religion as basically the effort of man to bring some order (cosmos) into the universal experience of disorder (chaos). Baffling, unpredictable, and uninterpretable chaos impinges upon man in many ways, but especially at points of unmitigated suffering, of persistent violation of principles of justice as expressed in a given society (for everywhere "the ungodly" appear to prosper), and of incomprehensible mystery and meaninglessness. These three manifestations of chaos are particular expressions of more general categories of human experience labelled, in the work of the American sociologist Talcott Parsons, desires, values, and ideas. In terms of his vocabulary of motives, responses to these experiences are cathectic, evaluative, and cognitive, which in religion refer positively to efforts to achieve salvation, justice, and truth, or, negatively, to avoid suffering, injustice, and error.

Many secular activities, however, are also involved in these quests. Religion is differentiated from them because of the ultimacy or final quality of the effort. Knowledge appears to fail man just when he most needs it—to make some sense out of existence. Efforts to produce a just society seem always to fall short. Man reduces some forms of suffering only to discover that his capacity to create new forms is impressive. Thus, man's intellectual, legal-moral, and magical systems seem inadequate unless they can be made to go beyond the crushing impact of the reality of experience. Religion is obliged, as Durkheim said, to "pass science and complete it prematurely" in the effort to keep the experience of chaos within limits that man can sustain. Religion, thus viewed, is the expression of man's attempt to formulate some ultimate solution, which is more abstract, less ephemeral, and, in the last analysis, beyond tragedy.

Religion, viewed here as a generic phenomenon and not as a series of specific religions, will now be considered in terms of the beliefs and practices by means of which a group (1) designates its deepest problems of meaning, suffering, and injustice; (2 j specifies its mosi fundamental ways of trying to reduce those problems; and (3) seeks to deal with the fact that, in spite of all efforts to eliminate them, meaninglessness, suffering, and injustice continue.

TYPES OF RELIGIOUS ORGANIZATIONS

Variations in religious organizations

Even a quick glance at the range of religious organizations reveals that they vary widely in a number of ways. They differ, for example, in the degree to which local groups are bound together into larger structures, recognizing their religious kinship; in the number and level of training of professional leaders; and in the extent to which a hierarchical or bureaucratic structure has been created. On the one hand there are small, independent local groups; on the other, there are international ecclesia (religious organizations) that bind thousands of local units together under a professionally trained, hierarchical leadership. A somewhat different contrast can be drawn between diffused and specialized religious patterns. In some societies (particularly isolated, preliterate societies) one finds an organizationally indistinct religious process, diffused through the social system, that is an integral part of familial, political, economic, and artistic activity. In more heterogeneous and complex societies, specialized religious institutions as well as religious specialists have appeared, with at least the possibilities for lines of action separate from other major institutions.

Viewing these facts, the social scientist asks two clusters of questions: Can the great diversity of religious organizations be classified; would some system of taxonomy (classification) help in viewing the range more clearly? And how does one account for the diversity; does it reflect, and in turn perhaps affect, secular diversity?

Church and sect types in Christianity. Fundamental answers to these questions were given by the German church historian Ernst Troeltsch, primarily with reference to Christian materials. Building on an earlier distinction between priest and prophet, he described two basic types of religious organization, the church and the sect. These are best understood as abstract types, marking the end points of a continuum that ranges from the most thoroughly established, formalized, and complex religious organization to the least established, least bureaucratized. These types express, in Troeltsch's judgment, differences in religious values and emphases; but more particularly, they express different ways of dealing with the secular world and its power.

The church, typologically, recognizes the strength of the secular world but neither abandons the attempt to influence it nor risks quick defeat by challenging the secular powers directly. The economic, political, and military powers that be are seen, in the Roman Catholic phrase, as part of the "relative natural law," as proximate goods, to be accepted, yet criticized. The church is built, therefore, on compromise. It

> . . . utilizes the State and the ruling classes, and weaves these elements into her own life; she then becomes an integral part of the existing social order; from this standpoint, then, the Church both stabilizes and determines the social order; in so doing, however, she becomes dependent upon the upper classes, and upon their development. (From Ernst Troeltsch, ***The Social Teaching of the Christian Churches;*** Olive Wyon ftr.], The Macmillan Co., 1931.)

Group and individual needs

Religion, viewed functionally, is connected with two partially competitive sets of needs. On the group level are the needs for stability, for predictability in the behaviour of others, and for processes that control the potentially destructive responses of individuals to tragedy and severe frustration. On the individual level is the need to handle problems of tension, guilt, anguish, and frustration. These societal and individual aspects are partially competitive, because processes that promote integration may do so only by demanding sacrifices from individuals that increase their anguish; or individuals may pursue salvation from their most grievous difficulties in ways that disrupt social order. The church, as a type, strives to be coextensive with society in its membership and thus to maximize integration. In the effort to achieve this, emphasis is placed on sacrament and creed, not on right behaviour. In the logically extreme case, the church type lends itself to the support of the secular order, even when it is authoritarian.

The sect, on the other hand, is best understood in connection with efforts to satisfy fundamental individual needs. Individuals who feel particularly deprived (and there are many forms of deprivation) may feel alienated from society and from the dominant religious organization closely connected with it. Participation in a sect expresses that alienation. In the extreme case, it is purely a religion of laymen, free from both worldly and ecclesiastical authority. The sect repudiates what it views as the compromises of the church, preferring isolation to compromise; its members cannot believe that a society that has filled them with such anguish or a church so well accommodated to that society is worthy of allegiance.

Troeltsch's description of the contrast between church and sect is a good starting point for the classification of religious organizations. It requires modification, however, in two ways: Since it is based wholly on Christian materials, its value in the analysis of other religious traditions must be examined. And as a dichotomy, even if the two types are understood as the extreme cases of a range, it leaves unexamined the wide variety of organizations within each type.

Structured and unstructured types in other religions. When one moves from a Christian to a universal perspective, the need for a distinction between formally organized, highly structured churches and religion as a pervasive quality of the social system becomes apparent. Joachim Wach distinguished between a religion that is coterminous with a "natural group" and "specifically religious groups." In a tribal or other homogeneous society, one does not belong to a church; he belongs to the society, which has certain religious qualities. In modern urban societies, the two memberships are separable, in fact and in imagination. The Chinese-American sociologist C.K. Yang suggests "diffused religion" and "institutionalized religion" as terms that convey this distinction.

Diffused and institutionalized religion

This distinction has influenced the comparative study of world religions in various ways. Hinduism and Buddhism, for example, have no "church" in the Troeltsch sense. This is not true if what is meant by "church" is an inclusive religious structure, accommodated to the world around it, to which individuals are affiliated by birth rather than choice. It is true, however, to a substantial degree if "church" conveys the connotations of formal organization and hierarchical structure, for these are less characteristic of Hinduism and Buddhism than of Islām and Christianity. Thus, churches should be identified as more or less institutionalized or diffused. Or perhaps the term ecclesia should be used to avoid the somewhat limiting connotations of "church."

Religions vary, however, historically and geographically in the degree of their institutionalization. Although Hinduism, for example, tends to be diffused, not institutionalized, and has no congregation in the Western meaning of the term, Brahminism (in the sense of an organized and partly hierarchical Hindu priesthood) has been characteristic of some periods in Indian history. Within the Buddhist tradition, the *saṅgha* (the organization of monks) represents a complex ecclesiastical structure in Theravāda (Way of the Elders, or southern Buddhist) countries, such as Thailand.

The concept of church, or ecclesia, thus has general analytic value when subtypes are described on the basis of their degree of institutionalization.

The sect type. An adequate formulation of the concept of "sect" requires a similar process of refinement and extension. The members of sects share in common a strong sense of deprivation and alienation. Since they can differ widely, however, in the forms of deprivation and alienation, they have quite different kinds of relationships to the surrounding society. In purely typological terms, sects can be described as shared and heightened expressions of three fundamental religious styles, each indicating a different form of deprivation. An American scholar, J. Milton Yinger, has suggested that these three styles are those of the prophet, the ascetic, and the mystic.

> The prophet sees himself outside the structure of power; he feels alienated from the social structure. He has not given up on the world; he wants power over it. The ascetic sees himself outside the cultural system; he feels alienated from prevailing values. The values he most esteems seem beyond realization within the human communities he knows, so he is ready to withdraw to achieve them. The mystic sees himself outside the usual motivational systems; he feels alienated from himself and lonely; he lacks morale; he feels guilty and baffled. He may join with others, without reference to the world around them, to achieve poise, insight, and control over their spiritual and physical malaise. (From J. Milton Yinger, ***The Scientific*** Study ***of Religion;*** The Macmillan Co., 1970.)

These prophetic, ascetic, and mystical sects are analytic types, not actual organizations. Particular sects may develop values and activities with reference to all three types of deprivation. There may be internal struggles within the group, accounting for frequent schisms, because individuals want different emphases. Nevertheless, description of pure types can facilitate cross-temporal and cross-cultural comparisons.

An appropriate taxonomy (classification system) for religious organizations is helpful in answering the problem of diversity. The social scientist looks for an answer not in the contrasting beliefs and rites---indeed, he sees these as the dependent variables to be explained--but in the structural and cultural differences within a group or between groups. The American theologian H. Richard Niebuhr, in his ***The Social Sources of Denominationalism,*** has stated that if men had similar needs, were placed in equivalent positions in the social structure. and were trained to share the same cultural heritage, their religious outlooks would be quite similar. Religious divisions and schisms occur, in fact, because of various combinations of the following highly interactive influences.

Individual variations. Individuals vary in their feelings of frustration and guilt, in the level of their intellectual development, in their tendencies toward authoritarianism or dogmatism, in their capacities for the several kinds of religious experience, and in many other ways. The beliefs and rites that are resonant and meaningful, therefore, vary among them.

Economic and political variations. There are differences in economic and political interests. Since religion is involved in social conflict as well as integration, it carries the imprint of these interests. The German philosopher Friedrich Nietzsche described religion as an expression of class-determined resentment, an–effort by the powerless to enchain the powerful. This was a great oversimplification, but it focusses attention on the interconnection of religion and stratification systems. The German sociologist Max Weber is closer to the mark in his discussion of the contrast between upper and lower class religious views. On the one hand, there is a theodicy (justification of evil in accordance with a view of a good and righteous God) of good fortune—the beliefs and rites that serve to justify one's relatively favourable position and give assurance that it will not disappear. On the other hand, there is a glorification of suffering—perhaps even the seeking out of suffering to achieve salvation—by the powerless; or the theodicy of rebirth, as in Hinduism and Buddhism; or the religious downgrading of worldly success, Struggles within a religious group often reflect these contrasting locations in the social structure. The prophetic and pharisaic (legalistic reinterpretive) elements in early Judaism, for example, match the economic and cultural conflicts between the seminomadic shepherds and the settled farmers, the lanaless and the landowning, and the artisans and the nobles.

Societal and nutional variations. Even the world religions manifest wide societal and national differences. They carry the imprint of the demographic, historical, and institutional facts of the societies of which they are a part. Thus, one must note not only the contrast between the Theravāda (southern) and Mahāyāna (northern) Buddhist traditions but also the national variation within each. It is not enough to note the distinctions between Buddhism in China, where it interacts in a large and complex society with many other major religious traditions, and Buddhism in Thailand, where its influence is relatively unalloyed. The Theravāda Buddhism of Burma differs from that of Thailand in important ways. Differences in the national expressions of religious traditions persist, often for several generations, even when they have been transplanted into new nations, as shown by continuing variation in religious beliefs and practices among the several Lutheran or Roman Catholic groups in the United States.

National contrasts within a religion

The student of such national and societal differences within a religious tradition must also be a student of variation in the interpretations of the differences. Most scholars would agree that it is only on a fairly abstract level that one can speak of Islām, without noting the contrasts among the African, Arabic, Pakistani, Indonesian, and other varieties; but it is well to note also that the interpretations of the differences vary, as between Islāmic and non-Islāmic scholars, and within each of these groups.

Social mobility and change variations. Social mobility and social change do not affect the members of a society in the same way or at the same speed; and one of the consequences of the differences in experience is variation in religious perspective. In the 20th century every major religious tradition is divided between traditionalists and modernists, literalists and revisionists, partly as a result of differences in the stability of individuals' social worlds.

Internal variations. Variation in religious organizations derives in part from the internal development of the religious system. The elaboration of theological ideas, inventiveness in the field of aesthetics or ritual, and differences in judgment concerning the most effective way to maintain or strengthen the faith all express themselves in different religious organizations. They do so, however, not because of some independent power but because of the responses of possible constituencies. Thus, these internal religious developments help to account for church–sect and other contrasts only as they are connected with the other previously noted factors and with the needs and desires of various audiences. In Weber's thought there is an "elective affinity" between particular religious ideas and the circumstances faced by given groups. This can result either in internal differentiation or in conversion, which may be defined as the shifting of loyalty across a major religious line. Thus, lower caste Hindus, when change and mobility have opened up to their imaginations new religious perspectives, become potential candidates for conversion to Buddhism; and some American Negroes, under similar circumstances, join a variant of the Muslims.

Contextual differentiations. Viewing the question from the perspective of religious leaders, there is a strategic factor to be noted as an important cause of religious differentiation. According to contemporary theories of administration, complex organizations are often faced with dilemmas, both in dealing with their own members —whose allegiances and motives are mixed—and in dealing with the external environment. As a religious organization (or any other group) seeks to maintain or broaden its influence, it often has to win the support or neutralize the opposition of others. The question thus arises: shall it oppose other groups and member motivations that are not supportive of the organization's purposes directly, compromise with them, or try to co-opt them into its own activities? The dilemma of religious organizations is not unlike that of political parties: direct opposition to powerful forces may bring defeat; compromise may require yielding on major values; co-optation raises the question of who will in fact dominate whom. The dilemma is clearly illustrated in the variety of relationships between church and state. According to Yinger:

Dilemmas of complex organizations

> Those who believe that clear separation of church and state increases the power of the church emphasize the freedom from political domination, the freedom to criticize the political process and the secular power structure. There is the danger, however, that such freedoms are closely connected with powerlessness. On the other hand, close institutional connection between church and state scarcely avoids the dilemma, because the union raises the likelihood that the church will be used to lend sanctity to a secular power

> structure. The problem is to **find** a way to be simultaneously in politics [thus to influence it] and beyond politics [thus free to challenge it]. (From J. Milton Yinger, ***Sociology Looks at Religion;*** The Macmillan Co., 1963.)

Those who support the churchlike response believe that compromise is better than isolation, that to press too hard is not to win but rather to alienate potential supporters. By accepting what in any event can scarcely be avoided or directly changed, at least by a religious movement (war, or inequalities of station, for example), churchmen hope to remain in a position of influence from which they can gradually remove the causes. Sectarians disagree with this strategy. They view compromise as a one-way street down which churches move until they have left their original values behind. The way to deal with the dilemma, in the sectarian view, is to maintain the purity of the religious system of values in an uncompromised community of believers. Some sectarians may then define the dilemma as irrelevant, for they seek no influence beyond the group. Others envisage victory for the sectarian ideal by human effort or miraculous event. Still others believe that the vocation of the religious person and group is to maintain the religious ideal in the hope and expectation that it will be a beacon for others.

These several sources of religious differentiation are highly interactive. Those who support a sectarian approach for strategic reasons, for example, are often those who are seriously deprived, and thus their class location is also involved. A society that treats them so poorly is clearly not one with which they want their religious group to compromise. These same persons may be among the most seriously affected by the stresses of social change.

Ecumenism

Contextual **affinities.** If certain cultural and structural facts tend to produce religious divisions, other facts tend to produce ecumenism (religious cooperation). Those concerned with the social aspects of religion must therefore be concerned with fusion as well as with fission, with the social sources of ecumenism as well as with the social sources of denominationalism. Although the forces of division are strong in the 20th century (reflecting the sharp differences in experience and in opinions about the conflicts of the day), tendencies toward ecumenism are also strong. A World Fellowship of Buddhists was formed in 1950; the Roman Catholic second Vatican Council (1962–65) gave serious attention to the question of the relationship of that church to other Christian churches and to other religions; Protestant and Orthodox bodies have extended the scope and the reach of the World Council of Churches; cooperation among Orthodox, Conservative, and Reform congregations in Judaism has increased.

The term ecumenism is thus used in a variety of ways, referring to activities ranging from tentative conversations among formerly separate religious groups to the achievement of full unity. Ecumenism may thus be viewed as a variable rather than as a condition, with points marked along a scale with familiar terms. The scale might therefore range from toleration—through conversation, cooperation, and federation—to integration, indicating progressively stronger ecumenical activity. This range is meaningful, however, only when another scale is used at the same time, a scale that indicates which religious groups are being drawn together. Integration of two Methodist bodies, for example, may be a smaller ecumenical step than serious conversation between a Christian and a Buddhist group.

Viewed sociologically, the trends of this period of relatively active ecumenism are a product of shared experiences, perspectives, and threats. Within the United States, for example, religious differentiation based on national origin, regional isolation, rural–urban contrasts, and language backgrounds has been reduced in the latter half of the 20th century. The appearance of powerful and competitive quasi-religions, communism in particular, has caused traditional groups to view their shared qualities —and their shared competition—more clearly. And the transportation and communication revolution of the 20th century has created and made dramatically visible a powerful ecumenical fact: inhabitants of the earth are vitally interdependent.

RELIGIOUS LEADERSHIP

Types and functions of religious leaders

There are many ways to describe the range of types of religious leaders. Wach distinguishes among the following: founder, reformer, prophet, seer, magician, diviner, saint, priest, and religiosus (a plain man who lives a highly religious life). Studies that focus on the contemporary scene emphasize the range of functions, more than types of leaders, although these may be partially equated by a division of labour. Marshal Sklare, a U.S. sociologist, designates eight functions: priest (conductor of public worship), preacher, cleric (a functionary of the state), rector (administrator of an organization), pastor (counsellor), father (head of a congregation in a psychological sense), parson (representative of the church to the community), and rabbi (teacher and interpreter of religious doctrines). Such diversity of functions is found only in highly differentiated societies in which religious organizations have become complex. Social aspects of religion are strongly influenced by the factors in the recruitment of these various types of leaders (*e.g.*, the social classes to which they belong, their educational levels and personality types), by the degree of self-selection, by the extent and nature of their training, and by the extent to which they are gathered up into or under the direction of complex organizations—*e.g.*, the local Muslim *imām* (religious leader) into the *ʿulamāʾ* (the body of men trained in Muslim law and theology), or the Buddhist monk into a regional or national *saṅgha.*

More analytic studies of religious leadership deal with the relationship of the leader to the surrounding society and to processes of social change. In anthropological work, a distinction is frequently drawn between the priest and the shaman, the former a religious leader trained to fulfill a culturally defined role in a complex organization, the latter a person possessed of personal divine power and involved in less structured, noncalendrical rites. In sociology, the distinction is more often drawn, as by Weber, between the priest and the prophet. Priests are trained functionaries of an established religious system. Prophets are the bearers of charisma—an ability to convince others that they possess supernatural or at least excep tional powers, not by virtue of their office but by virtue of their personal qualities. The concepts of charisma and prophet were not adequately developed by Weber. They have something of an ad hoc (particularistic) quality, making them appear to be designed to help account for drastic change in the face of Weber's emphasis on bureaucratization and rationalization. In view of this, the question as to who is ready to grant charisma to which candidates (for many prophets go unheard) and what their relationship to the tradition from which they come (however much they may challenge it) needs to be asked.

A further analytic approach to religious leadership refers to those persons who give intensive expression to one of the three primary roots of religion and who have become identified with a movement of which that expression is the focus of attention. The mystic promises enlightenment; the ascetic is most concerned with suffering; the prophet seeks justice. These are analytic types, however, not specific positions; individual leaders are likely to combine them. Moreover, they cut across the types of leaders described in terms of organizational structure. Each approach has its value in the study of leadership.

RELIGION AND SOCIETY

The concept of function

Religion as sociofunetional. The universality of religion leads readily to the assumption that it serves essential functions for individuals and society, or both. This observation expresses a theory that has been widely used, and sharply criticized, partly because the concept of "function" has several meanings. Two of these are critical for the student of religion. One meaning is a given structure or process that performs an essential service. In mechanics, for example, a carburetor functions to mix gasoline and oxygen to make combustion possible in an automobile engine. Without it an automobile could not run. Religion, functionally, serves to give men shared goals, thus reducing the sharpness of their competitive goals; it promises later rewards, thus softening the anguish of present frustration; it helps to define the meaning of suffering, thus reducing the threat of chaos. Religion, thus, is a functional prerequisite for society.

But function also means consequence. If religion is a function of society and of individual need, it varies with them, reflecting their changes; it is an effect. Functional theory ties these two meanings together, although some of its advocates have used only the first meaning and wondered why they were criticized; some critics have used only the second meaning and wondered how it could possibly be used to show the effects of religion. When the two meanings are used together, functional investigation focusses on those processes within a system in which a product of the system (in this case religion) helps to maintain that system by reducing or eliminating otherwise destructive processes. This is an influential theory in the biological and social sciences because they work with evolutionary and cybernetic, or "feedback," models. A system becomes what it is partly because of the effects it has had, which help it to adapt to its environment. In evolutionary terms, those societies that did not invent religion did not survive. In psychological terms, the pattern is reinforced because of its consequences. An effect triggers a signal or a process that contributes to adaptation, integration, and development.

Functionalism emphasizes in a valuable way the interdependence of society and religion. If it is narrowly interpreted, however, it leaves several critical questions unanswered. One of these is of special importance for the student of religion: if religion, which in the first instance is a product of certain social and individual forces, serves functions of adaptation and integration for the system out of which it comes, the question as to what are its consequences for smaller and larger systems, which are interdependent with the first and which may have conflicting needs and inclinations, must be asked. If religion integrates a society that is highly repressive for some of its members, its functional aspects for them must be investigated. If it integrates in such a way that it arms a society with righteous anger to attack another "heathen" society, the question concerning its functional aspects for the latter must also be investigated.

Religion as dysfunctional. These questions suggest that functionalism is not only incomplete without conflict theory but that it is one segment of a larger theory, of which the study of conflict is also an integral part. To simply add functional and conflict theories together, eclectically noting that both processes operate in society, is not a sufficient answer for the social scientist. They are best seen as a unit. Lack of conflict can be functional or dysfunctional for a system. (Pain can be a signal that some action is needed; the lack of such a signal can be catastrophic.) Conflict can be functional or dysfunctional. In assessing these various possibilities, it is essential that the system of reference be specified.

Any bland interpretation that religion preserves sociai order or lowers individual anxiety disregards the frequency with which religion expresses the deep rifts in society or increases the possibilities of anxiety. If one person's life is given meaning and coherence by a faith that constrains the life of another or makes him the object of attack, the total level of anxiety may not be lowered.

The role
of conflict
in religion

The conflict potential of religion is expressed in sectarian and reformation movements, in the development of new religions, and in the continuing confrontation of different religious communities within one society. The Levellers and Diggers (religiosocial reform movements of 17th-century England) used sectarian protest to challenge the establishment, both secular and religious. The Black Muslims in contemporary America demonstrate the use of a new religion to challenge a society, not to express integration with it. Catholic–Protestant conflicts for several centuries in Northern Ireland indicate that religious difference can symbolize and strengthen divisive tendencies within a society.

These are nor vaiuarive statements. Whether or no1 the potential of religion to contribute to conflict is desirable can be stated only in terms of specified values and with knowledge of total, long-run consequences. The harsh, antagonistic stance of the Black Muslims may help the members to rid themselves of self-doubt, encourage them to accept disciplines and responsibilities that a "white man's religion" could not do, and alert the dominant society to grave injustices and great suffering that it apparently has disregarded. Whether these will be the long-run outcomes, rather than increased tension and conflict, mainly depends not on the religious doctrines but on the social setting in which they are heard.

Reciprocity in society and religion. The relationships of religion to society, then, are best understood not by separate applications of functional and conflict theories but by consistent application of the whole system, or field, theory.

From the perspectives of the social sciences, religious organizations are part of the social structure, religious values part of the general culture, religious motivation part of individual character. For purposes of analysis, however, the religious element in social life can be separated from the rest in order to study its interactions with the polity, the economy, the stratification system, and other elements of society. These relationships are too complex to examine in detail but will be referred to briefly.

Religion and social organization. Religion everywhere is closely connected with family structures. All societies have culturally designated ways in which sexual behaviour and reproduction are regulated; status assignment is made (indicating who stands up for a child, giving him his name and social position), socialization is carried out, and affection and personal support are made available. Although a variety of structures can serve this complex of functions, they are so vital that no society leaves them simply to chance or to individual initiative. In virtually every instance, moreover, they are reinforced by the religious system. This is particularly true in those societies in which the kinship system is the centre of social organization. Family deities and ancestor worship may be important parts of the religious life in such societies. The ethical norms governing interpersonal behaviour within the family are often given religious sanction (*e.g.*, "Honor thy father and thy mother," as in the Ten Commandments of Judaism and Christianity; marriages, monogamous and permanent, are "made in heaven," as in popular piety). Even in societies in which occupational, locational, recreational, and other associations have modified the pre-eminent place of the family, there continue to be important religious elements in socialization; religious endogamy (marriage within one's group) remains common; and the basic structure of the family — the place of women, the importance of children, the possibility of divorce, and the like—continues to be supported by religious norms and training.

Religion
and social
stratification

Religion and social differentiation. The distinction between church and sect indicates that religion also is involved in social stratification. Whether it be a caste system, a majority–minority system, or a class system, there is religious variation among the strata. In Buddhism, the ways in which one seeks merit—*i.e.*, the means to rebirth in a higher life — vary widely by class because expectations, resources, training, and motivation vary. The wealthy have time and training for meditation, or they may make a large gift to a monastery. The poor also make gifts, perhaps at greater sacrifice than the wealthy; but they also express devotion — in which they are at no disadvantage.

The conflicts among social strata are also expressed, in part, religiously. In some instances this will be in the form of a deflected or displaced attack on the social order by the lower strata, who proclaim that "the last shall be first" in some heavenly kingdom. But religion may arm the lcwly for purely earthly battles (as it often arms the powerful). This happens under various sets of circumstances. In one case, religious protest develops shortly after some independent and prosperous era has ended. A giorious (often partly romanticized j past is seen through the frustrations of an oppressive present. Thus, the Ghost Dance swept through a number of American Indian tribes in 187%-90—particularly among the tribes of the plains—a period when their own independent cultural lives were still a vivid memory, though their contemporary lives were filled with frustration. Although there were several versions of the Ghost Dance, some more conflict-oriented than others, they all envisioned the restoration of the land to Indians, the return of their cultures, and the annihilation or the expulsion of the white man.

Religion may also support a protest movement among members of lower strata who are beginning to experience some improvement of status, perhaps after generations of repression. Their hopes may soar far more rapidly than opportunities open up to them. Their comparisons, then, are with a glorious future, not a glorious past; but these produce the same sense of unjust deprivation, relative to the standard before them. "I have a dream," the American civil rights leader Martin Luther King, Jr., proclaimed, and the dream was the basis of his militant, if nonviolent, attack on racial discrimination. In his "Letter from Birmingham Jail" he declared that civil disobedience and extremism were religiously prescribed:

> Jesus Christ was an extremist for love, truth and goodness, and thereby rose above his environment. Perhaps the South, the nation and the world are in due need of creative extremists.

Religion and economics. No aspect of the relationship between religion and society has been more carefully studied than the reciprocal influence of religion and economic activity. Religion influences an economic system in many ways. Worship and holidays affect the timing and rhythm of work. Occupational assignment, as in the Hindu caste system or, more individualistically, in the Christian doctrine of the "calling," may be religiously prescribed or supported. The economic value of goods and services is partly determined by the religious definition of their value, which can raise or lower both demand and supply. The distribution of wealth is partly governed by religious norms, which may indicate the deserving and the undeserving. Religious organizations, activities, personnel, and buildings may command a substantial proportion of the resources of a society. This is particularly true of technologically underdeveloped societies and of those with a precarious economic existence, for there is a negative correlation between economic productivity and the proportion of the resources of a society spent on religion.

The Protestant Ethic

The relationship between religion and economic development, however, has received the most intensive examination. The issue was raised in a decisive way by Max Weber in his essay ***The Protestant Ethic and the Spirit of Capitalism.*** Unconvinced by the Marxian interpretation of religion and of economic development, Weber carefully examined the relationship between the areas of rapid capitalist advancement and religious allegiance. In countries with both Protestant and Catholic citizens, he noted, business leaders, owners of capital, technically trained persons, and skilled workers were "overwhelmingly Protestant." He did not deny the importance of economic factors:

> . . . we have no intention whatever of maintaining such a foolish and doctrinaire thesis as that the spirit of capitalism . . . could only have arisen as the result of certain effects of the Reformation, or even that capitalism as an economic system is a creation of the Reformation.

Weber did emphasize, however, what he saw as the peculiar spirit of economic enterprise among early Protestants, a spirit characterized as disciplined, rational, and highly ascetic. These qualities expressed the 16th-century Reformer John Calvin's desire to be not simply a "vessel of the Holy Spirit," which expressed Martin Luther's mysticism, but "the tool of the divine will."

> Since Calvin viewed all pure feelings and emotions, no matter how exalted they might seem to be, with suspicion, faith had to be proved by its objective results in order to provide a firm foundation for the *certitudo salutis* [or certainty of salvation].

Religious ethics and economic action

Probably no book in the sociology of religion has aroused so much commentary and criticism as ***The Protestant Ethic.*** A substantial library has been written proclaiming it as a theoretical work of the first rank, denouncing it as absurd, substituting a different religious ethic for Protestantism, or qualifying the thesis in various ways. This is scarcely surprising, considering the complexity of the problem and the difficulty in excluding ideological perspectives. Weber skillfully demonstrated that religious ethics are part of a complex of forces that influence economic action. He less skillfully examined the evolution of the Protestant ethic in its step-by-step modification as it interacted with the economic setting. Or, to put it in another way, he paid insufficient attention to the selectivity of the groups who chose Protestantism (in part precisely because of its congeniality with emerging economic conditions), who modified it and who then were significantly influenced by its doctrines and obligations. Calvinism (Reformed or Presbyterian) had within it various potentialities for development into a religious attitude that would be meaningful to people who were also engaged in, or wanted to be engaged in, capitalist economic activity. These potentialities were much more limited in Lutheranism and Catholicism. The religious spirit of which Weber wrote, however, was not simply a product of economic motivation. It evolved partly as a result of its own "inner dialectic." It helped to shape the life style and the goals of its adherents and thus served to intensify a peculiarly rational and frugal economic pattern that was developing as a result of more mundane forces.

The Protestant Ethic, first published in 1904, was the initial volume in a series of studies in which Weber explored the significance of religion for economic matters. In studies of Confucianism and Taoism, Hinduism and Buddhism, and ancient Judaism, Weber examined the ways in which religious ideas blocked or supported the development of rational economic enterprise. In later writings, he took account of the interactions among economic, political, and religious influences, thus approaching the question from a broader perspective than that employed in ***The Protestant Ethic,*** in which he described only one side of the causal chain. In each of these later works, Weber concluded that the religious influence restrained economic development, although at different levels of intensity. Even in the face of extensive economic opportunity, as in China, traditionalism prevailed, partly as a result of the strength of Confucian traditionalism. Weber accounts for Europe's being freed from an equally or even more restrictive Christian traditionalism by reinterpretations of Protestant prophecy. A major theoretical issue then emerges: why did the prophets appear in Europe and why did Hinduism or Confucianism not have a Reformation? Thus, there is a need for recognition of the gradual and mutual modifications that appear in a situation of rapid change.

The study of the interaction of religion and economic development received renewed attention in the 1960s because of the efforts of many technologically underdeveloped societies to create economies with the capability for self-sustaining growth. The applicability of Weber's thesis to Japan and to contemporary United States has also been studied. The range of issues is too wide to examine here; but the following generalizations are broadly supported by contemporary research. Traditional religions are proving to be only minimally supportive of economic development in the technologically less advanced societies, although most of them are capable of reinterpretations that permit, if they do not promote, economic change. Japan's economic transformation did not require prior religious transformation and may have received some religious stimulation. In the United States, those brought up in the Protestant tradition adapted more quickly to the possibilities of economic achievement than did Catholics, but this contrast subsequently has been sharply reduced if not eliminated entirely. In the technologically underdeveloped world as a whole, the process of making new types of men and new societies has been led mainly by revolutionary quasi-religions, not by reformed traditional religions.

Power, prestige, and possessions

Religion and political development. Political and religious orders are everywhere interactive. Both are involved with questions that arise from the fact that many of the values for which men strive—power, prestige, and possessions—are in scarce supply. If each person pursued them by means of his own choosing, organized societies would be impossible. Political institutions, therefore, are the structures that assign ultimate coercive power--even to the extreme of the administration of death—to certain procedures and certain individuals in order to enforce approved ways of achieving life's values.

In some societies, religious and political structures are coterminous, a characteristic situation in relatively isolated, preliterate societies. The gods of the group guarantee or represent its values; the roles of citizen and believer cannot be separated. This does not mean that conflict is absent. Indeed, conflict often finds religious expression in such societies; but it is handled in such a way that it is less likely to increase or produce a permanent split in the social structure.

When societies become more heterogeneous, when some groups become acquainted through cuiture contact with different sets of values, and when sharper lines of stratification appear, the stage is set for a different pattern of relationship between religion and politics. Because these trends may loosen the hold of social norms, some members of the society will acquire an instrumental attitude toward religion. The ruling class begins to see it as a means of preserving order—*i.e.*, helping to guarantee their positions. This is not so much cynical manipulation as ready belief in a congenial faith. In Islāmic societies, to challenge the ruler may be to challenge the caliph (Islāmic political leader) as a religious leader. A U.S. sociologist, Kai Erikson, has noted that the antinomian (anti-legalistic) controversy in the Massachusetts Bay Colony (in 17th-century America), although couched in religious language, was primarily a political struggle. The elders could fight Anne Hutchinson and her supporters more effectively and more earnestly on doctrinal than on political grounds.

Separation of religious systems and the state

Implicit in the partial separation of religious and political identities is the possibility of a third pattern of relationship—religious systems that are set sharply apart from the state. Religious specialists may develop ideas that contradict secular claims and values. This seems to occur under conditions of persistent social change, social contact, and widespread and prolonged suffering. The religious imagination leaps the boundaries of the society within which that suffering has been felt and propounds a universal system. In the classic case, Yahweh worship was transformed by the 8th-century-BC prophets of Judaism from a tribal religion, with its burnt offerings and sacrifices, into a monotheistic religion that emphasized repentance. And, as Leonard Trelawney Hobhouse, a British social scientist, wrote in comparing the development of Judaism with the development of Buddhism,

> . . . by a very different road and with much difference of implied meaning, we are reaching the Buddhist doctrine of renunciation and humility—those cardinal points of spiritualized religion.

These three levels of relationship between religion and politics do not represent a historical sequence. One is built upon the other, and all three are visible in contemporary societies. *Gott* mit *uns* ("God with us," a phrase found on German army uniforms in World War II) can be written in many languages, to illustrate the first level. Pres. Nguyen Van Thieu of South Vietnam ordered authorities in the 1960s to arrest Buddhist monks and Catholic priests who make "political sermons" to "inflame the people." He thus indicated their opposition and his interest in maintaining a religious neutrality if he could not win active support—illustrative of the second level of interaction. There are elements of universalism—the third level—in ecumenism, pacifism, opposition to discrimination, and the symbolic affirmation of the brotherhood of man in all of the world religions.

The complexity and importance of the issues involved in the various degrees of the separation of church and state is indicated in the sharp separation of the duties of Brahmins (priestly caste) and Kṣatriyas (military caste) in Hinduism, which has institutionalized the separation of church and state and has created a context in which a secular state could emerge. That Brahmins have developed a minimum of ecclesiastical structure supports these trends. In theory, Buddhism equally supports the separation of church and state, but under some conditions this relationship has been extensively modified. When a predominantly Buddhist nation confronts a non-Buddhist imperial power, for example, the nationalist movement may gain strength from religious themes, as in the case of Burma. In most instances, Islām supports the close identity of church and state. Even some contemporary states have an explicitly religious character, as documented by the official title of Pakistan, the Islāmic Republic of Pakistan.

Political and religious structures

Political and religious structures were closely aligned in early Judaism, but through history the pattern changed drastically. There was no likelihood that Jews would identify their religious community with the Assyrian, Babylonian, or Roman empires that held political power over them at various times. Political subordination has continued to be the experience of many Jews through centuries of what some regard as a continuing diaspora (dispersion into foreign lands). The impact of the development of modern Israel—where some 15 percent of the world's Jews live—on the creed and practice of church–state relationships is too recent to measure. Citizenship and religious identity in Israel, however, have once again been brought close together.

Christianity shows perhaps the widest variety of relationships of church and state, from extremely close union (not only in the formal but in the operative sense) to separation tinged with antagonism. This last aspect tends to persist for relatively brief periods, however, as in Mexico and the Soviet Union in the first few decades after their revolutions. The more standard relationship today is formal separation accompanied by numerous points of contact, cooperation, and interpenetration, as in the U.S.

Religion anti *sociocultural* transformation. Almost every issue thus far discussed involves, directly or implicitly, the question of the relationship between religion and social change. There are several possible answers to the question: religion and social change are unconnected; religion is the dependent variable, merely reflecting changes in its environment; religion prevents change (whether as an "opiate" or a "conserver of tested values"); religion is the independent variable, the source of major changes; and religion is part of an interdependent system and thus reflects, carries, and causes changes, depending on the point in a sequence that is being examined and the perspective from which a problem is being studied.

RELIGION AND PERSONALITY

A major connecting link in the study of religion, culture, and society is the field of culture and personality, which relates individual motivation and behaviour to the social context in which socialization and learning take place. An individual learns the religious culture around him just as he learns other parts of culture. Were this the whole truth, however, religious change, conversion to other faiths, withdrawal of allegiance, or the range of individual religious experiences within a given tradition would be difficult to explain.

The significance of the fear of death

In noting individualistic theories of the origin of religion, the most important of the universal experiences connected with religion are the fear and fact of death. Everywhere, the ways in which men meet the problem of death are in the realm of the sacred or holy (the transcendent). In societies in which any one of the world religions is dominant, the religious response to death often proceeds along two paths. One is the reaching upward to what is called the "great tradition," which, however, may be more stern and ascetic than the average man can achieve; the other is a reaching out to what is called the "little tradition" that is more attuned to immediate human need and capabilities. Thus, the Buddha (the 6th-century BC Indian religious reformer) did not make what he felt were false promises about immortality; he stressed instead the strength that could come from recognition of man's common helplessness. Alongside this form of Buddhism, however, are found ancestral and national cults that soften its stern message by dealing with the day-to-day ills that threaten to overwhelm man.

Other individual needs and experiences that are less obviously universal also enter importantly into religion. They require that attention be paid to variation in the direction and intensity of religious inclinations. This variation is partly the result of differences in socialization. Those who need religion are, in some measure, those who have been trained to need it—the training comes first, not the need. The need is partly a consequence of social role assignment. Women may be more religiously inclined in some societies than men because they have been socialized that way, though this disposition is partly an indirect effect of their role and its frustrations. In other societies, however, men's roles are more likely to encourage the public forms of religious worship, as in Buddhist, Jewish, and Islāmic societies.

Social learning and role influences help to account for variation in religious needs and activities, but more individualistic factors also enter in. In ***The Varieties of Religious Experience,*** the U.S. philosopher William James distinguished between the religion of "healthy-mindedness" and the religion of "the sick soul." The healthy-minded individual, in James's terms, views life optimistically; he does not linger over the darker facts of human experience. But the sick soul maximizes evil,

> . . . based on the persuasion that the evil aspects of our life are of its very essence, and that the world's meaning most comes home to us when we lay them most to heart.

This distinction was illustrated more than tested in James's work, and it is considered to be subject to value-laden distortion. Nevertheless, it stimulated an extensive amount of study of the varieties of religious experience, of conversion, ecstasy, possession, and revelation; of joy, terror, wonder, mystical encounter, and prophetic power; and of less exuberant feelings of contentment, enlightenment, and expansion of self. Much of this work is purely descriptive; but the accumulated material for a more analytic treatment of religious experience is being developed in the latter part of the 20th century.

Religious experience

Religious experience, according to social scientists, is not the direct expression of inner forces but is rather a product of a transaction between an individual with certain tendencies and the surrounding social and physical environment. In a study of trance ***(Journal for the Scientific Study of Religion,*** April ***1962)*** by Alexander Alland, a U.S. anthropologist, for example, the interactive influence of heat, high levels of carbon dioxide, loud, rhythmic music that leads to sensory deprivation, the presence of "significant others" to serve as models, lack of information, and isolation are examined. The description of the social setting does not mean that individual motivational factors are uninvolved;

> . . . receptivity to the trance is most certainly influenced by personal differences such as: range of experience, needs of the individual, and tolerance for various physiological stresses.

Religious experiences of this kind are best understood not as sharply different from other experiences but as one possible outcome of tendencies in a particular situation. A different situation can support the expression of those tendencies in another way. Students of comparative psychiatry emphasize that the behavioral significance of individuai anxiety is a function not only of its intensity and form but also of the interpretation given it by others. Consequences as diverse as acute schizophrenia (split personality) and shamanism (psychic transformation) may follow from the same underlying individual tendencies, depending on the way they are acted upon by others. According to the U.S. scholar Julian Silverman,

> In primitive cultures in which such a unique life crisis resolution is tolerated, the abnormal experience (shamanism) is typically beneficial to the individual, cognitively and affectively; he is regarded as one with expanded consciousness. In a culture that does not provide referential guides for comprehending this kind of crisis experience, the individual [schizophrenic] typically undergoes an intensification of his suffering over and above his original anxieties. (From Julian Silverman, "Shamans and Acute Schizophrenia," ***American Anthropologist,*** February 1967.)

An empirically based science of the personality factors in religion requires more reliable and valid measures than have yet been designed. A great deal of work is being done on this topic, however, not only of the case study variety, discussed above, but of a more quantitative sort. Thus, in comparing the levels of educational motivation of Protestants, Catholics, and Jews, for example, care is being taken to see that other variables are controlled, that only those alike on other grounds are compared. The meaning of individual "religiosity" is also being differentiated since it has many dimensions. To compare two "church member" samples is to learn very little if one does not know how much they are alike or different in beliefs, religious knowledge, or tendencies toward particular religious experiences. When the various "dimensions" of religion have been isolated, each can be put on a scale. Church membership, for example, can range from nuclear (dedicated and active participation), to modal (usual), marginal (occasional participation), and dormant (in name only). The importance of religion to the individual can be used as a variable, and variations in concern for suffering, injustice, and meaninglessness, as basic religious categories, can be measured to obtain individual views. Both the causes and the consequences of variations in these concerns are fundamental questions for the student of religion and personality. To explore them fully, he will need to connect the study of personality with the study of society and culture.

BIBLIOGRAPHY. MICHAEL P. BANTON (ed.), ***Anthropological Approaches to the Study of Relrgion*** (1966), a series of essays on religion from the point of view of contemporary anthropology; ROBERT N. BELLAH, ***Beyond Belief: Essays on Religion in a Post-Traditional World*** (1970), includes papers on "Religious Evolution" and "Civil Religion in America"; EMILE DURKHEIM, ***Les Formes élémentaires de la vie religieuse*** (1912; Eng. trans., ***The Elementary Forms of the Religious Life,*** 1915), a significant work on the functional and sociological approach to religion; S.N. EISENSTADT (ed.), ***The Protestant Ethic and Modernization: A Comparative View*** (1968), essays dealing with Weber's thesis and with studies of religion and modernization; E.E. EVANS-PRITCHARD, ***Theorres of Primitive Religion*** (1965), an excellent brief examination of anthropological theories of religion; SIGMUND FREUD, "Totem und Tabu" (pub. in ***Imago,*** vol. 1–2, 1912–13; Eng. trans., ***Totem and Taboo,*** 1918), Freud's basic statement on religion; CLIFFORD GEERTZ, ***The Religion of Java*** (1960), analyzes religion in change; WILLIAM JAMES, ***The Varieties of Religious Experience*** (1902), a basic document in the psychology of religion—insightful, but highly speculative; ARI KIEV (ed.), ***Magic, Faith, and Healing: Studies in Primitive Psychiatry Today*** (1964), a functional interpretation of magical and religious practices related to healing; VITTORIO LANTERNARI, ***Movimenti religiosi di libertà e di salvezza dei popoli oppressi*** (1960; Eng. trans., ***The Religions of the Oppressed: A Study of Modern Messianic Cults,*** 1963), useful as a source of information on religion as a protest movement; GERHARD E. LENSKI, ***The Religious Factor: A Sociological Study of Religion's Impact on Politics, Economics, and Family Life*** (1961), one of the most extensive studies of religion-community interactions, in the Weberian tradition;.REUBEN LEVY, ***The Social Structure of Islam*** (1957), a basic description of Islām as a social system; BRONISLAW MALINOWSKI, "Magic, Science and Religion," in JOSEPH NEEDHAM (ed.), ***Science, Religion, and Reality,*** pp. 19–84 (1925), a classic, although somewhat exaggerated, functional interpretation; DAVID A. MARTIN, ***A Sociology of English Religion*** (1967), a brief, well-written sociological interpretation of English religion; H. RICHARD NIEBUHR, ***The Social Sources of Denominationalism*** (1929), Troeltsch's approach extended and applied to America; REINHOLD NIEBUHR, ***Moral Man and Immoral Society*** (1932), the most sociological of Niebuhr's works; TALCOTT PARSONS, ***The Social System*** (1951), contains the basic structure of Parsons' theory, although extensively revised in later works; M.N. SRINIVAS, ***Caste in Modern India, and Other Essays*** (1962), opposes the belief that caste is a changeless and uqiform system by studying it in social context; R.H. TAWNEY, ***Religion and the Rrse of Capitalism*** (1926), a valuable study of Weber's Protestant Ethic that qualifies and enriches Weber's argument;

ERNST TROELTSCH, *Die Soziallehren der christlichen Kirchen und Gruppen* (1912; Eng. trans., *The Social Teaching of the Christian Churches,* 2 vol., 1931), the source of the church-sect typology, valuable as church history and as sociology; EDWARD B. TYLOR, *Primitive Culture,* 7th ed. (1924), a basic document of early anthropology's approach to religion; JOACHIM WACH, *Sociology of Religion* (1944), valuable classifications and typologies, less useful as a theoretical work; ANTHONY F.C. WALLACE, *Religion: An Anthropological View* (1966), an overview from the perspective of contemporary anthropology; MAX WEBER, *Gesammelte Aufsatze zur Religionssoziologie,* vol. 1 (1920; Eng. trans., *The Protestant Ethic and the Spirit of Capitalism,* 1930), fundamental to understanding Weber and contemporary theories of religion; *Wirtschaft und Gesellschaft,* vol. 2, ch. 4 (1922; Eng. trans., *The Sociology of Religion* (1963), a series of essays that contain much of the substance of Weber's theory of religion; BRYAN R. WILSON, *Religion in Secular Society* (1966), a study of the impact of contemporary social trends on religion; J. MILTON YINGER, *The Scientific Study of Religion* (1970), a contemporary systems or field theoretical examination of religion, drawing on sociology, anthropology, and psychology.

(J.M.Yi.j

Religion, Study of

The study of religion, involving theological, historical, philosophical, literary, anthropological, sociological, psychological, and phenomenological approaches to the subject, has been a matter of scholarly concern from the time of the ancient Greeks until the present, though the subject has been of increasing importance in the second part of the 20th century.

This article is divided into the following sections:

I. Nature and significance

The history of mankind has shown the pervasive influences of religion, and thus the study of religion, involving the attempt to understand its significance and origins, has become increasingly important in modern times. The 19th century saw the rise of the study of religion in the modern sense, in which the techniques of historical enquiry, the philological sciences, psychology, anthropology, sociology, and other human studies were brought to bear on the task of estimating the history, origins, and functions of religion. Rarely, however, has there been unanimity among scholars about the nature of the subject, partly because assumptions about the revealed nature of the Christian (or other) religion or about the falsity of religion become entangled with questions about the historical and other facts of religion. Thus, the subject has, throughout its history, contained elements of controversy.

THE ESSENCE OF RELIGION AND THE CONTEXT OF RELIGIOUS BELIEFS, PRACTICES, AND INSTITUTIONS

Attempts to arrive at a definition of religion

An acceptable definition of religion itself is difficult to attain. Attempts have been made to find an essential ingredient in all religions (*e.g.*, the numinous, or spiritual, experience; the contrast between the sacred and the profane; belief in gods or in God), so that an "essence" of religion can be described. But objections have been brought against such attempts, either because the rich variety of men's religions makes it possible to find counterexamples or because the element cited as essential is in some religions peripheral. The gods play a very subsidiary role, for example, in most phases of Theravāda ("Way of the Elders") Buddhism. A more promising method would seem to be that of exhibiting aspects of religion that are *typical* of religions, though they may not be universal. The occurrence of the rituals of worship is typical, but there are cases, however, in which such rituals are not central. Thus, one of the tasks of a student of religion is to gather together an inventory of types of religious phenomena.

The fact that there is dispute over the possibility of finding an essence of religion means that there is likewise a problem about speaking of the study of religion or of religions, for it is misleading to think of religion as something that "runs through" religions. This brings to light one of the major questions of method in the study of the subject. In practice, a religion is a particular system, or a set of systems, in which doctrines, myths, rituals, sentiments, institutions, and other similar elements are interconnected. Thus, in order to understand a given belief that occurs in such a system, it is necessary to look at its particular context—that is, other beliefs held in the system, rituals, and other aspects. Belief in the lordship of Christ in the early Christian Church, for example, has to be seen in the context of a belief in the Creator and of the sacramental life of the community. This systematic character of a religion has been referred to by the 20th-century Dutch theologian Hendrik Kraemer as "totalitarian"; but a better term would be "organic." Thus, there arises the problem of whether or not one belief or practice embedded in an organic system can properly be compared to a similar item in another organic system. To put the matter in another way, every religion has its unique properties, and attempts to make interreligious comparisons may hide these unique aspects. Most students of religion agree, however, that valid comparisons are possible, though they are difficult to make. Indeed, one can see the very uniqueness of a religion through comparison, which includes a contrast. The importance of setting religions side by side is why the study of religions is sometimes referred to as the "comparative study of religion"—though a number of scholars prefer not to use this phrase, partly because some comparative work in the past has incorporated value judgments about other religions.

Uniqueness and similarities in religions

But even if an inventory of types of belief and practices can be gathered—so as to provide a typical profile of what counts as religion—the absence of a tight definition means that there will always be a number of cases about which it is difficult to decide. Thus, some ideologies, such as Soviet Marxism, Maoism, and Fascism, may have analogies to religion. Certain attempts at an essentialist definition of religion, such as that of the German-American theologian Paul Tillich (1886–1965), who defined religion in terms of man's ultimate concern, would leave the way open to count these ideologies as proper objects of the study of religion. Tillich, incidentally, calls them "quasi-religions." Though there is no consensus on this point among scholars, it is not unreasonable to hold that the frontier between traditional religions and modern ideologies represents one part of the field to be studied.

NEUTRALITY AND SUBJECTIVITY IN THE STUDY OF RELIGION

Discussion about religion has been complicated further by the attempt of some Christian theologians, notably Karl Barth (1886–1968), to draw a distinction between the Gospel (the proclamation peculiar to Christianity) and religion. This distinction depends, to some extent, upon taking a projectionist view of religion as a human product. This tradition goes back in modern times to the seminal work of the German philosopher Ludwig Feuerbach (1804–72), who proposed that God was the extension of human aspirations, and is found in the work of Karl Marx, Sigmund Freud, and others. The distinction attempts to draw a line between the transcendent, as it reveals itself to men, and religion, as a human product involved in the response to revelation. The difficulty of the distinction consists chiefly in a denial that God (*e.g.*, Yahweh or Christ) as the object of man's response is a "religious" being (*i.e.*, God is transcendent, not "religious" in the sense of being a part of the human product), and thus the question about revelation as a religious fact needs to be answered. This account of religion, however, incorporates a theory about it, which is characteristic of a number of definitions of religion and creates a difficulty in that the field—namely, the study of religion—is being defined in terms of a theory within it.

Subjectivity in the study of religion. There are, however, doubts about how far there can be neutrality and objectivity in the study of religion. Is it possible indeed to understand a faith without holding it? If it is not possible, then cross-religious comparisons would mostly break down, for normally it is not possible to be inside more than one religion. But it is necessary to be clear about what objectivity and subjectivity in religion means. Religion can be said to be subjective in at least two senses. First, the practice of religion involves inner experiences and sentiments, such as feelings of the gracious presence of God in the heart, guiding the life of the devotee. Here religion involves subjectivity in the sense of individual experience. Second, religion may be thought to be subjective because the criteria by which its truth is decided are obscure and hard to come by, so that there is no obvious "objective" test in the way in which there is for a large range of empirical claims, such as that there are mountains on the moon. As to the first sense, one of the challenges to the student of religion is the problem of evoking its inner, individual side, which is not observable in any straightforward way. In considering a religion, however, the scholar is not only conceined with individual responses but also with communal ones. In any case, very often he is confronted only with texts describing beliefs and stories, so that he needs to infer the inner sentiments that these both evoke and express. The adherent of a faith is no doubt authoritative as to his own experience, but he is not necessarily so in regard to the communal significance of the rites and institutions in which he participates. Thus, the matter of coming to understand the inner side of a religion involves a dialectic between participant observation and dialogical (interpersonal) relationship with the adherents of the other faith. Consequently, the study of religion has strong similarities to, and indeed overlaps with, anthropology. General agreement upon scholarly methods, however, does not exist, partly because different scholars have come to the study of religion from different disciplines and points of view—such as history, theology, philosophy of religion, and sociology.

Study of individual and communal responses

The other sense of the subjectivity of religion is properly a matter for philosophy of religion and theology (Christian and otherwise). The study of religion can roughly be divided between descriptive and historical enquiries, on the one hand, and normative inquiries, on the other. The latter primarily concern the truth of religious claims, the acceptability of religious values, and other such normative aspects; the former, only indirectly involved with the normative elements of religion, are primarily concerned with its history, structure, and similar descriptive elements. The distinction, however, is not an absolute one, for, as has been noted, descriptions of religion may sometimes incorporate theories about religion that imply something about the truth or other normative aspects of some or all religions. Conversely, theological claims may imply something about the history of a religion. The dominant sense in which one speaks nowadays of the study of religion is the descriptive sense.

Neutrality in the study of religion. The attempt to be descriptive about religious beliefs and practices, without judging them to be valuable or otherwise, is often considered to involve *epochē*—that is, the suspension of belief and the "bracketing" of the phenomena under investigation. The idea of *epochē* is borrowed from the philosophy of the German thinker Edmund Husserl (1859–1938), the father of Phenomenology, and the procedure is regarded as central to the phenomenology of religion.

Phenomenology as a method attempting objectivity

In this context the term phenomenology refers first to the attempt to describe religious phenomena in a way that brings out the beliefs and attitudes of the adherents of the religion under investigation, but without either endorsing or rejecting these beliefs and attitudes. Thus, the bracketing means forgetting about one's own beliefs that might endorse or conflict with what is being investigated. Second, phenomenology of religion refers to the attempt to devise a typology of religious phenomena—to classify religious activities, beliefs, and institutions.

To some extent the emphasis on neutral description arises in modern times as a reaction against "committed" accounts of religion, which were for long the norm and still exist where religion is treated from a theological point of view. The Christian theologian, for example, may see a particular historical process as providential or as providing significance for Christian living. This is a legitimate perspective from the standpoint of faith. But the historical process itself has to be investigated in the first instance "scientifically"—that is, by considering the evidence, using the techniques of historical enquiry and other scientific methods. Conflict sometimes arises because the committed point of view is likely to begin from a more conservative stance—*i.e.*, to accept at face value the scriptural accounts of events—whereas the "secular" historian may be more skeptical, especially of records of miraculous events. The study of religion may thus come to have a reflexive effect on religion itself, such as the manner in which modem Christian theology has been profoundly affected by the whole question of the historicity of the New Testament.

The reflexive effect of the study of religion on religion itself may in practice make it more difficult for the student of religion to adopt the detachment implied by bracketing. Scholars generally agree, however, that the pursuit of objectivity is desirable, provided this does not involve sacrificing a sense of the inner aspect of religion. Thus, the stress on the distinction between the descriptive and normative approaches is becoming more frequent among scholars of religion.

The study of religion may thus be characterized as concerned with man's religious behaviour in relation to the transcendent, to God or the gods, and whatever else is regarded as sacred or holy, and as a study that attempts to be faithful both to the outer and inner facts. Its present-day concern is predominantly descriptive and explanatory and hence embraces such various disciplines as history, sociology, anthropology, psychology, and archaeology. Traditionally, however, the study has been more oriented toward truth claims in religion—these being properly the concern of theology and philosophy of religion. Needless to say, there are different sorts of theology, related to the different religious traditions, such as Christian, Muslim, and Buddhist. But insofar as the theologian expresses and articulates a tradition, he belongs to it and thus is part of the subject matter studied by the student of religion.

II. History of the study of religion

No single history of the study of religion exists since the major cultural traditions (Europe, the Middle East, India, China) have been mutually independent over long periods. The main impulse to study religion, however, happens to be the Western one, especially because other cultural traditions utilized categories other than that denoted in the Western concept of "religion." On the whole, in the ancient world and in the Middle Ages the various approaches to religion grew out of attempts to criticize or defend particular systems and to interpret religion in harmony with changes in knowledge. The same is true of part of the modern period, but increasingly the idea of the nonnormative (descriptive-explanatory) study of religions, and at the same time the attempt to understand the genesis and function of religion, has become established. Viewed thus, the 19th century is the formative period for the modern study of religion. The ensuing accounts here of the history of the subject take it up to the modern period and then consider the various disciplines connected with religion in detail—*i.e.,* in relation to their development since the 19th century.

Western orientation to the study of religion

THE GRECO-ROMAN PERIOD

Early attempts to study religion. An early attempt to systematize the seemingly conflicting Greek myths and thus to bring order into the Greek tradition was the *Theogony* of the Greek poet Hesiod (flourished c. 800 BC), who rather laboriously put together the genealogies of the gods. His work remains an important source book of ancient myth. The rise of speculative philosophy among the Ionian philosophers (*e.g.*, Thales, Heracleitus, and Anaximander) led to a more critical and, to some extent, rationalistic treatment of the gods. Thus, Thales (6th century BC) and Heracleitus (flourished c. 500 BC) considered water and fire, respectively, to be the first substance, out of which everything is made, though Aristotle reported mysteriously in the 4th century BC that Thales believed that everything was full of the gods. Anaximander (6th century BC) called the primary substance the infinite (apeiron). In these various schemes of belief, there is a unitary something that transcends the many clashing forces in the world, including the gods. Heracleitus refers to the controlling principle as logos, or reason, though the philosopher, poet, and religious reformer Xenophanes (6th–5th centuries BC) directly assaulted the traditional mythology as immoral, out of his concern to express a monotheistic religion. This theme of criticism of the myths was taken over and elaborated in the 4th century BC by the philosopher Plato. More conservatively, the poet Theagenes (6th century BC) allegorized the gods, treating them as standing for natural and psychological forces. To some extent, this line was pursued in the works of the Greek tragedians and by the philosophers Parmenides and Empedocles (5th century BC). Criticism of the ancient Greek tradition was reinforced by the reports of travellers as Greek culture penetrated widely into various other cultures. The Greek historian Herodotus (5th century BC) attempted to solve the problem of the plurality of cults by identifying foreign deities with Greek deities (*e.g.*, those of the Egyptian Amon with Zeus). This kind of syncretism was widely used in the merging of Greek and Roman culture in the Roman Empire (*e.g.*, Zeus as the Roman god Jupiter).

Skepticism in the study of religion

The plurality of cults and gods also induced skepticism, as with the Sophist Protagoras (c. 481–411 BC), who was driven from Athens because of his impiety in questioning the existence of the gods. Prodicus of Ceos (5th century BC) gave a rationalistic explanation of the origin of deities that foreshadowed Euhemerism (see below Later attempts to *study* religion), and another Sophist, Critias (5th century BC), considered that religion was invented to frighten men into adhering to morality and justice. Plato was not averse to providing new myths to perform this alleged function—as is seen in his conception of the "noble lie" (*i.e.*, the invention of myths to promote morality and order) in *The* Republic. He was strongly critical, however, of the older poets' (*e.g.*, Homer's) accounts of the gods and substituted a form of belief in a single creator, the Demiurge or supreme craftsman. This line of thought was developed in a stronger way by Aristotle, in his conception of a supreme intelligence that is the unmoved mover. Aristotle combined elements of earlier thinking in his account of the genesis of the gods (coming from the observation of cosmic order and stellar beauty and from dreams).

Later attempts to study religion. Later Greek thinkers tended to vary between the positions adumbrated in the earlier period. The Stoics (philosophers of nature and morality) opted for a form of naturalistic monotheism, while the philosopher Epicurus (341–270 BC) was skeptical of religion as ordinarily understood and practiced, though he did not deny that there were gods who, however, had no transactions with men. Of considerable influence was Euhemerus (c. 330–c. 260 BC), who gave his name to the doctrine called Euhemerism, namely, that the gods are divinized men. Though Euhemerus' own argument was based largely upon fantasy, there are certainly some examples, both in Greek religion (*e.g.*, the god Heracles) and elsewhere, of the tendency to make men into gods, but it is obviously not universal.

Most of the Greek concepts about religion proved to be influential in the Roman world also. The atheistic Atomism of the Roman natural historian Lucretius (*c.* 95–55 BC) owed much to Epicurus. The eclectic thinker and politician Cicero (106–43 BC), in his De *natura deorum* ("Concerning the Nature of the Gods"), criticized Stoic, Epicurean, and later Platonic ideas about religion, but the book remains, however, incomplete. Much of the skepticism about the gods in the ancient world was concerned with the older traditional religions, whether of Greece or Rome. But in the early empire, the mystery cults, ranging fom 'the Eleusinian mysteries of Greece to those of the Anatolian Cybele and the Persian Mithra, together with philosophically based religions such as Neoplatonism and Stoicism, had the greatest vitality. The patterns of religious belief were complex and of different levels, with various types of religion existing side by side. Into this situation Christianity was injected, and in its encounter with classical civilization it absorbed a number of the critiques of the gods of the older thinkers. In particular, Euhemerism was fashionable among the Church Fathers (the religious teachers of the early church) as an account of paganism. On the "pagan" side, there were persistent attempts to justify the popular cults and myths by the extensive use of allegory—a technique well adapted to the attempt to synthesize philosophical and popular religion. Christianity's own contribution to theories of the genesis of polytheism was through the doctrine of the fall of man, in which pure monotheism was believed to have become overlaid by demonic cults of the gods. Such an account could help to explain some underlying similarities between the Judeo-Christian tradition and certain aspects of Greco-Roman paganism. In this view there is the germ of an evolutionary account of religion. On the whole, however, the theories of religion in the ancient world were naturalistic and rationalist in emphasis.

THE MIDDLE ACES TO THE REFORMATION

Theories of the Middle Ages. The spread of Christianity into northern Europe and elsewhere beyond the confines of the Roman Empire presented similar problems to those encountered in the pagan world. Similar solutions were offered—*e.g.*, the identification of northern and Roman and Greek gods, sometimes using etymologies owing much to superficial resemblances of names. Thus, the Icelandic historian Snorri Sturluson (1179–1241) made use of this method in his handbook of Icelandic mythology—a work necessitated by the need to pass on the myth-laden Norse poetic lore that had survived the Christianization of the north—by adding to it Euhemeristic elements.

European contacts with other religions

Meanwhile, Islāmic theology had had an impact on Western Christianity, notably upon medieval Scholastic philosophy, in which the values of both reason and revelation were maintained. Muslim knowledge of other religions was in advance of European knowledge, notably in the work of the theologian Ibn Hazm (994–1064). Nevertheless, the reports of some European travellers, such as the Italian Marco Polo (1254?–?1324) and also Odoric of Pordenone (14th century), gave Westerners some knowledge of Asian religions. This opened the way toward a more inductive treatment of the phenomena of other religions, based on factual knowledge. Though most Christian, as well as Islamic and Jewish, theologians tended to consider the question of whether or not natural religion gives insight in God's nature—treating religion as a relation to the first cause of the universe—the English philosopher Roger Bacon (c. 1220–c. 1292) preferred to categorize the various manifest types of religion as a preliminary effort to establishing a true theology. Theorists of the medieval period continued to accept the thesis that polytheism had its origin in the Fall of man, but two new theories modified attitudes of Christians to other faiths. First, the theory arose that God adapts customs and rites having a pagan style in order to combat paganism itself—as a concession to the human condition. This theory could be used to explain the divergencies of practice within Christendom and to show points of contact between Christianity and paganism. Second, the doctrine of man's innate capacity to know God by reason enabled thinkers to discern some measure of truth in other religions. The questions raised by such theories were intensified during the Renaissance.

Theories of the Renaissance and Reformation. The Renaissance consisted in the invigoration of European culture through the rediscovery of the Greek and Roman classics, art, and architecture and thus was bound to set up tensions among Christians about paganism. The Italian Humanist Giovanni Boccaccio (1313–75) attempted to resolve these tensions in a medieval way by extensively allegorizing the ancient myths. The Dutch Humanist Erasmus (c. 1466–1536) and others, however, went further in stating that the ancient thinkers had a direct knowledge of the highest truth and sometimes in comparing them favourably with Scholastic theologians. One of the interlocutors in his *Convivium* Religiosum suggests that it would be better to lose the Scholastic theologian Duns Scotus than the ancient Roman thinkers Cicero or Plutarch, while another speaker restrains himself with difficulty from praying to the Greek philosopher Socrates (c. 470–399 BC) as though he were a Catholic saint. But a new turn to the arguments about idolatry, which were essentially apologetic, was given by the Protestant Reformers' attack on idolatry within the Roman Catholic Church and by their comparison between what they took to be the Christianity of the New Testament and the religion of Rome.

The need for a comparative study of religion

Thus, the need for a comparative treatment of religion became clear, and this prepared the way for more modern developments. Also preparatory for the modern study of religion was the new trend toward more or less systematic compilations of mythological and other material, stimulated partly by the Renaissance itself and partly by the discovery of America and other lands—conveying to the inhabitants of Europe a new perspective on the richness and variety of man's customs and histories. The most important figures in the exploration of the religions of the non-European world were the Spanish monk Bernardino de Sahagún (c. 1499–1590), who conscientiously gathered information in New Spain, J. Lafitau (1685–1740), a French missionary in Canada, and the Italian Jesuits Roberto de Nobili (1577–1656) and Matteo Ricci (1552–1610). The last two, who brought to bear a deep understanding of Indian and Chinese cultures, were unparalleled in that area of study until modem times. Thus, some of de Nobili's discussions with Brahmins were probably the first profound dialogues between Hindus and Christians. The inquiries of the 16th to 18th centuries thus initiated an accumulation of data about other cultures that stimulated studies of other men's religions and went beyond apologetic concerns, which hitherto had been dominant.

THE BEGINNINGS OF THE MODERN PERIOD

The late 17th and 18th centuries. Attempts at a developmental account of religion were begun in the late 17th and 18th centuries. Notable was the scheme worked out, though not in great detail, by the Italian philosopher Giambattista Vico (1668-1774), who suggested that Greek religion passed through various stages: the divinization of nature, then of those powers that man had come to control (such as fire and crops), then of institutions (such as marriage), and finally the process of humanizing the gods, as in the works of Homer. The English philosopher David Hume (1711–76) gave another account in his Natural History of Religion, which reflected the growing Rationalism of the epoch. For Hume, original polytheism was the result of a naïve anthropomorphism (conceiving the divine in human form) in the assignment of causes to natural events. The intensification of propitiatory and other forms of worship, he believed, led to the exaltation of one infinite divine Being. His "Essay upon Miracles" was also important in posing vital questions about the historical treatment of sacred texts, a set of problems that was to preoccupy 19th- and 20th-century Christian theologians.

The Rationalism of the period often involved a rejection both of paganism and dogmatic Christianity in the name of "natural religion." This natural religion, also called Deism, was the intellectual counterpart of the more emotional antidogmatic faith of the Pietists, who advocated "heart religion" over "head religion." Among the French Philosophes and Encyclopaedists, Voltaire (1694–1778) espoused an anticlerical Deism, which viewed the genesis of polytheism in the work of priests—a point also developed by another Encyclopaedist, Denis Diderot (1713–1784). Voltaire was, incidentally, somewhat influenced and impressed by reports of the ethics of the Chinese social and religious sage Confucius (6th century BC).

The culmination of 18th-century Rationalism was found in the works of the German philosopher Immanuel Kant (1724–1804), but it was a rationalism modified to leave room for religion, which he based essentially on ethics. He held that all men in their awareness of the categorical imperative (*i.e.*, the notion that one must act as though what one does can become the universal law for mankind) and reverence for it share in the one religion and that the pre-eminence of Christianity lay in the conspicuous way in which Jesus enshrined the moral ideal. A series of reactions against the highly influential Kantian account paved the way for the various approaches to religion in the 19th century. In the meantime, the first beginnings of the development of Oriental studies and of ethnology and anthropology were making available more data about religion, though discussion in the 18th century continued, as in earlier centuries, the concern for the problems of religions other than those of the Judeo-Christian tradition largely in terms of the paganism of the ancient world. In this connection, the French scholar and politician Charles de Brosses (1709–77) attempted to explain Greek polytheism partly through the fetishism (belief in the magical powers of certain objects) found in West Africa. This foreshadowed later attempts to use comparative material in the elucidation of Greek myths. The French abbé Bergier (1718–90) explained primitive religions by means of a belief in spirits arising from a variety of psychological causes, which thus was a precursor of animism (a belief in souls in persons or certain natural objects).

One of the critics of Kant's view of religion was the German philosopher Johann Gottfried von Herder (1744–1803), who adopted an evolutionary account of the human race and who saw in mythology something much deeper and more significant for the understanding of human language and thought than a record of follies. His concern with symbolic thinking makes him the first modern student of myth. The German philosopher Friedrich Schelling (1775–1854) continued this positive approach, in the tradition of Romanticism. Furthermore, the advances in the knowledge of non-European, especially Indian, religion gave a wider perspective to discussions of the nature of religion, as was clear in the work of the German philosopher G.W.F. Hegel (1770–1831). The latter's self-confidence, in supposing that his philosophy represented the culmination of the history of philosophy, may amuse present-day scholars in view of the fact that many changes have occurred in philosophical enquiry since his day. Hegel was, nevertheless, immensely influential over a wide range of scholarship, including the study of religion. His followers were in large measure the founders of modern scientific history. Admittedly his theory of the historical dialectic—in which one movement (the thesis) is countered by another (the antithesis), both in interplay giving rise to a third (the synthesis), which now becomes the thesis of a new dialectical interplay, and so on—has been viewed as too artificial. But in providing a theoretical skeleton, it inspired attempts to make sense of the multitude of historical data, so that scholars were driven to the investigation and discovery of particular facts that might exhibit the universal patterns postulated. Hegel also had a modified relativism, which implied that each phase of religion has a limited truth. This, together with his dialectic scheme, led to a general theory of religions, which though dated, much too neat, and based on imperfect information, nevertheless represents an important attempt at a comparative treatment, and one that was evolutionary or developmental.

Beginnings in the modern study of myth and evolutionary theories

The early 19th century. Hegel, as an Idealist, stressed the formative power of the spiritual on human history. By contrast, the French social philosopher Auguste Comte (1798–1857), from a Positivistic and Materialist point of view, devised a different evolutionary scheme in which there are three stages of human history: the theological, in which the supernatural is important; the metaphysical, in which the explanatory concepts become more abstract; and the positivistic—*i.e.*, the empirical. A rather different Positivism was expressed by the English philosopher Herbert Spencer (1820–1903), in which religion has a place beside science in attempting to refer to the unknown, and unknowable, Absolute. Evolutionary accounts were much boosted in the latter part of the 19th century by the new theory of biological evolution and had a marked effect both on history of religions and anthropology.

Meanwhile, the German philosopher Ludwig Feuerbach (1804–72) propounded, in his Lectures on the Essence of Religion, a view of religion as a projection of the aspirations of men, a thesis that was to be taken up in various ways by, among others, Marx, Freud, and Barth.

These various movements were supplemented by the growth of scientific history, archaeology, anthropology, and other sciences, which increased comparative knowledge of civilizations and cultures. The major figures and trends in the relevant disciplines are dealt with below.

Though the 19th-century theories that form the starting point of the modern study of religion were often based directly on metaphysical schemes in competition with Christian and other theologies, there was a notably different atmosphere in comparison with preceding periods, and the stage was set for a more complex and mutual attempt to understand the history and nature of religion.

III. Basic aims and methods

The growth of various disciplines in the 19th century, notably psychology and sociology, stimulated a more analytic approach to religions, while at the same time theology became more sophisticated and, in a sense, scientific as it began to be affected by and thus to make use of historical and other methods. The interrelations of the various disciplines in relation to religion as an area of study can be described as follows.

The various dimensions of religion

Religions. being complex, have different aspects or dimensions. Thus, the major world religions typically possess doctrines, myths, ethical and social teachings, rituals, social institutions, and inner experiences and sentiments. These dimensions lie behind the creation of buildings, art, music, and other such extensions of basic beliefs and attitudes. But not all religions are like Christianity and Buddhism, for example, in possessing institutions such as the church and the *saṅgha* (Buddhist monastic order), which exist across national and cultural boundaries. In opposition to such institutionalized religions, tribal religion, for example, is not usually separately institutionalized but in effect is the religious side of communal life and is not treated as distinct from other things that go on in the community.

The various dimensions of religion noted above represent a cross section of a tradition; but to see the latter in a well-balanced perspective it is necessary to view it as historical—as a religion having a past and the capacity for development in the future ("dead" religions, obviously enough, being the exception). Thus, there are various disciplines that may examine a religion cross-sectionally to find its basic patterns or structures. Psychology views religious experience and feelings and to some extent the myths and symbols that express experience; sociology and social anthropology view the institutions of religious tradition and their relationship to its beliefs and values; and literary and other studies seek to elicit the meanings of myths and other items. These structural enquiries sometimes benefit from being comparative—as when recurrent motifs in the doctrines of different religions are noticed. On the other hand, the aforementioned disciplines need to be supplemented by history, archaeology, philology, and other such disciplines, which have their own various methods of elucidating the past. Philosophy generally has attempted wide-ranging accounts of the nature of religion and of religious concepts, but it is not always easy to disentangle these enquiries from issues raised by normative theology.

HISTORICAL, ARCHAEOLOGICAL, AND LITERARY STUDIES

Historical and literary studies. The expansion of European empires in the early 19th century and the growth of scientific methods in history and philology combined to place Oriental and other non-European studies on a new basis. Another stimulus to the new approach to history and philology was Napoleon's expedition to Egypt, which was accompanied by scholars and scientists; it was a notable attempt to gather knowledge of a culture systematically. The discovery and editing of sacred and other texts from other cultures also had profound effects upon European thinking. A notable publishing venture was the series Sacred Books of *the* East, edited under the leadership of the German Orientalist and philologist Max Miiller (1823–1900), which placed at the disposal of Westerners translations of the major literary sources of the non-Christian world. Earlier, Muller had published translations of the more important Vedic texts (Hindu sacred works), of which the Rgveda was given a complete scholarly edition in 1861–77. Interest in these ancient Indian texts was intense among Europeans and Americans in that earlier reports had suggested that these represented a world outlook from the "dawn of humanity" and that the origin of polytheism lay in nature worship. The Vedas, however, turned out to be of a very different character. The length of human history and prehistory, as implied by evolutionary theory and the growing archaeological discoveries, precluded looking upon the Vedic hymns as anything but late; though the contents showed them to be highly artificial and complex compilations for use in a priest-dominated ritual context, they were not at that time seen as spontaneous outpourings of the human spirit. Muller himself reacted rather sharply by adopting a different theory, which expressed his philological slant—namely, that polytheism was the result of a disease of language, in which the terms for natural phenomena came to be treated as having independent and personal reality: *nomina* ("names") became *numina* ("spirits"). The theory was in vogue for a time but was later replaced by more realistic insights drawn from anthropology. Furthermore, study of the greater part of the corpus of Indian sacred writings, including those in vernacular languages (especially Tamil), gradually modified the preoccupation with the earliest texts—the Vedic hymns and the *Upaniṣads* (philosophical treatises).

Discovery and publication of non-Western sacred writings

Throughout the development of the study of non-European languages there was a supposition that a non-Christian equivalent of the Bible could be found, a sacred writing that would thus provide the authoritative key to the beliefs, practices, and institutions of the religion under consideration. Gradually, however, it became apparent that sacred scriptures play very different roles in different religious cultures. Somewhat later in developing were studies of the Buddhist canon in Pāli (an ancient Indian language), which, through the work of such scholars as the English Orientalist T.W. Rhys Davids (1843–1922) and of the Pāli Text Society, which he founded, had a remarkable impact in revealing to the West the full range of Theravādin (southern Buddhist) religious literature; it tended to make Western scholars look upon the Theravāda as the earlier, "purer" form of Buddhism; but the editing of early Mahāyāna ("Greater Vehicle," or northern Buddhist) texts and the recognition of the different strata in the PBli canon have modified this view. Buddhist studies were enhanced by the growth of Tibetan, Chinese, and Japanese studies. Some of the more important modern scholars of Zen Buddhism (a Mahāyāna sect) have been Japanese, notably the philosopher D.T. Suzuki (1870–1966), sometimes called the apostle of Zen Buddhism to America, whose editions and interpretations have been widely influential.

The productivity of the study of religious literature of the late 19th century was immense, for it was not confined to the foregoing literary and archaeological activities but to the investigation of the Chinese Classics and the roots of Chinese civilization as well. Thus, by the early 20th century, Western scholars were in a position to study the main range of non-Western literary cultures. The wave of interest in these texts and the freeing of their dissemination from some of their traditional constraints (*e.g.*, the restriction of Vedic revelation to the upper classes of the Indian caste system) contributed to the revival of other

Interest in textual studies

religious cultures—notably Hinduism and Buddhism, under the stimulus of the Western challenge. Modern scholarship thus provided the basis for a new self-understanding among such religious traditions.

Meanwhile, the texts of Zoroastrianism, an Iranian religion originating in the 6th century BC, were being discovered and edited (from 1850 onward). The disentangling of different layers of varying antiquity indicated the complex ways in which the religion of Zoroaster had developed.

During the latter part of the 19th and early part of the 20th century, there was a remarkable flowering of ancient Near Eastern studies. Archaeology contributed to the unravelling of non-Jewish and Jewish religious history. The discovery of the *Epic of Gilgamesh,* a major work of Mesopotamian religious literature, and other materials brought a whole new perspective to the development of ideas in Mesopotamia; and in Egypt archaeological and papyrological studies brought to light the famous and revealing Egyptian funerary text, the Book of the Dead. These various ancient Near Eastern discoveries have thrown light on the evolution of Judaism, and Semitic studies have likewise illuminated the origins and background of Islām. Furthermore, classical and European studies assembled data about the pre-Christian religions of the West so that scholars might gain a more detailed and scientific understanding of them. Compilations such as the *Corpus Inscriptionum Graecorum* and the *Corpus Inscriptionum Latinarum,* assembled in the mid-19th century, and the publication of Germanic, Celtic, and Scandinavian texts provided the tools for a reappraisal of these older traditions. Throughout the period intense researches into the composition and milieu of the Old and New Testaments reflected a new and "scientific" spirit of enquiry—which was, however, not without its controversial elements, sometimes because of the intimate tie between religious positions and evaluations of the Bible and sometimes because of the application of speculative patterns in the history of (non-Christian) religions to the New Testament. Meanwhile, the assemblage of materials extended forward into Christian history through the application of classical philological methods to patristic texts (the writing of the early Church Fathers) and to the corpus of Reformation writings.

The significance of 19th-century archaeological discoveries

Archaeological studies. The great archaeological discoveries of Heinrich Schliemann, the German excavator of Troy; the English archaeologists Arthur Evans in Crete and Wm. M. Flinders Petrie in Egypt; the French archaeologist Jacques de Morgan in Elam; the German Orientalist Hugo Winckler in Boğazköy (Anatolia); the French archaeologists Claude Schaeffer and C. Virolleaud in Ras Shamra (Ugarit); and other archaeologists greatly enlightened modem knowledge of the Greco-Roman and ancient Near Eastern worlds. Biblical archaeology, culminating perhaps in the discovery of Masada, the Judaean hill fortress where the Jews made their last stand against the Romans in the revolt of AD 66–73 and that was mainly excavated in 1963, has given a new perspective to Old Testament, intertestamental, and later studies of ancient Judaism. The spectacular discovery by the English archaeologist John Marshall and others of the Indus Valley civilization pushed back knowledge of Indian prehistory to about 3500 BC and called into question the earlier theory of the primacy of Vedic culture in the formation of the Indian tradition, many features of which appear to have their first manifestation in the Indus Valley cities.

Archaeology made another profound impact on the study of religion when in 1841 the discovery of prehistoric human artifacts and later finds gave clues to early man's magico-religious beliefs and practices. These discoveries, notably the cave paintings in the Dordogne, northern and eastern Spain, and elsewhere, gave scholars encouragement to work out the course of man's religious evolution from earliest times. Spectacular as prehistoric archaeology was proving to be, however, it could only yield fragments of a whole that is difficult to reconstruct. Even the famous cave paintings of Trois Frères, in the Dordogne, for example, which portray among other things a dancing human with antlers on his head and a stallion's tail decorating his rear, does not yield an unambiguous interpretation: is the dancing figure a sorcerer, a priest, or what? He very likely is a priest presenting himself as a divine figure connected with animal fertility and hunting rites—but this remains as only an educated guess. Hence, it became attractive to many scholars of religion to try to supplement ancient archaeological evidence with data drawn from contemporary primitive peoples—*i.e.*, to interpret the prehistoric Stone Age through present-day stone age cultures. This procedure has several pitfalls—partly because contemporary "primitives" are themselves the product of a long historical process and because their culture may have changed over the millennia in many and various ways.

The work of the archaeologists has not merely stimulated new thinking about the early stages of religious history but it has also been a factor in drawing attention to the roles of buildings and art objects in religion. During the present century, spectacular religious monuments of the past, such as Angkor Wat (Cambodia), Borobodur (Indonesia), Ellora and Ajantā (India), and the Acropolis (Athens), have been officially preserved for scholarly and public viewing. Though iconography (the study of content and meaning in visual arts) has been better developed among art historians, students of religion are now paying increased attention to the religious decipherment of the visual arts. By contrast, very little has been done in the sphere of music, despite the considerable role it plays in so many religions. This is a further way in which the study of texts and ideas needs to be supplemented by knowledge of the milieu in which they have their meaning.

ANTHROPOLOGICAL APPROACHES TO THE STUDY OF RELIGION

Theories concerning the origins of religion. To draw a clear line between anthropology and sociolonv is difficult, and the two disciplines are divided more–by tradition than by the scholarly methods they employ. Anthropology, however, has tended to be chiefly concerned with nonliterate and technologically primitive cultures and thus has stressed a certain range of techniques, such as the use of participant observation. Much anthropological investigation, however, has been carried out recently in more complex societies, such as in various Hindu areas of India, where there are different layers of society, ranging from an educated elite to illiterate workers who carry out the traditional menial tasks of the lowest castes and the outcastes. Because of the anthropologists' interest in tribal and "primitive" societies, it has not been unnatural for them to try to use the data gained in the study of such societies to speculate about the genesis and functions of religion.

Evolutionary theories about religion

An early attempt to combine archaeological evidence of prehistoric peoples, on the one hand, and anthropological evidence of primitive peoples, on the other, was that of the English anthropologist John Lubbock (1834–1913). His book, *The Origin of Civilization and the Primitive Condition of Man*, outlined an evolutionary scheme, beginning with atheism (the absence of religious ideas) and continuing with fetishism, nature worship, and totemism (a system of belief involving the relationship of specific animals to clans), shamanism (a system of belief centring on the shaman, a religious personage having curative and psychic powers), anthropomorphism, monotheism (belief in one god), and, finally, ethical monotheism. Lubbock recognized a point later made by the German theologian and philosopher Rudolf Otto (1869–1937) in distinguishing between the unique holiness (separateness) of God and his ethical characteristics. Unfortunately, much of his information was unreliable, and his schematism was open to question; he foreshadowed, nevertheless, other forms of evolutionism, which were to become popular both in sociology and anthropology. The English ethnologist E.B. Tylor (1832–1917), who is commonly considered the father of modem anthropology, expounded, in his book *Primitive Culture,* the thesis that animism is the earliest and most basic religious form. Out of this evolves fetishism, belief in demons, polytheism, and, finally, monotheism, which derives from the exaltation of a great god, such as the sky god, in a polytheistic context. A somewhat similar system was advanced by Herbert Spencer (1820–1903) in his *Principles of Sociology,* though he stresses ancestor worship rather than animism as the basic consideration.

The classifications of religion—polytheism, henotheism (*i.e.*, the worship of one god as supreme without necessarily excluding the possibility of other groups' gods), and monotheism—begin from concern with gods and often imply the superiority of monotheism over other forms of belief. Naturally, the anthropologists of the 19th century were deeply influenced by the presuppositions of Western society.

The English anthropologist R.R. Marett (1866–1943), in contrast to Tylor, viewed what he termed animatism as of basic importance. He took his clue from such ideas as mana, mulungu, orenda, and so on (concepts found in the Pacific, Africa, and America, respectively), referring to a supernatural power (a kind of supernatural "electricity") that does not necessarily have the personal connotation of animistic entities and that becomes especially present in certain men, spirits, or natural objects. Marett criticized Tylor for an overly intellectual approach, as though primitive men used personal forces as explanatory hypotheses to account for dreams, natural events, and other phenomena. For Marett, primitive religion is "not so much thought out as danced out," and its primary emotional attitude is not so much fear as awe (in this he is close to Otto, whom he influenced).

Theories of Sir James Frazer

Another important figure in the development of theories of religion was the English folklorist Sir James Frazer (1854–1941), in whose major work, *The Golden Bough,* is set forth a mass of evidence to establish the thesis that men must have begun with magic and progressed to religion and from that to science. He owes much to Tylor but places magic in a phase anterior to belief in supernatural powers that have to be propitiated—this belief being the core of religion. Because of the realization that magical rituals do not in fact work, primitive man then turns, according to Frazer, to reliance on supernatural beings outside his control, beings who need to be treated well if they are to cooperate with human purposes. With further scientific discoveries and theories, such as the mechanistic view of the operation of the universe, religious explanations gave way to scientific ones. Frazer's scheme is reminiscent of that of the French "father of sociology," Auguste Comte.

These and other evolutionary schemes came in for criticism, however, in the light of certain facts about the religions of primitive peoples. Thus, the Scottish folklorist Andrew Lang (1844–1912) discovered from anthropological reports that various primitive tribes believed in a High God—a creator and often legislator of the moral order. Marett and other anthropologists contended that Lang's attempt to argue for an *Urmonotheismus* (primordial monotheism) was contrary both to evolutionary ideas and to the established view of the lack of sophistication and half-animal status of the so-called savage. Since Lang was more of a brilliant journalist than an anthropologist, his view was not taken with as much seriousness as it should have been.

The German Roman Catholic priest and ethnologist Wilhelm Schmidt (1868–1954), however, brought anthropological expertise to bear in a series of investigations of such primitive societies as those of the Tierra del Fuegians (South America), the Negrillos of Rwanda (Africa), and the Andaman Islanders (Indian Ocean). The results were assembled in his *Der Ursprung der Gottesidee* ("The Origin of the Idea of God"), which appeared in 12 volumes from 1912 to 1955. Not surprisingly, Schmidt and his collaborators saw in the high gods, for whose cultural existence they produced ample evidence from a wide variety of unconnected societies, a sign of a primordial monotheistic revelation that later became overlaid with other elements (this was an echo of earlier Christian theories invoking the Fall to similar effect). The interpretation is controversial, but at least Lang and Schmidt produced grounds for rejecting the earlier rather naive theory of evolutionism.

Theories of primordial monotheism

Modern scholars do not, on the whole, accept Schmidt's scheme. Some, such as the Italian anthropologist Raffaele Pettazzoni (1883–1959), have stressed merely that a sky god has a certain natural pre-eminence; others emphasize that the high god is often a *deus otiosus* ("idle god")—*i.e.*, not active in the world and hence not the recipient of a functioning cult. In any event, it is a very long jump from the premise that primitive tribes have high gods to the conclusion that the earliest men were monotheists.

Others who have looked at religions from an anthropological point of view have emphasized the importance, in a number of cultures, of the mother goddess (as distinct from the male sky god). A pioneer work in this direction was that of the Swiss anthropologist and jurist J.J. Bachofen (1815–87), whose *Das Mutterrecht* ("The Mother Right") unravelled some puzzles in ancient law, mythology, and art in terms of a matriarchal society.

Functional and structural studies of religion. The search for a tidy account of the genesis of religion in prehistory by reference to primitive societies was hardly likely to yield decisive results. Thus, anthropologists became more concerned with functional and structural accounts of religion in society and relinquished the apparently futile search for origins.

Notable among these accounts was the theory of the French sociologist Émile Durkheim (1858–1917). According to Durkheim, totemism was fundamentally significant (he wrongly supposed it to be virtually universal), and in this he shared the view of some other 19th-century savants, notably Salomon Reinach (1858–1932) and Robertson Smith (1846–94), not to mention Sigmund Freud. Because Durkheim treated the totem as symbolic of the god, he inferred that the god is a personification of the clan. This conclusion, if generalized, suggested that all the objects of religious worship symbolize social relationships and, indeed, play an important role in the continuance of the social group.

Analyses of social patterns and institutions

Various forms of functionalism in anthropology—which understood social patterns and institutions in terms of their function in the larger cultural context—proved illuminating for religion, such as in the stimulus to discover interrelations between differing aspects of religion. The Polish-British anthropologist Bronisław Malinowski (1884–1942), for instance, emphasized in his work on the Trobriand Islanders (New Guinea) the close relationship between myth and ritual—a point also made emphatically by the "myth and ritual" school of the history of religions (see below *Other studies and emphases).* Furthermore, many anthropologists, notably Paul Radin (1883–1959), moved away from earlier categorizations of so-called primitive thought and pointed to the crucial role of creative individuals in the process of mythmaking.

A rather different approach to myths was made by the 20th-century French anthropologist Claude Levi-Strauss, whose rather formalistic structuralism tended to reinforce analogies between "primitive" and sophisticated thinking and also provided a new method of analyzing myths and stories. His views had wide influence, though they are by no means universally accepted by anthropologists.

Specialized studies. The impact of Western culture, including missionary Christianity, and technology upon a wide variety of primitive and tribal societies has had profound effects and represents a specialized area of study closely related to religious anthropology. One pioneering work is *Religions of the Oppressed* by the Italian anthropologist and historian of religion Vittorio Lanternari. What is striking is the way in which similar types of reaction, creating new religious movements, occur at different points across the world. There are, thus, many possibilities of a comparative treatment.

Among a number of contemporary anthropologists, including the American Clifford Geertz, there is a concern, after a period of functionalism, with exploring more deeply and concretely the symbolism of cultures. The English social anthropologist E.E. Evans-Pritchard (1902–73), noted among other things for his work on the religion of Nuer people (who live in The Sudan), produced in his *Theories* of Primitive Religion a penetrating critique of many of the earlier anthropological stances. Though it has always been difficult to confirm theories in view of the complexity of the data, a statistical approach has been attempted—*e.g.*, by G. Swanson (1922–) in his *Birth* of the Gods, which attempts to exhibit correlations between types of social arrangement and religious beliefs, such as the caste system and belief in reincarnation.

Because of the nature of the societies that typically have come under the scrutiny of anthropology, the discipline has necessarily had to come to terms with religion. In terms of the methods used, the anthropological approach is of considerable interest to historians of religion and is a corrective to overintellectual, text-based accounts of religions. Also, the present concerns for comparative studies and symbolic analysis coincide with existing concerns in the phenomenology of religion (see below History *and* phenomenology of religion).

SOCIOLOGICAL STUDIES OF RELIGION

Theories of stages. Auguste Comte (1798–1857) is usually considered the founder of modem sociology. His general theory hinged substantially on a particular view of religion, and this view has somewhat influenced the sociology of religion since that time. In his *Cours* de *philosophie* positive (The Positive Philosophy of Auguste *Comte)* Comte expounded a naturalistic Positivism and sketched out the following stages in the evolution of thought. First, there is what he called the theological stage, in which events are explained by reference to supernatural beings; next, there is the metaphysical stage, in which more abstract unseen forces are invoked; finally, in the positivistic stage, men seek causes in a scientific and practical manner. To seek for scientific laws governing human morality and society is as necessary, in this view, as to search for those in physics and biology—hence Comte's role in advocating a science of society, namely sociology. Among the leading figures in the development of sociological theories were Spencer and Durkheim (see above *Anthropological approaches to* the study of religion).

Marxist theories of religion

A rather separate tradition was created by the German economic theorist Karl Marx (1818–83). A number of Marxists, notably Lenin (1870–1924) and K. Kautsky (1854–1938), have developed social interpretations of religion based on the theory of the class struggle. Whereas sociological functionalists posited the existence in a society of some religion or a substitute for it (Comte, incidentally, propounded a positivistic religion, somewhat in the spirit of the French Revolution), the Marxists implied the disappearance of religion in a classless society. Thus, in their view religion in man's primordial communist condition, at the dawn of the historical dialectic, reflects ignorance of natural causes, which are explained animistically. The formation of classes leads, through alienation, to a projection of the need for liberation from this world into the transcendental or heavenly sphere. Religion, both consciously and unconsciously, thus becomes an instrument of exploitation. In the words of the young Marx, religion is "the generalized theory of the world . . . , its logic in popular form." The modem intellectualist accounts of religion, tending to ignore the rituals, experiences, and institutions but concentrating rather on the doctrines and myths, have proved something of a problem for later Marxist applications of their theory. Since the theory was a product of a rather early and unsophisticated stage of theorizing about religion, it was not adapted particularly well to deal with other cultures—hence a considerable debate in modem China on the status of Chinese religion in the light of Marxism, some holding that Marx's critique did not, for example, fit Buddhism.

Theories of Max Weber

Comparative studies. One of the most influential theoreticians of the sociology of religion was the German scholar Max Weber (1864–1920). He observed that there is an apparent connection between Protestantism and the rise of capitalism, and in The Protestant Ethic and the Spirit of Capitalism he accounted for the connection in terms of Calvinism's inculcating a this-worldly asceticism—which created a rational discipline and work ethic, together with a drive to accumulate savings that could be used for further investment. Weber noted, however, that such a thesis ought to be tested; and a major contribution of his thinking was his systematic exploration of other cultural traditions from a sociological point of view. He wrote influentially about Islām, Judaism, and Indian and Chinese religions and, in so doing, elaborated a set of categories, such as types of prophecy, the idea of charisma (spiritual power), routinization, and other categories, which became tools to deal with the comparative material; he was thus the real founder of comparative sociology. Because of his special interest in religion, he can also be reckoned a major figure in the comparative study of religion (though he is not usually reckoned so in most accounts of the history of religions). Though he made significant contributions to the study of religion, his judgments on Indian and other religions are not all or mostly accepted now—since he necessarily based his views on secondary sources—and some of his categorial distinctions are open to debate, such as his rather broad use of the category of prophet.

Weber's comparative method in the scientific sociology of religion introduced an analogue to experimentation (*i.e.*, looking at similar patterns in independent cultures with varying contextual conditions). Since the 1950s there has been considerable emphasis on statistical methods, side by side with the more theoretical discussions arising from classical sociology. Typical of the trend is the American sociologist Gerhard Lenski's Religious Factor, which delineates the relations between religious allegiance and other factors in a large city in the United States.

Other sociological studies. An extensive literature on religious sects and similar groups has also developed. To some extent this has been influenced by the German theologian Ernst Troeltsch (1865–1923) in his distinction between church and sect (see below Theological studies). Notable among modem investigators of sectarianism is the British scholar Bryan Wilson (1929–). Church organizations also have attempted to use the insights of sociology in the work of evangelism and other church-related activities—a use of the discipline that is sometimes called "religious sociology" to distinguish it from the more theoretical and "objective" sociology of religion.

Coordination between sociology and the history of religions is not usually very close, since the two disciplines operate as separate departments in most universities and often in different faculties. From the sociological end, Weber represents one kind of synthesis; from the history-of-religions end, the writings of the German-American scholar Joachim Wach (see below The "Chicago school") were quite influential. In his book Sociology of Religion he attempted to exhibit the ways in which the community institutions of religion express certain attitudes and experiences. This view was in accordance with his insistence on the practical and existential side of religion, over against the intellectualist tendency to treat the correlate of the group as being a system of beliefs.

Recent sociological studies

Among the more recent theorists of the sociology of religion is the influential and eclectic American scholar Peter Berger (1929–). In The Sacred Canopy he draws on elements from Marx, Durkheim, Weber, and others, creating a lively theoretical synthesis. One problem is raised by his method, however; he espouses what he calls "methodological atheism" in his work, which appears to presuppose a view about religion. Despite Berger's sympathy in dealing with religious phenomena, the methodological stance adopted in this book seems to imply a reductionist position—namely, one in which religious beliefs are explained by reference to basically nonreligious sentiments, sociopsychological circumstances, and other factors. In itself, this is a theory having possibilities, for the study of religion cannot rule out a priori the thesis that religion is a projection—*e.g.*, that it rests upon an illusion—or other such theses; but the question arises as to whether or not the methods espoused in the scientific study of religion have already secretly prejudged the is-

On the whole, modern sociology is largely geared to dealing with Western religious institutions and practices, though some notable work has been done, especially since World War II, in Asian sociology of religion. Emphasis has been placed upon the process of secularization in a number of Western sociological studies (which have had some impact on the formation of modern Christian theology), notably in *The Secular City* of the American theologian Harvey Cox (1929–). There are indications that the process of secularization does not occur in the same degree or occurs in a different manner in non-Western cultures.

In general, the main question of the sociology of religion concerns the effectiveness with which it can relate to other studies of religion. This question is posed in *The Scientific Study of Religion,* by the American sociologist J. Milton Yinger (1916–). A similar tendency is noted in the synthesis between the history and the sociology of religion in a new-style evolutionism propounded by another American scholar, Robert Bellah (1927–).

THE PSYCHOLOGY OF RELIGION

The study of religious psychology involves both the gathering and classification of data and the building and testing of various (usually rather wide-ranging) explanations. The former activity overlaps with the phenomenology of religion, so it is to some extent an arbitrary decision under which head one should include descriptive studies of religious experience and related subjects.

Investigations of religious experience

Psychological studies. Notable among investigations by psychologists was *The Varieties of Religious Experience,* by the American philosopher and psychologist William James (1842–1910), in which he attempted to account for experiences such as conversion through the concept of invasions from the unconscious. Because of the clarity of his style and his philosophical distinction, the work has had a lasting influence, though it is dated in a number of ways and his examples come fiom a relatively narrow selection of individuals, largely within the ambit of Protestant Christianity. This points to a recurring problem—that of relating individual psychology to the institutions and symbols of different cultures and traditions.

More radical, but drawing from a rather larger range of examples, was the American psychologist J.H. Leuba (1868–1946). In A *Psychological Study of Religion,* he attempted to account for mystical experience psychologically and physiologically, pointing to analogies with certain drug-induced experiences. Leuba argued forcibly for a naturalistic treatment of religion, which he considered to be necessary if one was to look at religious psychology scientifically. Others, however, have argued that psychology is in principle neutral, neither confirming nor ruling out belief in the transcendent. Most scholars would, however, consider the problem to be a complex philosophical one, which goes beyond psychology as such.

Among those who have attempted a fairly detailed classification of mystical experience, but not necessarily from a scientific-psychological point of view, mention should be made of the English scholar Evelyn Underhill (1875–1941), drawing on examples from the Jewish, Christian, and Muslim traditions. Recently, systematic explorations (taking into account Eastern mysticism as well) have been undertaken. Rudolf Otto was important in elucidating the nature of numinous experience, and there has also been a certain amount of scholarly work performed in the description and classification of types of shamanism, spirit possession, and similar phenomena.

Theories of Freud and Jung

Psychoanalytical studies. More influential than James and Leuba and others in that tradition were the psychoanalysts. Thus, Sigmund Freud (1856–1939) gave explanations of the genesis of religion in various of his writings. In *Totem and Taboo* he applied the idea of the Oedipus complex (involving unresolved sexual feelings of, for example, a son toward his mother and hostility toward his father) and postulated its emergence in the primordial stage of human development. This stage he conceived to be one in which there were small groups, each dominated by a father. According to Freud's reconstruction of primordial society, the father is displaced by a son (probably violently), and further attempts to displace the new leader bring about a truce in which incest taboos (proscriptions against intrafamily sexual relations) are formed. The slaying of a suitable animal, symbolic of the deposed and dead father, connected totemism with taboo. In *Moses and Monotheism* Freud reconstructed biblical history in accord with his general theory, but biblical scholars and historians would not accept his account since it was in opposition to the point of view of the accepted criteria of historical evidence. His ideas were also developed in *The Future of an Illusion.* Freud's view of the idea of God as being a version of the father image and his thesis that religious belief is at bottom infantile and neurotic do not depend upon the speculative accounts of prehistory and biblical history with which Freud dressed up his version of the origin and nature of religion. The theory can still stand as an account of the way in which religion operates in individual psychology, though of course it har also attracted criticism on grounds other than historical ones (*e.g.*, Buddhism does not have a father figure to worship).

A considerable literature has developed around the relationship of psychoanalysis and religion. Some argue, despite the atheistic mood of Freud's writing and his critique of religious belief, that the main theory is compatible with faith—on the grounds, for instance, that the theory describes certain mechanisms operative in people's religious psychology that represent modes in which people respond to the challenge of religious truth. Even if this position can be sustained, it is clear, nevertheless, that acceptance of Freudian insights makes a considerable difference to the way in which one views religious experience and behaviour. Questions have arisen about the range of applicability of Freud's ideas—*e.g.*, whether or not his theories apply outside the Western milieu, such as in Theravāda Buddhism, which does not possess a father figure or worship a god. Various attempts have been made to test Freud's theory of religion empirically, but the results have been ambiguous.

The Swiss psychoanalyst C.G. Jung (1875–1961) adopted a very different posture, one that was more sympathetic to religion and more concerned with a positive appreciation of religious symbolism. Jung considered the question of the existence of God to be unanswerable by the psychologist and adopted a kind of agnosticism. Yet he considered the spiritual realm to possess a psychological reality that cannot be explained away, and certainly not in the manner suggested by Freud. Jung postulated, in addition to the personal unconscious (roughly as in Freud), the collective unconscious, which is the repository of human experience and which contains "archetypes" (*i.e.*, basic images that are universal in that they recur in independent cultures). The irruption of these images from the unconscious into the realm of consciousness he viewed as the basis of religious experience and often of artistic creativity. Religion can thus help men, who stand in need of the mysterious and symbolic, in the process of individuation—of becoming individual selves. Some of Jung's writings have been devoted to elucidating some of the archetypal symbols, and his work in comparative mythology, the history of alchemy, and other similar areas of concern has proved greatly influential in stimulating the investigations of other interested scholars. Thus, the Eranos circle, a group of scholars meeting around the leadership of Jung, has contributed considerably to the history of religions. Associated with this circle of scholars have been Mircea Eliade, the eminent Romanian-French historian of religion, and the Hungarian-Swiss historian of religion Karl Kerényi (1897–). This movement has been one of the main factors in the modern revival of interest in the analysis of myth.

Among other psychoanalytic interpreters of religion, the American scholar Erich Fromm (1900–) has modified Freudian theory and has produced a more complex account of the functions of religion. Part of the modification is viewing the Oedipus complex as based not so much on sexuality as on a "much more profound desire°—namely, the childish desire to remain attached to protecting figures. The right religion, in his estimation, can, in principle, foster men's highest potentialities, but religion in practice tends to relapse into being neurotic. Authoritarian religion, according to Freud, is dysfunctional and alienates man from himself.

Other studies. Apart from Jung's work, there have been various attempts to relate psychoanalytic theory to comparative material. Thus, the English anthropologist Meyer Fortes, in his Oedipus and Job in West *African* Religion, combined elements from Freud and Durkheim, and G.M. Carstairs (a British psychologist), in The T ice *Born*, investigated in depth the inhabitants of an Indian town from a psychoanalytic point of view and with special reference to their religious beliefs and practices. Among the more systematic attempts to evaluate the evidences of the various theories is Religious Behaviour, by Michael Argyle, another British psychologist.

A certain amount of empirical work in relation to the effects of meditation'and mystical experience—and also in relation to drug-induced "higher" states of consciousness—has also been carried on. Investigation of religious responses as correlated with various personality types is another area of enquiry; and developmental psychology of religion, largely under the influence of the French psychologist Jean Piaget (1896–), has played a prominent part in educational theory in the teaching of religion. Most scholars agree, however, that more needs to be done to make results in the psychology of religion more precise; and also, for reasons that are unclear, very few people recently have concerned themselves with the field, which thus is in a state of suspension after a flurry of activity in the late 19th and early 20th centuries.

PHILOSOPHY OF RELIGION

Trends in the philosophy of religion

The concerns of the philosophy of religion. The scope of the philosophy of religion has changed somewhat in the last century and a half—that is, in the time since it came to he recognized as a separate branch of philosophy. Its nature is, as is typically the case in philosophy, open to debate. Three main trends, however, can be noted: (1) the attempt to analyze and describe the nature of religion in the framework of a general view of the world; (2) the effort to defend or attack philosophically various religious positions; and (3) the attempt to analyze religious language. Philosophical materials are also often incorporated into theologies—a modern example being the use of Existentialism in the theology of Rudolf Bultrnann, the German New Testament scholar (see below *Neo-orthodoxy and demythologization*), and others; an older example is the medieval theologian Thomas Aquinas' use of Aristotle and of his (Aquinas' own) insights in the service of a systematic Christian theology. The different activities mentioned above overlap substantially. The second of them is usually taken to include the exploration of natural theology (*i.e.*, the truths about God that can be known, as it is claimed, by the aid of men's reasoning and insight, independently of the truths vouchsafed by revelatior'. Metaphysical systems (concerning the nature of reality) sometimes function as analogues to natural theology and thus provide a kind of support for a revealed religious belief system. Thus, much of philosophy of religion is concerned with questions not so much of the description of religion (historically and otherwise) as with the truth of religious claims. For this reason philosophy can easily become an adjunct of theology or of antireligious positions. To this extent, philosophy lies outside the main disciplines concerned with the descriptive study of religion; thus, it is often difficult to disentangle descriptive problems from those bearing on the truth of the content of what is being described. Feuerbach's "projection" theory of religion, for example, possessed a metaphysical framework, but it also included empirical claims about the nature of religion. The following brief account of philosophical trends is necessarily selective, leaning toward those philosophical theories that have a stronger content of, or relevance to, descriptive claims about religion.

Theories of Schleiermacher and Hegel. Immanuel Kant's powerful critique of traditional natural theology appeared to rob religion of its basis in reason and to make it an adjunct to morality. But Kant's system depended on drawing certain distinctions, such as that between pure and practical reason, which were open to challenge. One reaction that attempted to place religion in a more realistic position (*i.e.*, as neither primarily to do with pure nor with practical reason) was that of the German theologian and philosopher Friedrich Schleiermacher (1768–1834) in his On Religion: *Speeches* to *Its Cultured* Despisers. He attempted there to carve out a separate territory for religious experience, as distinct both from science and morality. For him the central attitude in religion is "the feeling of absolute dependence." In drawing attention to the affective and experiential side of religion, usually neglected in preceding philosophical discussions, Schleiermacher set in motion the modem concern to explore the subjective or inner aspect of religion. Schleiermacher's main goal, however, was not the exploration of religion as such but rather the construction of a new type of theology—the "theology of consciousness." In so doing he relegated doctrines to a secondary role, their function being to express and articulate the deliverances of religious consciousness. Thus, incidentally, it became important for New Testament historians who were influenced by Schleiermacher to penetrate the religious consciousness of Jesus—this becoming, in effect, the reputed locus of his divinity.

Philosophical approach to religious experience

G.W.F. Hegel had, as noted above, a profound effect upon the development of historical and other studies. His own system, the system of the Absolute, contained a view of the place of religion in human life. According to this notion, religion arises as the relation between man and the Absolute (the spiritual reality that undergirds and includes the whole universe), in which the truth is expressed symbolically, and so conveyed personally and emotionally to the individual. As the same truth is known at a higher—that is, more abstract—level in philosophy, religion is, for all its importance, ultimately inferior to philosophy. The relationship between abstract and concrete truth was, incidentally, taken up in the 19th-century Hindu renaissance as a parallel to the doctrine of the Absolute in the Hindu philosopher Sankara (c. 780–820), the Advaita (nondualism), the dominant expression of Hindu metaphysics. The Hegelian account of religion was worked out in the context of the dialectical view of history, according to which opposites united in a synthesis, which in turn produced its opposite, and so on. Hegel was influential in the interpretation of Christian history: Jesus as thesis, Paul as antithesis, and early Catholicism as the synthesis, the latter becoming a new thesis that would elicit a new antithesis, Protestantism.

Hegel attracted some radical criticism, however. One such was that of the aforementioned German philosopher Ludwig Feuerbach (1804–72), whose ideas have been sketched above. Another was that of the Danish philosopher and theologian Søren Kierkegaard (1813–55), sometimes regarded as the father of modern Existentialism, who reacted against the metaphysical and "rational" approach to Christianity in Hegel's thought. Kierkegaard's penetrating psychological insights were put to the service of philosophy and theology and threw new light on the nature of religious experience and its relation to features of man's inner life, such as dread and despair. Kierkegaard's main concern, however, was prophetic rather than descriptive. From a very different standpoint (*i.e.*, that of liberal Protestantism), the German theologian Albrecht Ritschl (1822–89) made an apologetic defense of Christianity in his attempt to analyze theological utterances as essentially affirming value judgments.

Schleiermacher's delineation of religious experience was complemented by attempts among the Romantics and by the German philosopher Ernst Cassirer (1874–1945) to exhibit the nature of symbolic thinking and in particular the special character of religious symbolism. This was some distance from the rationalism of Kant, though Cassirer was nevertheless influenced by the Neo-Kantian tradition.

Influence of Existentialism and Linguistic Analysis

Empiricism and Pragmatism. The Hegelian school, very influential in the 19th century, entered a period of rapid decline in the early part of the 20th. The common sense and scientifically oriented philosophy of the English scholars G.E. Moore (1873–1958) and Bertrand Russell (1872–1970) introduced a period of Empiricism in Britain, while William James's Pragmatism had a similar effect in America. Theologically, there was an antimetaphysical revolution during and after World War 1. On the continent of Europe, the increasing Influence of Existentialism was hostile to the old type of metaphysics. British Empiricism was expressed very strongly in Logical Positivism (maintaining the exclusive value of scientific knowledge and the denial of traditional metaphysical doctrines) and its linguistic aftermath. This stimulated the analysis of religious language, and the movement was complicated by the transformation in the thought of the Austrian-English philosopher Ludwig Wittgenstein (1889–1951), who in his later thought was very far removed from his early, rather formalistic treatment of language.

Theoretically, the Analytic attempt to exhibit the nature of religious language could have been a chiefly descriptive task, but, in fact, most analyses have occurred in the context of questions of truth—thus some scholars have been concerned with exhibiting how it is possible to hold religious beliefs in an Empiricist framework, and others with showing the meaninglessness or incoherence of belief. A landmark was the publication, in 1955, of New Essays in Philosophical Theology, edited by the English philosophers A.G.N. Flew and A. MacIntyre. Though Wittgenstein stressed the idea of "forms of life," according to which the meaning of religious beliefs would have to be given a practical and living contextualization, little has been done to pursue the idea empirically. The discovery by the English philosopher J.L. Austin (1911–60) and others of performative uses of language has stimulated some enquiry in this direction. On the whole, however, the Analytic philosophy of religion has been pursued rather independently of the descriptive study and history of religion.

Influence of Heidegger and Sartre

Modern Existentialist and Phenomenological studies. Since linguistic philosophy tends to be considered by its proponents to be a method or a group of methods, internal diversity within the area of concern is not surprising. Similarly, Existentialism, which is less of an "-ism" than an attitude, expresses itself in a variety of ways. The most influential modern Existentialists have been the German philosopher Martin Heidegger (1889–1976) and the French philosopher, dramatist, and novelist Jean-Paul Sartre (1905–80); the former was especially important in the development of modem continental theology, particularly for the use made of some of his ideas by Rudolf Bultmann.

According to Heidegger, man's existence is characterized as "care." This care is shown first in possibility: man makes things instrumental to his concerns and so projects forward. Secondly, there is his facticity, for he exists as a finite entity with particular limitations (his "thrownness"). Thirdly, man seeks to avoid the anxiety of his limitations and thus seeks inauthentic existence. Authenticity, on the other hand, involves a kind of stoicism (positive attitude toward life and suffering) in which death is taken up as a possibility and man faces the "nothing." The structure of man's world as analyzed by Heidegger is revealed, in a sense, affectively—*i.e.*, through care, anxiety, and other existential attitudes and feelings.

Sartre's thought has had less direct impact on the study of religion, partly because his account of human existence represents an explicit alternative to traditional religious belief. Sartre's analysis begins, however, from the human desire to be God: but God is, on Sartre's analysis, a self-contradictory notion, for nothing can contain the ground of its own being. In searching for an essence man fails to see the nature of his freedom, which is to go beyond definitions, whether laid down by God or by other human beings.

The French philosopher Gabriel Marcel (1889–1973) is not individualistic like Sartre (or at least the early Sartre, whose thinking was modified by Marxism); instead, he stresses the communal character of human existence—the highest virtue being fidelity. Marcel also emphasizes the mysterious (as distinguished from the empirically problematic) character of love, evil, hope, freedom, and, above all, being. His work provides a rich analysis and interpretation of the religious dimensions of human experience and thus is a philosophical basis for the study of religious experience.

The Existentialist approach attempts to describe and evoke the way human beings are and thus can lay claim to be phenomenological. It is clear, however, from the divergencies among Existentialists, that they contain speculative and idiosyncratic elements, and one question raised about the general applicability of their characterizations is how far they are bounded by the product of a particular mood in Western culture.

Influence of Phenomenology

The German philosopher Edmund Husserl (1859–1938) has had, as the main exponent of Phenomenology, a wide effect on the study of religion. His program of describing experience and "bracketing" the objects of experience, in the pursuit of essences of types of experience, was in part taken up in the phenomenology of religion. Husserl distinguished Phenomenology from psychology, however, because, in his view, the latter concerns facts in a spatio-temporal setting, whereas Phenomenology uncovers timeless essences. This aspect of Husserl's thinking has not always or wholly been accepted by phenomenologists of religion, who have been much more oriented toward facts, though Husserl's emphasis on essences often has tended to make religious phenomenology lean toward a static typology.

Relationship between Western and non-Western philosophy in regard to religion. Western philosophy has thus had a significant influence on the study of religion. It has also come into contact with non-Western traditions and has thus stimulated concern with the problem of the nature of religious truth in a world perspective. The most influential product of this interplay has most likely been the neo-Advaitin philosophy (a new version of Advaita, or nonduality) espoused by a number of modem Indians, such as Swami Vivekananda (1863–1902), who made a sensational appearance at the Parliament of Religions in Chicago in 1893, and the Indian philosopher Sarvepalli Radhakrishnan (1888–1975). Both of these thinkers attempted to reveal the underlying unity in the great religions—a unity described from a point of view drawing on the thought of Śaṅkara.

The idea of a world faith

The U.S. philosopher William Ernest Hocking (1873–1966) pursued similar interests in the construction of a world faith that he considered might come about through the mutual modification of, and interchange between, the great religious traditions. These concerns have raised important questions about the criteria of truth between religions, the tests of whether one religion is truer than others, and the extent to which valid identifications of belief can be made between one faith and another. The various elements of the philosophical traditions of the last two centuries have thus had a bearing on religious questions, and most scholars consider that though the philosophy of religion tends to be normative rather than descriptive, it is a necessary adjunct to descriptive studies. Philosophical insights and expertise are of significant relevance to the numerous questions of method that arise in the study of religion.

THEOLOGICAL STUDIES

Historical-critical studies. The major feature in the development of Christian theology during the 19th and 20th centuries has been the impact of historical enquiry on the biblical sources of belief (there has also been a similar effect on Jewish and other theologies, but Christian theology has been the most influential in the development of Western culture). A pioneer in the attempt to understand the mythological elements in the New Testament was the German theologian David F. Strauss (1808–74), whose controversial Life of Jesus (published in German, 1835–36) was an attempt to sift out the historical Jesus from the overlay of myth created by the poetic imagination of the early church. Similarly, the German church historian Adolf von Harnack (1851–1930), influenced by Albrecht Ritschl, intended to penetrate the accretions of dogma attached to the historical Jesus. Such attempts were later to come under radical criticism from, among others, the Alsatian philosopher-theologian and Nobel laureate Albert Schweitzer (1875–1965) for describing the alleged Jesus of history in terms tailored to fit the presuppositions of liberal Protestantism. Thus was raised an important methodological question on how to deal with such material as the Gospels.

Principles of historical criticism

Important in trying to spell out principles for dealing with the material was Ernst Troeltsch, who argued that history has to be written in accordance with the following principles: first, the principle of criticism—*i.e.*, the sifting of the evidences and testing of conclusions (thus historical certainty about much in the ancient witnesses to Jesus is impossible); second, the principle of analogy—*i.e.*, in the absence of first-hand experience, scholars must treat reports of miraculous events with skepticism since men do not encounter such events in their own experiences (here Troeltsch adopts the position of David Hume); and third, the principle of correlation—*i.e.*, events in history are continuous with one another in a causal nexus, which rules out irruptions into the causal order by God: if he works in history he is immanently in all of it. Troeltsch, it may be noted, had some effect on the sociology of religion—*e.g.*, in his distinction between church-type and sect-type organizations in the history of Christianity, a distinction that has formed the starting point of considerable researches in recent times, as noted above. The implications of Troeltsch's historical treatment of religion seemed to be relativistic. Christianity, at any rate, is viewed as a part of religious history as a whole, a point that had not always been clearly recognized by theologians. Troeltsch thereby raised some important questions about the relationship between Christianity and other religions and showed how Christian theology was beginning to take a more realistic view of mankind's religious experience and history, in distinction to the earlier rather simplistic dichotomies between special (*i.e.*, Judeo-Christian) and general (*i.e.*, natural) revelation.

Discoveries about ancient Near Eastern religions were also bound to affect biblical studies, and a well-defined school developed in Germany—the Religionsgeschichtliche Schule (History of Religions school)—which was critical of the rather unhistorical treatment of Jesus by Ritschl and others. This school emphasized the degree to which biblical ideas were the product of the ancient cultural milieu. Important in this line of development was Albert Schweitzer, in whose Quest of the Historical Jesus the eschatological teachings (statements about the "last times," or end of the world as it is now understood) of Jesus are emphasized, together with the dissimilarity of his thought world from our own. Criticism of Harnack also came from a different direction. The French theologian Alfred Loisy (1857–1940), from a Catholic point of view but taking into account the work of Protestant biblical critics, found the essence of Christianity in the faith of the developed church, which could not be found simply by trying to discover the nature of the historical Jesus. The founder, in effect, of Catholic Modernism, Loisy was condemned by his own church, and this was a main factor in discouraging some of the livelier Catholic studies of the New Testament until after the epochal ecumenical second Vatican Council (1962–65).

Influence of Barth and Bultmann

Neo-orthodoxy and demythologization. Liberal Protestantism of the Harnack type was severely criticized by Karl Barth (1886–1968), the founder of Neo-orthodoxy; liberalism's optimism, in any event, came under a cloud through the outbreak of World War I. Barth's *Epistle* to the Romans and his later *Church* Dogmatics became highly influential. His theology depended in part on a distinction between the Word (*i.e.*, God's self-revelation as concretely manifested in Christ and in preaching) and religion. The latter, according to Barth, is the product of human culture and aspirations and is not to be identified with saving revelation (for salvation cannot come from man, only from God). This rather uncompromising view made use of the projectionist theory of religion expressed by Feuerbach and others. Barth's conclusion was challenged somewhat by another Swiss theologian, Emil Brunner (1889–1966), who allowed a modicum of insight for fallen man into God's nature. The concession was, however, a slight one. The Dutch theologian Hendrik Kraemer (1888–1965) applied the doctrine of the theology of the Word to non-Christian religions in The *Christian* Message in a Non-Christian Would, which had a wide impact on the overseas mission field. Since man's religions are cultural products and since each system of belief is organic and particular, there are, according to Kraemer, no points of contact between them and the Gospel (even Christianity as an empirical religion must be distinguished from it: its only advantage is to have been continuously under the judgment and influence of the Gospel). Kraemer's position has come under some criticism from students of comparative religion; one of the theological problems it poses is that it seems to shut off the possibilities of dialogue between religions.

After Barth, the most influential theologian in the 20th century has been Rudolf Bultmann (1884–). Though mainly concerned with the presentation of the faith, his project of "demythologization" has a wide significance for the historian of religions, for it involves a theory of myth. Bultmann comes to the New Testament material partly as a historian and partly as a theologian influenced by the Existentialism of Heidegger. He centres his interest on the difference between the style of thinking in the early church, as expressed in the New Testament writings, and modern thought. Modern man, he holds, cannot think in the mythological terms employed in the New Testament presentation of the Gospel. Therefore, it is necessary to demythologize the New Testament message. For Bultmann, the mythological elements are belief in the pre-existence of Christ, the three-layer universe (heaven, earth, and hell), miracles, ascension into heaven, demonology, and various other elements of the Judeo-Christian-Hellenistic world view. The inner meaning of the myths, he claims, has to be explicated in existential terms and purged of the objectifications that they contain. Thus, his theory contains an empirical claim, namely about the original function of myths (expressing existential attitudes through objectified representations). Bultmann's theory, however, has not yet been brought together with anthropological and other theories of myth.

A follower of Bultmann, Fritz Buri (1907–), considers Bultmann's stance to be insufficiently radical, for Bultmann still differentiates between the kerygma (the essential proclamation of the early church) and the myths, desiring to retain the former, but not the latter. Buri has attempted to overcome this distinction. Authentic existence is not, according to Buri, distinctively Christian, and he has been led to a position not altogether different in principle from that of Troeltsch. Buri's views have also led him into considering in some depth the significance of other religions.

Inter-religious dialogue

The relationship of Western Christianity to other religions. Since World War II, Western Christianity has found it difficult, from a cultural point of view, to ignore the challenge of other religions; and the mood has changed somewhat from the more rigorous climate in which the theology of the Word (*i.e.*, Barth's position) was dominant. The "theology of religions" (analogous to the "history of religions") has moved in the direction of dialogue, which sometimes simply refers to mutual acquaintance in charity so that men of differing faiths can come to understand more deeply the meaning of each other's religions. More significantly, it means a kind of mutual theologizing. Among the more prominent writers who have been involved one way or another in the process of dialogue have been the Jewish philosopher Martin Buber (1878–1965), the English Islāmic scholar Kenneth Cragg (1913–), and the Canadian Islāmic scholar Wilfred Cantwell Smith. In effect, modern dialogue continues an earlier tradition that emphasized some continuities between religions, notably the work of the British theologian John Oman (1860–1939), who was influenced both by Schleiermacher and Otto, though critical of the latter. Oman contrasted prophetic and mystical religion and considered that the former had the highest conception of the supernatural. There are analogies between his position and that of the important Swedish theologian, historian of religion, and archbishov Nathan Soderblom (1866–1931).

A rather different theory of myth and symbolism from that of Bultmann was expressed by Paul Tillich, who viewed religion as having to do with what concerns man ultimately. He taught that symbolic and mythological language, used by all religions, points beyond itself to the being in which the symbols participate. Tillich used the term being in an existential sense (one related directly to human experience and commitment) rather than a strictly metaphysical sense. Also, he claimed that it is not possible to dispense with the symbolic, which is essential to the task of speaking about ultimate reality, but the myths are to be "broken"—that is, they are to be seen as not being literally true.

Christian theology, in the 19th and 20th centuries, has been more concerned with intellectual and social challenges, however, than with the analysis of religion, which has been secondary to that concern.

HISTORY AND PHENOMENOLOGY OF RELIGION

The history of religions and the phenomenology of religion are generally understood by scholars to be nonnormative—that is, they attempt to delineate facts, whether historical or structural, without judging them from a Christian or other standpoint. At any rate, their tasks are considered to be different from that of articulating and systematizing a faith. The same, in principle, is true for the comparative study of religion, though this sometimes is thought to cover the theology of other religions, such as the Christian appraisal of Hindu history. Needless to say, the fact that a discipline aims to be nonnormative does not mean that it will succeed in being so. Also, the history and phenomenology of religion tend to raise essentially philosophical questions of explanation, where the issues are often debatable.

Modern origin and development of the history and phenomenology of religion. The history of religions on a cross-cultural basis, though it has quite an ancient pedigree, came into its own in a modern sense from about the time of the German comparative philologist Max Miiller (1823–1900). During the latter part of the 19th century an attempt was made to place comparative methodology on a systematic basis (often called the Science of Religion), and in this connection the work of the Dutch theologians P.D. Chantepie de la Saussaye (1848–1920) and C.P. Tiele (1830–1902) was important. During this period, various lectureships and chairs in the subject were instituted. In The Netherlands, following the reform of the theological faculties in 1876, four chairs in the history of religions were founded. In 1879 a chair was founded at the Collège de France (followed by others elsewhere in France), while a number were created in Switzerland. The subject also spread to Great Britain (where chairs at Manchester and London were instituted), the United States (at Harvard and Chicago), and elsewhere in the Western world. In Germany, on the other hand, there was strong resistance, notably from Adolf von Harnack, who thought that theology should avoid what he regarded as dilettantism and that the subject was sufficiently covered in the study of biblical religion.

The first congress of *Religionswissenschaft* (Science of Religion) took place in Stockholm in 1897, and a similar one in the history of religions at Paris in 1900. Later, the International Association for the History of Religions, dedicated to a mainly nonnormative and nontheological approach, was formed. Also important was the compilation of encyclopaedias, notably Hastings' *Encyclopaedia of Religion and Ethics,* with many distinguished contributions. Thus, there were development and progress in the new subject in the latter part of the 19th and early part of the 20th century. In the 1960s came the next major burst of expansion.

A great amount of the work of scholars in the field has been devoted to exploring particular histories—piecing together, for instance, the history of Gnosticism (a Hellenistic-Christian heretical sect that emphasized dualism) or of early Buddhism. In principle, Christianity is considered from the same point of view, but much significant work has also been comparative and structural. This can range from the attempt to establish rather particular comparisons, such as Otto's comparison (in his *Mysticism East and West)* of the medieval German mystic Meister Eckehart and the medieval Hindu philosopher Śaṅkara, to a systematic typology, as in *Religion in Essence and Manifestation* by the Dutch historian of religion Gerardus van der Leeuw.

The concept of the Holy

There have been many significant scholars in the history and phenomenology of religion since Max Miiller. Rudolf Otto (1869–1937) made a profound impression on the scholarly world with the publication of *The Idea of the Holy* (in its German edition of 1917), which showed the influence of Schleiermacher, Marett, Edmund Husserl, and the Neo-Kantianism of Jakob Fries (1773–1843). More important than the philosophical side of his enterprise, however, was the excellent delineation of a central experience and sentiment and the elucidation of the concept of the Holy. The central experience Otto refers to is the numinous (Latin *numen*, "spirit") in which the Other (*i.e.*, the transcendent) appears as a *mysterium tremendum et fascinans*—that is, a mystery before which man both trembles and is fascinated, is both repelled and attracted. Thus, God can appear both as wrathful or awe inspiring, on the one hand, and as gracious and lovable, on the other. The sense of the numinous, according to Otto, is *sui generis,* though it may have psychological analogies, and it gives an access to reality, which is categorized as holy. Otto stresses what he calls the nonrational character of the numinous, but he does not deny that rational attributes may be applied to God (or the gods or other numinous powers), such as goodness and personality. The impact of Otto's work, however, does not depend on the now rather curious Neo-Kantian scheme into which he presses his data. Not all scholars would agree that the numinous is universal as a central element in religion, as Otto seems to have supposed: early Jainism and Theravāda Buddhism, for example, have other central values. Otto's treatment of mysticism, which is central to Buddhism, wavers somewhat, and the notions of the "wholly Other" and of the *tremendurn* do not easily apply to the experience of Nirvana (the state of bliss) or to other deliverances of the contemplative mystical consciousness.

Friedrich Heiler (1892–1967), like Otto a professor at Marburg (Germany), was a strong proponent of the phenomenological and comparative method, as in his major work on prayer. Heiler, however, went beyond the scientific study of religion in attempting to promote interreligious fellowship, partly through the Religioser Menschheitsbund (Union of Religious Persons), which he helped to found. Heiler believed in the essential unity of religions—a recurring theme in various guises in the period, though open to question because of the widely apparent divergences between prophetic and other religions, such as Theravāda Buddhism and Jainism, which do not believe in a supreme personal being.

The phenomenologist of religion who probably has had the greatest influence after Otto, partly because he is fairly explicit about method, is Gerardus van der Leeuw (1890–1950), who was somewhat influenced by the French anthropologist Lucien Lévy-Bruhl (1857–1939) and his notion of prelogical mentality, which he applied to primitive cultures to distinguish them from civilized cultures. Van der Leeuw emphasized power as being the basic religious conception. His major work, *Religion in Essence and Manifestation,* is an ambitious and wide-ranging typology of religious phenomena, including the kinds of sacrifice, types of holy men, categories of religious experience, and other types of religious phenomena. The work has been criticized, however, as being unhistorical. Partly because of his philosophical presuppositions, borrowed chiefly from Husserl, van der Leeuw held the disputable doctrine that Phenomenology knows nothing of the historical development of religion: it picks out timeless essences of religious phenomena. Apparently it is not necessary, however, to hold this doctrine, since one could as well classify types of religious change (*i.e.*, temporal sequences), as indeed Max Weber attempted to do. Classificatory and historical techniques and conclusions are not incompatible, however. Thus, the work of Nathan Soderblom, who, as well as being a historian of religions, was prominent in the ecumenical movement, combined the two aspects in his Living God.

Emphases of Joachim Wach and Mircea Eliade

The "Chicago school." The phenomenological method was brought to the United States primarily by the German-American historian of religions Joachim Wach (1898–1955), who established *Religionswissenschaft* (Science of Religion) in Chicago and was thus the founder of the modern "Chicago school" (though his successor, Mircea Eliade, has a rather different slant). Wach was concerned with emphasizing three aspects of religion—the theoretical (or mental; *i.e.*, religious ideas and images), the practical (or behavioral), and the institutional (or social); and because of his concern for the study of religious experience, he interested himself in the sociology of religion, attempting to indicate how religious values tended to shape the institutions that expressed them. Wach, however, was not committed to a religious neutralism in his use of the idea of a "science of religion." For him, Religionswissenschaft deepens the sense of the numinous and strengthens, rather than paralyzes, religious impulses.

Mircea Eliade (1907–), a Romanian scholar who emigrated to the United States after World War II, has had a wide influence, partly because of his substantive studies on yoga (a Hindu meditation technique) and on shamanism (both these major works are now regarded as classical studies of their subjects) and partly because of his later writings, which attempt to synthesize data from a wide variety of cultures. The synthesis incorporates a theory of myth and history. Eliade was also a founder of the journal History of Religions, which expresses the "Chicago school" viewpoint. Eliade has been somewhat influenced by Jung, both in his psychological interpretations of certain religious experiences (such as those attained in the practice of yoga) and more importantly in his attempt to give an interpretation in depth to the mythic material over which he ranges so widely. He also affirms strongly the importance of the history of religions in the intellectual world and is thus concerned to emphasize its unique and positive role in providing a "creative hermeneutics" (critical interpretive method) of man's religious and existential condition. Two important elements in the theory of Eliade are, first, that the distinction between the sacred and the profane is fundamental to religious thinking and is to be interpreted existentially (the symbols of religion are, typically, profane in literal interpretation but are of cosmic significance when viewed as signs of the sacred); and, second, that archaic religion is to be contrasted with the linear, historical view of the world. The latter essentially comes from biblical religion; the former viewpoint tends to treat time cyclically and mythically—referring to foundational events, such as the creation, the beginning of the human race, and the Fall of man, on to *illud tempus* (the sacred primordial time), which is re-enacted in the repetitions of the ritual and in the retelling of the myth. Though Christianity has contained archaic elements, in essence it is linear and historical. Thus, faith in Christianity involves a kind of fall from archaic timelessness, and secularization—in which the overt symbolism of religion is driven underground into the unconscious—is a second fall. Eliade is not very explicit about his meaning beyond this point. Not only is he concerned with descriptive phenomenology, in which context his analysis of the religious functions of time and space is most illuminating, but also with a kind of metaphysical speculation (as exemplified in his idea of the "fall").

The myth and ritual school

Other studies and emphases. Though not always giving a detailed account of the correlation between myth and ritual, Eliade is indebted to the so-called myth and ritual school, which has influenced thinking in the history of religions and which was important in the 1930s, especially in the interpretation of Near Eastern mythology. Thus, the *Enuma* elish, the Babylonian creation epic, was discovered to be no mere set of stories but rather a mythic drama re-enacted every year at the spring festival, at which time the foundation of the world is ritually renewed. More generally, it was seen that for a wide range of sacred stories it was important to discover the ritual context. The most influential statement of the school's position is to be found in *Myth* and Ritual (1933), edited by the English biblical scholar and Orientalist Samuel Hooke.

Meanwhile, the categorization of types of religion (*e.g.*, as polytheism, henotheism, or other) continued to stimulate attempts at a deeper understanding of the emergence of monotheism. To some extent scholars remained under the influence of the older evolutionism. An important work in this connection was Dio: Formazione *e sviluppo del monoteismo nella* storia *delle* religioni ("God: Formation and Development of Monotheism in the History of Religions"), by the Italian historian of religion Raffaele Pettazzoni (1883–1959), who emphasized the importance of the divinized sky in the development of monotheism. He was critical of the *Urmonotheismus* of Wilhelm Schmidt, considering that the latter's theory of an original monotheism went very far beyond the evidence. At best, the facts could only support the conclusion that primitive peoples believed in a supreme celestial being. Pettazzoni, in his concern for problems of method, was critical of the sharp division between phenomenology and history. He considered that the former cannot exist without the historical sciences—*e.g.*, history, philology, and archaeology—but that it supplies scholars in the latter fields a sense of the religious significance of what they discover. This point of view has also been more vigorously espoused by the Swedish scholar Geo Widengren (1907–), who has specialized mainly in Iranian religions. The need to integrate historical and structural studies has caused some debate in recent years; and there has also been some contrast made between historical approaches and contemporary sociological and (essentially theological) dialogical approaches to religion. To some extent, such debates represent different ideals of scholarship; but it is difficult to note where the essential incompatibilities lie. For many scholars, the multidisciplinary way of studying religion is difficult to comprehend.

Studies on the sociological functions of deities

Meanwhile, the longstanding interest in the Indo-European group of religions was given a new impetus in the work of the French comparative philologist and mythologist Georges Dumézil (1898–), who broke away from an etymological (analysis of word derivations) approach and sought instead the thematic traits of the gods in the mythical material. This approach, pioneered by others before Dumézil, also was skeptical of the easy identification of gods with natural forces and emphasized the sociological functions of the divinities—without, however, holding to a reductionist theory. Dumézil's theory was partly stimulated by discoveries in the Near East, notably that of Boğazköy (Turkey), which revealed a similarity between some of the chief gods of the Indo-European Mitannians and those of the Aryans of the Indian Vedic tradition. His theory correlated the functions of the gods with the tripartite division of Indo-European societies—namely the priestly regal, the nobility, and the producers (agriculturalists, craftsmen). Though his work has been controversial (there are, for instance, some difficulties about its application to ancient Greece, despite the fact that the analysis seems to apply to the threefold division of society into philosophers, warriors, and producers in Plato's Republic), there is no doubt that the search for correlated functions of the kind Dumézil postulated has been significant in the area of Indo-European mythology.

Dumézil's work is one example of a thematic, comparative study. The interest in such studies has grown since World War II. Examples can be found in the writings of such thinkers as the English scholar S.G.F. Brandon (1907–1971) in his treatment of ideas such as creation and time in different religions, but with special reference to the ancient Near East, and the English Indo-Iranian scholar R.C. Zaehner (1913–), notably in his work on mysticism, as in his *Mysticism Sacred and Profane.* Zaehner's is a definitely Christian approach rather than a scientific-descriptive one; and his concern is to distinguish between theistic and other forms of mysticism, such as monistic mysticism as found, according to him, in Yoga, Advaita, and even Theravāda Buddhism.

Apart from the comparative, phenomenological studies, there has also been a strong growth of historical work in regard to particular religions. This has been most obvious in Indian religions—in Hinduism and Buddhism especially. In part, this is the result of a general growth in non-Christian religions in the post-World War II era and of the need to come to terms with Asian and African cultures after the demise of European hegemony.

IV. **Conclusion**

Problems and directions of the study of religions

The foregoing, a necessarily rather selective account of some of the principal developments and scholars in the various disciplines related to the study of religion, emphasizes the artificiality of some of the divisions between traditional disciplines. Thus, Dumézil's work could as easily fall under sociology or anthropology as under the history of religions; and there are obvious connections between philosophy and sociology in, for example, Marxist interpretations of religion. Again, the description and typology of religious experience belong as much to psychology as to the phenomenology of religion, and the analysis of the nature of symbolism requires a variety of disciplinary approaches. To some extent, the study of religion has suffered from the barriers between disciplines, and this fact is increasingly recognized in the formulations, notably in the United States, of the idea of religion as a subject that should be institutionalized in a university department or program in which historians, phenomenologists, sociologists, and members of other disciplines work together. There are some, however, who consider that there are dangers in such an arrangement; thus Eliade prefers to work rather tightly within the framework of the history of religions, concerned lest the social sciences overwhelm and distract the interpreter of religious meanings. Similarly, the theological tradition in the West remains powerfully operative (quite legitimately) in regard to the articulation of the Christian faith and sometimes resists any attempt to treat Christianity itself in the manner dictated by the history and phenomenology of religion. Thus, the history of religions and the comparative study of religion still tend to mean in practice "the study of religions other than Judaism and Christianity." Educational and social pressures have arisen, however, within a secularistic, increasingly pluralistic society and (in effect) a shrunken world, increasing the tendency toward a pluralism in the study of religion that expands the viewpoints of traditional faculties and departments of theology, both in universities and theological seminaries.

A further problem about the multidisciplinary study of religion is that little has been done to explore the problem of the people to whom religions are interpreted—the clientele for the subject. Hitherto, the main assumption has been that the study is for Westerners, though a number of distinguished Asian and African scholars are working in the field. Until recently, owing to the unequal cultural and political relationship between Western and non-Western religions, however, some of the most vital contributions have been primarily attempts to articulate (for the new apologetic situation) the old traditions. This has been a main concern of scholars of Asian religions such as Sarvepalli Radhakrishnan, T.R.V. Murti, and K.N. Jayatilleke. The prospect is, however, that an intellectual community will be the clientele of the subject. To this extent the study of religions will most likely involve, as it does already to some extent, a complex dialogue between religions.

Another problem is the need to elucidate the basis of a dynamic typology of religion in which phenomenology and history are properly brought together. The tendency toward a rift between the historians and phenomenologists is unnecessary and causes harm to the pursuit of the subject.

Meanwhile, some emergent tendencies within the various disciplines can be perceived. There is an increased concern in anthropological theory for the content of religious symbolism, such as in the work of the English anthropologist Mary Douglas; and the sociology of religion is, in a sense, returning to the method of Max Weber in stressing the comparison of cultures. The important development of Oriental and African studies since World War II has made this task earier—American sociologists have, for example, examined in some detail Japanese culture and religion. The interest in symbolism and mythology coincides with developments in the philosophy of religion, which, under the influence of Wittgenstein (in his later, more open phase), is concerned with explicating different functions of language. One area of the study of religion that is seriously underdeveloped at the present time—other than in respect to the psychoanalytic approaches—is the psychology of religion, although current interest in mysticism and other forms of religious experience has stimulated the collection and interpretation of data. One of the difficult problems to be solved is the extent to which cultural conditioning exerts an influence on the actual content of such experience.

In many ways the present position promises well for an expanding multidisciplinary approach to problems in the study of religion. Historians of religion are recognizing some of the contributions to be made by modern sociology, and sociologists—partly because of the development of the sociology of knowledge—have become more aware of the need for accounting for the particular systems of meaning in religion. An area that may very well exhibit the new synthesis is the study of new religious movements.

After a period of relative unconcern, Christian theology is increasingly aware of the challenge of other religious beliefs, so that there are greater impulses toward blending Christian and other studies—often kept rather artificially apart, though biblical studies, especially Old Testament studies, have usually been quite closely ielated to the history of the relevant religions of the ancient Near East.

Meanwhile, in a number of Western countries (chiefly in Europe, but also to some extent in the United States), the study of religion on a pluralistic and multidisciplinary basis is being increasingly viewed as an important element in the education of secondary school students. This, together with the popularity of the subject in universities, may ensure that the study of religion will increase in significance.

BIBLIOGRAPHY. JAN DE VRIES, *The Study of Religion* (1967), a fairly useful and brief historical survey of the development of the subject; H. PINARD DE LA BOULLAYE, *L'Étude comparée des religions,* 2 vol. (1922–25), a thorough and excellent account; J. MILTON YINGER, *The Scientific Study of Religion* (1970), a recent attempt to indicate the multidisciplinary approach to the study of religion; J. HASTING~(ed.), *Encyclopaedia of Religion and Ethics,* 13 vol. (1908–26), dated in many respects but still an enormously important reference work; PAUL EDWARDS (ed.), *Encyclopedia of Philosoplty,* 8 vol. (1967), many entries on world religions, doctrines, and religious thinkers; JOHN MACQUARRIE, *Twentieth Century Religious Thought* (1963), a survey, despite its title, of both 19th- and 20th-century thinkers, including many important in the history and phenomenology of religion (also a good general guide to issues about religion in modern Western theology); G. VAN DER LEEUW, *Phänomenologie der Religion* (1933; Eng. trans., *Religion in Essence and Manifestation,* 1938), the most wide-ranging and ambitious attempt at a systematic and classificatory phenomenology of religion; RUDOLF OTTO, *Das Heilige* (1917; Eng. trans., *The Idea of the Holy,* 2nd ed., 1950), a highly influential classic; MIRCEA ELIADE has written widely from a standpoint that combines elements drawn from depth psychology, phenomenology, and the history of religions: his *Sacred and the Profane* (1961) and *The Quest* (1969) give an insight into his general approach; J. WACH, *The Comparative Study of Religion* (1958) and *Sociology of Religion* (1962), still useful compendiums; J. HINNELLS (ed.), *The Comparative Study of Religion in Education* (1970), contains some recent thinking about the comparative method relating it to education; MICHAEL BANTON (ed.), *Anthropological Approaches to Religion* (1966), and E. EVANS-PRITCHARD, *Theories of Primitive Religion* (1965), indicate the main issues about the genesis and function of religion debated by anthropologists; recent work in the sociology of religion is surveyed usefully in THOMAS O'DEA, *The Sociology of Religion* (1966); MAX WEBER, *Religionssoziologie* (1922; Eng. trans., *The Sociology of Religion,* ed. by TALCOTT PARSONS, 1963), a good introduction to the thought of Weber. PETER BERGER, *The Sacred Canopy* (Brit. title, *The Social Reality of Religion,* 1969), more speculative but stimulating example of modern sociological theorizing about religion; V. LANTERNARI, *The Religions of the Oppressed* (1963), an example of comparative sociology of religion; J. HICK, *The Philosophy of Religion* (1963), a useful survey of the issues raised in the philosophy of religion.

(N.Sm.)

Religions, Classification of

The classification of religions, the attempt to systematize and bring order to a vast range of knowledge about man's religious beliefs, practices, and institutions, has been the goal of students of religion for many centuries but especially so with the increased knowledge of the world's religions and the advent of modern methods of scientific inquiry in the last two centuries.

The classification of religions involves: (1) the effort to establish groupings among historical religious communities having certain elements in common or, (2) the endeavour to group similar religious phenomena in categories that serve to reveal the structure of human religious experience as a whole.

FUNCTION AND SIGNIFICANCE

Classification as a tool in understanding religion

The many schemes suggested for classifying religious communities and religious phenomena all have one purpose in common; *i.e.*, to bring order, system, and intelligibility to the vast range of knowledge about human religious experience. Classification is basic to all science as a preliminary step in reducing data to manageable proportions and in moving toward a systematic understanding of a subject matter. Like the zoologist who must distinguish and describe the various orders of animal life as an indispensable stage in the broad attempt to understand the character of such life as a whole, the student of religion also must use the tool of classification in his outreach toward a scientific account of man's religious experience. The growth of scientific interest in religion in Western universities over the past 130 years has compelled most leading students of religion to discuss the problem of classification or to develop classifications of their own.

The difficulty of classifying religions is accounted for by the immensity of religious diversity that history exhibits. As far as scholars have discovered, there has never existed any people, anywhere, at any time, who were not in some sense religious. The individual who embarks upon the arduous task of trying to understand religion as a whole confronts an almost inconceivably huge and bewilderingly variegated host of phenomena from every locale and every era. Empirically, what is called religion includes the mythologies of the preliterate peoples on the one hand and the abstruse speculations of the most advanced religious philosophy on the other. Historically, religion, both ancient and modem, embraces both primitive religious practices and the aesthetically and symbolically refined worship of the more technologically progressive and literate human communities. The student of religion does not lack material for his studies; his problem is rather to discover principles that will help him to avoid the confusion of too much information. Classification is precisely the appeal to such principles; it is a device for making the otherwise unmanageable wealth of religious phenomena intelligible and orderly.

The classification of religions thus has significance for the method of a general science of religion. Broadly speaking, the scientific study of religion comprehends two aspects: assembling information and interpreting the material gathered in order to elicit its meaning. The first aspect involves the psychological and historical study of religious life and must be supplemented by such auxiliary disciplines as archaeology, ethnology, philology, literary history, and other similar disciplines. The facts of religious history and insight into the development of the historical religious communities are the foundation of all else in the study of religion. Beyond the historical basis lies the task of seeing the entirety of human religious experience from a unified or systematic point of view. The student of religions attempts not only to know the variety of beliefs and practices of *homo religiosus* ("religious man"), but also to understand the structure, nature, and dynamics of religious experience. The student of religion attempts to discover principles that operate throughout religious life--on the analogy of a sociologist seeking the laws of human social behaviour—to find out whether there are also laws that operate in the religious sphere. Only with the attempt to discern the system and structure binding together the differentiated historical richness of religion does a true science of religion, or *Religionswissenschaft*, begin.

The endeavour to group religions with common characteristics or to discover types of religions and religious phenomena belongs to the systematizing stage of religious study. According to F. Max Müller, an Anglo-German comparative philologist of the 19th century,

> All real science rests on classification and only in case we cannot succeed in classifying the various dialects of faith, shall we have to confess that a science of religion is really an impossibility.

PRINCIPLES OF CLASSIFICATION

The criteria employed for the classification of religions are far too numerous to catalogue completely. Virtually every scholar who has considered the matter has evidenced a certain amount of originality in his view of the interrelationships among religious forms. Thus, only some of the more important principles of classification will be discussed.

True and false religion

Normative. Perhaps the most common division of religions—and in many ways the most unsatisfactory—distinguishes true religion from false religion. Such classifications may be discovered in the thought of most major religious groups and are the natural, perhaps inevitable, result of the need to defend particular perspectives against challengers or rivals. Normative classifications, however, have no scientific value, because they are arbitrary and subjective, inasmuch as there is no agreed method for selecting the criteria by which such judgments should be made. But because living religions always feel the need of apologetics (systematic intellectual defenses), normative classifications continue to exist.

Many examples of normative classification might be given. The early Church Fathers (*e.g.*, Clement of Alexandria, 2nd century AD) explained that Christianity's Hellenistic (Greco-Roman culture) rivals were the creations of fallen angels, imperfect plagiarisms of the true religion, or the outcome of divine condescension that took into account the weaknesses of men. The greatest medieval philosopher and theologian, Thomas Aquinas, distinguished natural religion, or that kind of religious truth discoverable by unaided reason, from revealed religion, or religion resting upon divine truth, which he identified exclusively with Christianity. In the 16th century Martin Luther, the great Protestant Reformer, forthrightly labelled the religious views of Muslims, Jews, and Roman Catholic Christians to be false and held the view that the gospel of Christianity understood from the viewpoint of justification by grace through faith was the true standard. In Islām, religions are classified into three groups: the wholly true, the partially true, and the wholly false, corresponding with Islām, the Peoples of the Book (Jews, Christians, and Zoroastrians), and polytheism. The classification is of particular interest because, being based in the Qur'ān, (the Islāmic sacred scripture), it is an integral part of Islāmic teaching, and also because it has legal implications for Muslim treatment of followers of other religions.

Although scientific approaches to religion in the 19th century discouraged use of normative categories, elements of normative judgment were, nonetheless, hidden in certain of the new scientific classifications that had emerged. Many evolutionary schemes developed by anthropologists and other scholars, for example, ranked religions according to their places on a scale of development from the simplest to the most sophisticated, thus expressing an implicit judgment on the religious forms discussed. Such schemes more or less clearly assume the superiority of the religions that were ranked higher (*i.e.*, later and more complex); or, conversely, they serve as a subtle attack on all religion by demonstrating that its origins lie in some of humanity's basest superstitions, believed to come from an early, crude stage. A normative element is also indicated in classification schemes that preserve theological distinctions, such as that between natural and revealed religion. In short, the normative factor still has an important place in the classification of religions and will doubtless always have, since it is extraordinarily difficult to draw precise lines between disciplines primarily devoted to the normative exposition of religion, such as theology and philosophy of religion, and disciplines devoted to its description or scientific study.

Geographical categories

Geographical. A common and relatively simple type of classification is based upon the geographical distribution of religious communities. Those religions found in a single region of the earth are grouped together. Such classifications are found in many textbooks on comparative religion, and they offer a convenient framework for presenting man's religious history. The categories most often used are: (1) Near Eastern religions, including Judaism, Christianity, Islām, Zoroastrianism, and a variety of ancient cults; (2) Far Eastern religions, comprising the religious communities of China, Japan, and Korea, and consisting of Confucianism, Taoism, Mahāyāna ("Greater Vehicle") Buddhism, and Shintō; (3) Indian religions, including early Buddhism, Hinduism, Jainism, and Sikhism, and sometimes also Theravāda Buddhism and the Hindu- and Buddhist-inspired religions of South and Southeast Asia; (4) African religions, or the cults of the tribal peoples of black Africa, but excluding ancient Egyptian religion, which is considered to belong to the ancient Near East; (5) American religions, consisting of the beliefs and practices of the Indian peoples indigenous to the two American continents; (6) Oceanic religions—*i.e.*, the religious systems of the peoples of the Pacific islands, Australia, and New Zealand; (7) classical religions of ancient Greece and Rome and their Hellenistic descendants. The extent and complexity of a geographical classification is limited only by the classifier's knowledge of geography and his desire to seek detail and comprehensiveness in his classification scheme. Relatively crude geographical schemes that distinguish Western religions (usually equivalent to Christianity and Judaism) from Eastern religions are quite common.

Although religions centred in a particular area often have much in common because of historical or genetic connections, geographical classifications present obvious inadequacies. Many religions, including some of the greatest historical importance, are not confined to a single region (*e.g.*, Islām), or do not have their greatest strength in the region of their origins (*e.g.*, Christianity, Buddhism). Further, a single region or continent may be the dwelling place of many different religious communities and viewpoints that range from the most archaic to the most sophisticated. At a more profound level, geographical classifications are unacceptable because they have nothing to do with the essential constitutive elements or inner spirit of religion. The physical location of a religious community reveals little of the specific religious life of the group. Though useful for some purposes, geographical classifications contribute minimally to the task of providing a systematic understanding of man's religions and religiousness.

Use of the comparative study of languages

Ethnographic-linguistic. F. Max Miiller, often called the "Father of the history of religions," stated that "Particularly in the early history of the human intellect, there exists the most intimate relationship between language, religion, and nationality." This insight supplies the basis for a genetic classification of religions (associating them by descent from a common origin), which Miiller believed the most scientific principle possible. According to this theory, in Asia and Europe dwell three great races, the Turanians (including the Ural-Altaic peoples), the Semites, and the Aryans, to which correspond three great families of languages. Originally, in some remote prehistory, each of these races formed a unity, but with the passage of time they split up into a myriad of peoples with a great number of distinct languages. Through careful investigation, however, the original unity may be discerned, including the unity of religion in each case. Miiller's principal resource in developing the resulting classification of religions was the comparative study of languages, from which he sought to demonstrate similarities in the names of deities, the existence of common mythologies, the common occurrence of important terms in religious life. and the likeness of religious ideas and intuitions among the branches of a racial group. His efforts were most successful in the case of the Semites, whose affinities are easy to demonstrate, and probably least successful in the case of the Turanian peoples, whose early origins are hypothetical. Miiller's greatest contribution to scholarship, however, lay in his study of Aryan languages, literatures, and comparative mythology.

Because Müller was a scholar of the first rank and a pioneer in several fields, his ethnographic-linguistic (and genetic) classification of religions has had much influence and has been widely discussed. The classification has value in exhibiting connections that had not been previously observed. Miiller (and his followers) discovered affinities existing among the religious perspectives of both the Aryan and Semitic peoples and set numerous scholars on the path of investigating comparative mythology, thus contributing in a most direct way to the store of knowledge about religions.

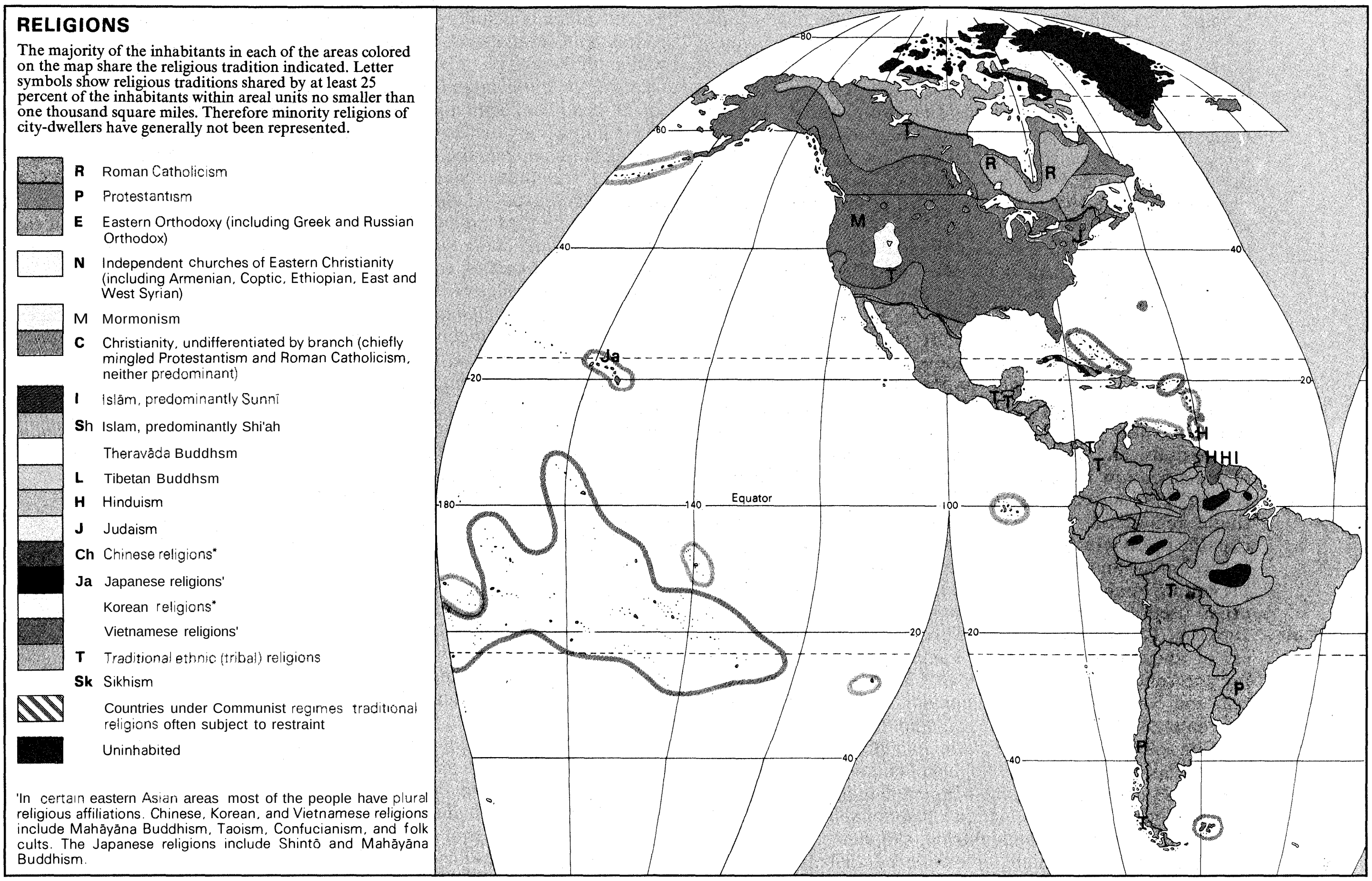
There are, nevertheless, difficulties with the ethnographic-linguistic classification. To begin with, Miiller's evidence was incomplete, a fact that may be overlooked given the state of knowledge in his day. More important is the consideration that peoples of widely differing cultural development and outlook are found within the same racial or linguistic group. Further, the principle of connection among race, language, and religion does not take sufficiently into account the historical element or the possibility of developments that may break this connection, such as the conversion of the Aryan peoples of Europe to a Semitic religion, Christianity.

Other scholars have developed the ethnographic classification of religion to a much higher degree than did Miiller. The German scholar Duren J.H. Ward, for example, in The Classification of Religions (1909) accepted the premise of the connection between race and religion, but appealed to a much more detailed scheme of ethnological relationship. He says: ". . . religion gets its character from the people or race who develop or adopt it . . ." and further that

> . . . the same influences, forces, and isolated circumstances which developed a special race developed at the same time a special religion, which is a necessary constituent element or part of a race . . . .

In order to study religion in its fullness and to bring out with clarity the historical and genetic connections between religious groups, the ethnographic element must thus have adequate treatment. Ward devised a comprehensive "Ethnographico-historical Classification of the Human Races to facilitate the Study of Religions—in five divisions." These major divisions were (1) the Oceanic races, (2) the African races, (3) the American races, (4) the Mongolian races, and (5) the Mediterranean races, each of which has its own peculiar religion. The largest branch, the Mediterranean races, he subdivided into primeval Semites and primeval Aryans, in order to demonstrate in turn how the various Semitic, Indo-Aryan, and European races descended from these original stocks.

Philosophical. The past 150 years have also produced several classifications of religion based on speculative and abstract concepts that serve the purposes of philosophy. The principal example of these is the scheme of G.W.F. Hegel, a seminal German philosopher, in his famous Lectures on the Philosophy of Religion (1832). In general, Hegel's understanding of religion coincided with his philosophical thought; he viewed the whole of


RELIGIONS
The majority of the inhabitants in each of the areas colored on the map share the religious tradition indicated. Letter symbols show religious traditions shared by at least 25 percent of the inhabitants within areal units no smaller than one thousand square miles. Therefore minority religions of city-dwellers have generally not been represented.
R Roman Catholicism
P Protestantism
E Eastern Orthodoxy (including Greek and Russian Orthodox)
N Independent churches of Eastern Christianity (including Armenian, Coptic, Ethiopian, East and West Syrian)
M Mormonism
C Christianity, undifferentiated by branch (chiefly mingled Protestantism and Roman Catholicism, neither predominant)
I Islām, predominantly Sunnī
Sh Islam, predominantly Shi'ah
Theravāda Buddhsm
L Tibetan Buddhsm
H Hinduism
J Judaism
Ch Chinese religions*
Ja Japanese religions*
Korean religions*
Vietnamese religions*
T Traditional ethnic (tribal) religions
Sk Sikhism
Countries under Communist regimes traditional religions often subject to restraint
Uninhabited
*In certain eastern Asian areas most of the people have plural religious affiliations. Chinese, Korean, and Vietnamese religions include Mahāyāna Buddhism, Taoism, Confucianism, and folk cults. The Japanese religions include Shintō and Mahāyāna Buddhism.
80
60
40
20
180
140
100
Equator
20
40
T
R
R
M
J
Ja
TT
T
T
H
HHI
T
P
P

Geographical distribution of the religions of the world.
Old World religions from D. Sopher, Geography of *Religions* (© 1967)

Hegelian dialectics

human history as a vast dialectical movement toward the realization of freedom. The reality of history, he held, is Spirit, and the story of religion is the process by which Spirit—true to its own internal logical character and following the dialectical pattern of thesis, antithesis, and synthesis (the reconciliation of the tension of opposite positions in a new unity that forms the basis of a further tension)—comes to full consciousness of itself. Individual religions thus represent stages in a process of evolution (*i.e.*, progressive steps in the unfolding of Spirit) directed toward the great goal at which all history aims.

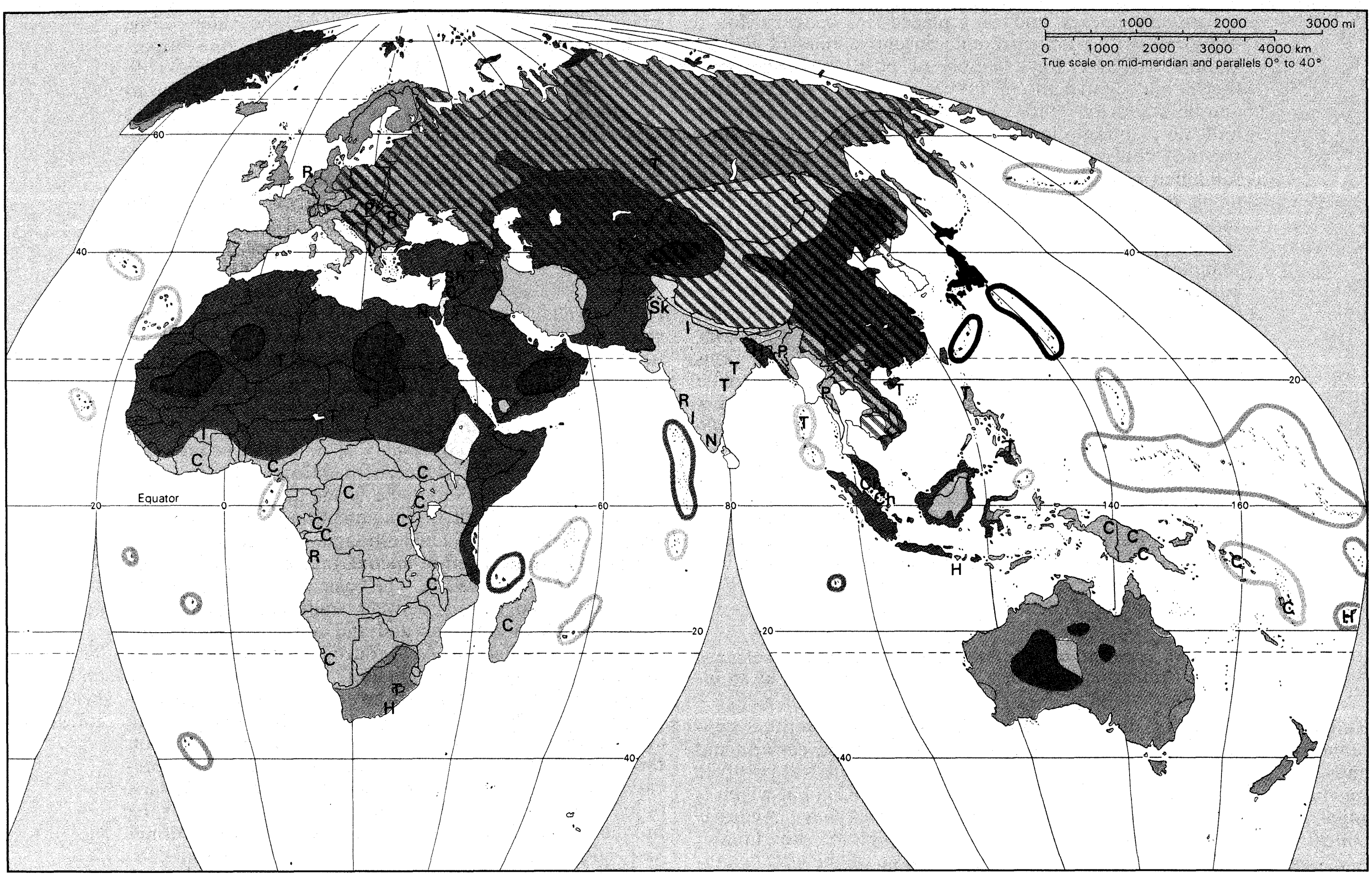
Hegel classified religions according to the role that they have played in the self-realization of Spirit. The historical religions fall into three great divisions, corresponding with the stages of the dialectical progression. At the lowest level of development, according to Hegel, are the religions of nature, or religions based principally upon the immediate consciousness deriving from sense experience. They include: immediate religion or magic at the lowest level: religions, such as those of China and India plus Buddhism, that represent a division of consciousness within itself; and others, such as the religions of ancient Persia, Syria, and Egypt, that form a transition to the next type. At an intermediate level are the religions of spiritual individuality, among which Hegel placed Judaism (the religion of sublimity), ancient Greek religion (the religion of beauty), and ancient Roman religion (the religion of utility). At the highest level is absolute religion, or the religion of complete spirituality, which Hegel identified with Christianity. The progression thus proceeds from man immersed in nature and functioning only at the level of sensual consciousness, to man becoming conscious of himself in his individuality as distinct from nature, and beyond that to a grand awareness in which the opposition of individuality and nature is overcome in the realization of Absolute Spirit.

Many criticisms have been offered of Hegel's classification. An immediately noticeable shortcoming is the failure to make a place for Islām, one of the major historical religious communities, which numbers one-seventh of the human race among its constituents. The classification is also questionable for its assumption of continuous development in history. The notion of perpetual progress is not only doubtful in itself but is also compromised as a principle of classification because of its value implications.

Nevertheless, Hegel's scheme was influential and was adapted and modified by a generation of philosophers of religion in the Idealist tradition. Departure from Hegel's scheme, however, may be seen in the works of Otto Pfleiderer, a German theologian of the 19th century, who held a particular interest in problems of classification. Pfleiderer believed it impossible to achieve a significant grouping of religions unless, as a necessary preliminary condition, the essence of religion were first isolated and clearly understood. Essence is a philosophical concept, however, not a historical one. Pfleiderer considered it indispensable to have conceptual clarity about the underlying and underived basis of religion from which all else in religious life follows. In ***Die Religion, ihr Wesen und ihre Geschickte*** ("Religion, Its Essence and History"), Pfleiderer held that the essence of religious consciousness exhibits two elements, or moments, perpetually in tension with one another: one of freedom and one of dependence, with a number of different kinds of relationships between these two. One or the other may predominate, or they may be mixed in varying degrees.

Religions of freedom and dependence

Pfleiderer derived his classification of religions from the relationships between these basic elements. He distinguished one great group of religions that exhibits extreme partiality for one over against the other. The religions in which the sense of dependence is virtually exclusive are those of the ancient Semites, the Egyptians, and the Chinese. Opposite these are the early Indian, Germanic, and Greek and Roman religions, in which the sense of freedom prevails. The religion of this group may also be



seen in a different way, as nature religions in the less-developed cultures or as culture or humanitarian religions in the more advanced. A second group of religions exhibits a recognition of both elements of religion, but gives them unequal value. These religions are called supernatural religions. Among them Zoroastrianism gives more weight to freedom as a factor in its piety, and Brahmanism and Buddhism are judged to have a stronger sense of dependence. The last group of religions is the monotheistic religions: Islam, Judaism, and Christianity, which are divided again into two sub-groups, *i.e.*, those that achieve an exact balance of the elements of religion and those that achieve a blending and merging of the elements. Both Judaism and Islam grant the importance of the two poles of piety, though there is a slight tendency in Islam toward the element of dependence and in Judaism toward freedom. It is Christianity alone, he claimed, that accomplishes the blending of the two, realizing both together in their fullness, the one through the other.

The intellectual heritage that lies behind this classification will be immediately apparent. The classification reflects its time (19th century) and place (western Europe) of conception in the sense that the study of religion was not yet liberated from its ties to the philosophy of religion and theology.

Morphological. Considerable progress toward more scientific classifications of religions was marked by the emergence of morphological (evolution of religious forms) schemes. These classifications assume that religion in its history has passed through a series of discernible stages of development, each having readily identifiable characteristics and each constituting an advance beyond the former stage. So essential is the notion of progressive development to morphological schemes that they might also be called evolutionary classifications. Recent trends in the comparative study of religions, however, have retained the interest in morphology but have decisively rejected the almost universal 19th-century assumption of unitary evolution in the history of religion. The crude expression of evolutionary categories such as the division of religions into lower and higher or primitive and higher religions has been subjected to especially severe criticism.

Influence of E.B. Tylor

The pioneer of morphological classifications was E.B. Tylor, a British anthropologist, whose *Primitive Culture* (1871) is among the most influential books ever written in its field. Tylor developed the thesis of animism, a view that the essential element in all religion is belief in spiritual beings. According to Tylor, the belief arises naturally from elements universal in human experience (*e.g.*, the phenomena of death, sleep, dreams, trances, and hallucinations), and leads through processes of primitive logic to the belief in a spiritual reality distinct from the body and capable of leading an independent existence. In the development of the idea, this reality is identified with the breath and the life principle; thus arises the belief in the soul, in phantoms, and in ghosts. At a higher stage, the spiritual principle is attributed to aspects of reality other than man, and all things are believed to possess spirits that are their effective and animating elements; it is the spirits, for example, that primitive men generally believe to cause sickness and control their destinies.

Of immediate interest is the classification of religions drawn from E.B. Tylor's animistic thesis. Ancestor worship, a common occurrence among less developed peoples, is obeisance to the spirits of the dead. Fetishism, the veneration of objects believed to have magical or supernatural potency, springs from the association of spirits with particular places or things and leads in turn to idolatry, in which the image is looked upon as the symbol of a spiritual being or deity. Totemism, a system of belief in which there is an association between particular groups of people and certain spirits that serve as guardians of those peoples, arises when the entire world is conceived as peopled by spiritual beings. At a still higher stage, polytheism, the interest in particular deities or spirits, disappears and is replaced by concern for a "species" deity who represents an entire class of similar spiritual realities. By a variety of possible means polytheism may evolve into monotheism, a belief in a supreme and unique deity. Tylor's theory of the nature of religions and the resultant classification were so logical, convincing, and comprehensive that for a number of years they remained virtually unchallenged.

The morphological classification of religions received more sophisticated expression from C.P. Tiele, a 19th-century Dutch scholar and an important pioneer in the scientific study of religion. His point of departure was a pair of distinctions made by the philosophers of religion A.K. Kuenen and W.D. Whitney. In the Hibbert Lectures for 1882, *National Religions and Universal Religions,* Kuenen had emphasized the difference between religions limited to a particular people and those that have taken root among many peoples and qualitatively aim at becoming universal. Whitney saw the most marked distinction among religions as being between race religions ("the collective product of the wisdom of a community") and individually founded religions. The first are the result of nature's unconscious working through long periods of time, and the latter are characterized by a high degree of ethical awareness. Tiele agreed strongly with Whitney in distinguishing between nature and ethical religions. Ethical religion, in Tiele's views, develops out of nature religion,

> But the substitution of ethical religions for nature-religions is, as a rule, the result of a revolution; or at least of an intentional reform.

Religions of nature and spiritualistic–ethical religions

Each of these categories (*i.e.*, nature or spiritualistic–ethical) may be further subdivided. At the earliest and lowest stage of man's spiritual development was polyzoic religion, about which there is no information but which is based on Tiele's theory that man must have regarded natural phenomena as endowed with life and superhuman magical power. The first known stage of the nature religions is called polydaemonistic (many spirits) magical religion, which is dominated by animism and characterized by a confused mythology, a firm faith in magic, and the pre-eminence of fear above other religious emotions. At a higher stage, but still in the type of nature religions, occurs therianthropic polytheism, referring to the form of the deities that are normally of mixed animal and human composition. The highest stage of nature religion is anthropomorphic polytheism, in which the deities appear in human form but have superhuman powers. These religions have some ethical elements, but their mythology portrays the deities as indulging in all sorts of shocking acts. None of the polytheistic religions, thus, was able to raise itself to a truly ethical point of view.

In their further division the ethical religions fall into two subcategories. First are the national nomistic (legal) religions that are particularistic, limited to the horizon of one people only and based upon a sacred law drawn from sacred books. Above them are the universalistic religions, qualitatively different in kind, aspiring to be accepted by all men, and based upon abstract principles and maxims. In both subtypes of ethical religion, doctrines and teachings are associated with the careers of distinct personalities who play important roles in their origin and formation. Tiele could find only three examples of this highest type of religion: Islām, Christianity, and Buddhism.

Influence of Nathan Soderblom

Tiele's classification enjoyed a great vogue and influenced many who came after him. Nathan Soderblom, a Swedish archbishop who devoted much energy to problems of classification, accepted the division of higher religions into two great groups but used a varied terminology that pointed to some of the characteristics of the two types of religion. In addition to natural religion and revealed religion, or religions of nature and religions of revelation, Soderblom spoke of culture religions and prophetic religions, of culture religions and founded religions, and of nature religions and historical religions. The highest expression of the first category is the "mysticism of infinity" that is characteristic of the higher aspects of Hindu and Buddhist religious experience. The apex of genuine prophetic religion is reached in the "mysticism of personality." All these distinctions mean the same thing, and all are indebted to Tiele's thought. In one important matter, however, Soderblom sharply disagreed with Tiele, for he could not accept the thesis of continuous development in the history of religion. In Soderblom's view, the line between nature religion and prophetic religion is a deep and unbridgeable chasm, a qualitative difference so enormous that one type could never evolve by natural historical processes into the other. Prophetic religion can be explained only as a radical and utterly new incursion into history. Soderblom, it must be remembered, was a churchman and theologian as well as a distinguished historian of religions, and there is without doubt an element of theological judgment influencing his stand on this matter. Soderblom was eager to defend the uniqueness of biblical religion, and he believed that his historical and scientific studies provided an objective basis for asserting, not only the uniqueness, but also the superiority of Christianity.

Tiele's enduring influence may also be seen in the classification of religions advanced by Mircea Eliade, a Romanian–American scholar who is one of the most prolific contemporary students of religion. Eliade, who in other respects might be considered among the phenomenologists of religion, is interested in uncovering the "structures" or "patterns" of religious life. The basic division of religions that Eliade recognizes is between traditional religions, in which primitive religions and the archaic cults of the ancient civilizations of Asia, Europe, and America are included, and historical religions. The meaning of the distinction is better revealed, however, in the terms cosmic religion and historical religion. In Eliade's estimation, all of traditional religion shares a common outlook upon the world. The chief element in this outlook is the deprecation of history and the rejection of profane, mundane time. Religiously, traditional man is not interested in the unique and specific, but rather exclusively in those things and actions that repeat and restore transcendental models. Only those things that participate in and reflect the eternal archetypes or the great pattern of original creation by which cosmos came out of chaos are real in the traditional outlook. The religious activities of traditional man are the recurring attempts to return to the beginning, to the Great Time, to trace again and renew the process by which the structure and order of the cosmos were established. Traditional religions may, therefore, find the sacred in any aspect of the world that links man to the archetypes of the time in the beginning; their typical mode of expression is in consequence repetitive. Further, their understanding of history, as far as they are concerned with it at all, is cyclical. The world and what happens in it are devalued, except as they show forth the eternal pattern of the original creation.

Traditional and historical religions

Modem, postarchaic, or historical religions (*e.g.*, Judaism, Christianity, Islām) show markedly other features. They tend to see a discontinuity between God and the world and to locate the sacred not in the cosmos but somewhere beyond it. Moreover, they hold to linear views of history, believing it to have a beginning and an end, with a definite goal as its climax and to be by nature unrepeatable. Thus, the historical religions are world affirming in the double sense of believing in the reality of the world and of believing that meaning for man is worked out in the historical process. By reason of these views, the historical religions alone have been monotheistic and exclusivist in their theologies. Although Eliade outstrips his predecessors in delineating the qualities of traditional religion in particular, much of his thought was anticipated in Soderblom's descriptions of nature religion and prophetic religion.

Phenomenological. The principles thus far discussed have all had reference to the classification of religions in the sense of establishing groupings among historical religious communities having certain elements in common. Attempts have been made to deal with entire religions or religious communities as the units to be classified. In recent times, however, the interest in classifying entire religions has markedly declined, partly because of an emerging interest in the phenomenology of religion.

urposes f the henomnology f religion

This new trend in studies, which dominates the field at the present, claims its origin in the Phenomenological philosophy of Edmund Husserl, a German Jewish–Lutheran scholar, and has found its greatest exponents in The Netherlands. Phenomenology of religion has at least two aspects. It is first of all an effort at devising a taxonomic (classificatory) scheme that will permit the comprehensive cataloging and classifying of religious phenomena across the lines of religious communities, but it is also a method that aims at revealing the self-interpretation by religious men of their own religious responses. Phenomenology of religion thus rejects any overview of religion that would interpret religion's development as a whole, confining itself rather to the phenomena and the unfolding of their meaning for religious men. Phenomenologists are especially vigorous in repudiating the evolutionary schemes of past scholars; whom they accuse of imposing arbitrary semiphilosophical concepts in their interpretation of the history of religion. Phenomenologists also have little interest in history for its own sake, except as a preliminary stage of material gathering for the hermeneutical (critical–interpretive) task that is to follow.

One of the earliest Dutch Phenomenologists, W. Brede Kristensen (1867–1953), spoke of his work as follows:

> Phenomenology of Religion attempts to understand religious phenomena by classifying them into groups . . . . we must group the phenomena according to characteristics which correspond as far as possible to the essential and typical elements of religion . . . .

The material with which Phenomenology is concerned is all the different types of religious thinking and action, ideas about divinity, and cultic acts. Kristensen's systematic organization of religious phenomena may be seen in the table of contents of his *Meaning of Religion* in which he divides his presentation of material into discussions of: (1) cosmology, which includes worship of nature in the form of sky and earth deities, animal worship, totemism, and animism; (2) anthropology, made up of a variety of considerations on the nature of man, his life, and his associations in society; (3) cultus, which involves consideration of sacred places, sacred times, and sacred images; and (4) cultic acts, such as prayer, oaths and curses, ordeals, and other acts. Kristensen was not concerned with the historical development, or the description of a particular religion, or even a series of religions, but rather with grouping the typical elements of the entire religious life, no matter in what community they might occur.

ignifiance f G. van er Leeuw

Probably the best known Phenomenologist is G. van der Leeuw. another Dutch scholar. In his ***Religion in Essence and Manifestation,*** van der Leeuw categorized the material of religious life under the following headings: (1) the object of religion, or that which evokes the religious response; (2) the subject of religion, in which there are three divisions: the sacred man, the sacred community, and the sacred within man, or the soul; (3) object and subject in their reciprocal operation both as regards outward reaction and inward action; (4) the world, ways to the world, and the goals of the world; and (5) forms, which must take into account religions and the founders of religions. Van der Leeuw was not interested in grouping religious communities as such, but rather in laying out the types of religious expression. He discussed distinct religions only because religion in the abstract has no existence. He classified religions according to 12 forms: (1) religions of remoteness and flight (ancient China and 18th-century deism); (2) religion of struggle (Zoroastrianism); (3) religion of repose, which has no specific historical form but is found in every religion in the form of mysticism; (4) religion of unrest or theism, which again has no specific form but is found in many religions; (5) dynamic of religions in relation to other religions (syncretism and missions); (6) dynamic of religions in terms of internal developments (revivals and reformations); (7) religion of strain and form, the first that van der Leeuw characterizes as one of the "great" forms of religion (Greece); (8) religion of infinity and of asceticism (Indian religions, but excluding Buddhism); (9) religion of nothingness and compassion (Buddhism); (10) religion of will and of obedience (Israel); (II) the religion of majesty and humility (Islām); and (12) the religion of love (Christianity). It is apparent that the above is not a classification of religions as organized systems. Categories 3, 4, 5, and 6 relate to elements found in many if not all historical religious communities, and the categories from 7 onward are not classifications but are attempts to characterize particular communities by short phrases that express what van der Leeuw considered to be their essential spirit. This classification leaves out of account altogether the "primitive" religions of less developed peoples.

Psychological and sociological types

Other principles. William James, the American philosopher and psychologist, in his well-known book ***The Varieties of Religious Experience,*** differentiated two types of religion according to the attitude of the individual toward life. There is the religion of healthy-mindedness, which minimizes or ignores the evil of existence, and the religion of morbid-mindedness, which considers evil as the very essence of life. Max Weber, a German sociologist, called attention to the differences between religions that express themselves primarily in mythopoetic ways and those that express themselves in rational forms. The distinction comes very close to that between traditional and historical religions, though its emphasis is somewhat different.

Nathan Soderblom, in his prolific scholarly career, devised several classifications other than the principal one discussed above. In his great work on primitive religions, ***Das Werden des Gottesglaubens*** ("Development of the Belief in God"), Soderblom divided religions into dynamistic, animistic, and theistic types according to the way primitive peoples apprehend the divine. In others of his works (***Einführung in die Religionsgeschichte,*** or "Introduction to the History of Religion," and ***Thieles Kompendium der Religionsgeschichte neu bearbeitet,*** or "Tiele's Compendium of the History of Religion Revised") he contended that Christianity is the central point of the entire history of religions and, therefore, classified religions according to the historical order in which they came into contact with Christianity. A somewhat similar approach was adopted by Albert Schweitzer, the French theologian, medical missionary, and Nobel laureate, in ***Christianity arid the Religions of the World,*** in which he grouped religions as rivals or nonrivals of Christianity. Still another scheme may be seen in Söderblom's Gifford Lectures, ***The Living God,*** in which religions were divided according to their doctrines of the relation between human and divine activity in the achievement of salvation. Thus, among higher religions there are those in which man alone is responsible for salvation (Buddhism), God alone is responsible (the Bhakti cults of India), or God and man cooperate (Christianity).

Attention may also be called to a recent classification of religions by the American sociologist, Robert Bellah. Having in mind the advances of the social sciences in their understanding of religions, Bellah offers a refurbished and more highly sophisticated version of an evolutionary scheme that he thinks to be the most satisfactory possible in the present state of scholarly knowledge. He views religion in its evolution as having passed through five stages, beginning with the primitive, and proceeding through the archaic, the historical, and the early modern to the modern stage. The religious complexes that emerge in each stage of this evolution have identifiable characteristics that Bellah studies and differentiates according to the following categories: symbol systems, religious actions, religious organizations, and social implications. Two basic concepts run through Bellah's classification, providing the instruments for the division of religions along the evolutionary scale. The first is that of the increasing complexity of symbolization as one moves from the bottom to the top of the scale, and the second is that of increasing freedom of personality and society from their environing circumstances or, in other words, the growing secularization of the religious field. Bellah's classification is important because of the wide discussion it has awakened among social scientists.

In addition to those already listed, one may find classifications based upon the content of religious ideas, the forms of religious teaching. the nature of cultus, the character of piety, the nature of the emotional involvement in religion, the character of the good toward which religions strive, the relations of religions to the state, to art, to science, to morality, and a host of others.

CONCLUSION

The classification of religions that will withstand all criticism and serve all the purposes of a general science of religions has not yet been devised. Each of the classifications presented above has been attacked for its inadequacies or distortions, yet each is useful in bringing to light certain aspects of religion. Even the crudest and most subjective classifications throw into relief various aspects of religious life and contribute in this way to the cause of understanding. At the present state of advance in this matter the most fruitful approach for a student of religion appears to be that of employing a number of diverse classifications of religions, each one for the insight that it may yield. Though each may have its shortcomings that must be recognized, each also offers a positive contribution to the store of knowledge and its systematization. The error that must be avoided is that of insisting upon the exclusive validity of any single taxonomic effort. To confine oneself to a single determined framework of thought about so rich and variegated a subject as religion is to risk the danger of missing much that is important. Classification should be viewed as a method and a tool only.

Criteria for making and judging classifications

Although a perfect classification lies at present beyond scholars' grasp, certain criteria may be suggested for building and judging classifications. These are both positive and negative in nature. First, it may be insisted that classifications should not be arbitrary, subjective, or provincial. A first principle of the scientific method is that it should strive for objectivity to the extent possible and that its findings should be capable of confirmation by other observers. Second, an acceptable classification should deal with the essential and typical in the religious life, not with the accidental and the unimportant. The contribution to understanding that a classification may make is in direct proportion to the penetration of the bases of religious life exhibited in its principles of division. A good classification must concern itself with the fundamentals of religion and with the most typical elements of the units it is seeking to order. Third, it may be held that a proper classification should be capable of presenting both that which is common to religious forms of a given type and that which is peculiar or unique to each member of the type. No classification can serve its proper purpose if it ignores the concrete historical individuality of religious manifestations in favour of that which is common to them all, nor can it be approved if it neglects to demonstrate the common factors that are the bases for the very distinction of types of religious experience, manifestations, and forms. Classification of religions involves both the systematic and the historical tasks of the general science of religion. Fourth, it is desirable in a classification that it demonstrate the dynamics of religious life both in the recognition that religions as living systems are constantly changing and in the effort to show, through the categories chosen, how it is possible for one religious form or manifestation to develop into another. Few errors have been more damaging to the understanding of religion than that of viewing religious systems as static and fixed, as, in effect, ahistorical. Adequate classifications should possess the flexibility to come to terms with the flexibility of religion itself. Fifth, a classification must solve the problem of what exactly it is that is to be classified. If the purpose is to develop types of religions as a whole, the questions of what constitutes a religion and what constitutes various individual religions must be asked. Since no historical manifestation of religion is known that has not exhibited an unvarying process of change, evolution, and development, these questions are far from easily solved. With such criteria in mind it should be possible continuously to construct classification schemes that illuminate man's religious history.

BIBLIOGRAPHY. Two monographs dealing specifically with the classification of religions, each of which offers a survey of previous classifications in addition to the author's own scheme, are: DUREN J.H. WARD; *The Classification of Religions: Different Methods, Their Advantages and Disadvantages* (1909); and FRED LOUIS PARRISH, *The Classification of Religions: Its Relation to the History of Religions* (1941), containing a full survey of classification schemes with brief characterizations of each and the best bibliographical guide for one who wishes to pursue the subject in depth. Other books useful for further study are as follows: P.D. CHANTEPIE de la SAUSSAYE, *Lehrbuch der Religionsgeschichte,* 2 vol. (1887–89; Eng. trans. of vol. 1, *Manual of the Science of Religion,* 1891), which includes some pages on classification problems at the beginning of vol. 1; C.P. TIELE, *Elements of the Science of Religion,* 2 vol. (1897–99), a classic work by an important scholar on this subject; and F. MAX MUELLER, *Introduction to the Science of Religion* (1873), another classic work. Of more recent origin is GUSTAV MENSCHING, *Die Religion: Erscheinungsformen, Strukturtypen und Lebensgesetze* (1959), a popular manual of the history of religions that includes a long section on classification problems.

(C.J.A.)

Religious Dress and Vestments

The meaning religious dress and vestments

Religious dress and vestments, broadly understood, include a wide range of attire, accoutrements, and markings used in religious rituals that may be corporate, domestic, or personal in nature. They comprise types of coverings all the way from the highly symbolic and ornamented eucharistic (Holy Communion) vestments of Eastern Orthodox Christianity to tattooing, scarification, or body painting of members of primitive (preliterate) societies. Some types of religious dress may be used to distinguish the priestly from the lay members of a religious group, or they may also be used to signify various orders or ranks within a priesthood. Some religious communities may require that religious personages (*e.g.*, priests, monks, nuns, shamans, priestesses, and others) garb themselves with appropriate types of religious dress at all times, whereas other religious communities may only request that religious dress be worn during rituals.

In theocratic societies, such as Judaism and Islām, religious sanctions govern what may and may not be worn by members of the community; and religious dress embraces not only what is worn by a prayer leader but also what is worn by his congregation outside as well as inside a place of worship. In many traditions, habits serve to identify monastic groups. Indeed, in the latter case, the function of religious dress is more akin to heraldry as a form of symbolic identification than to liturgy, with its ritualistic symbolic motifs.

In a more restricted sense, religious vestments articulate a liturgical language as part of a figurative idiom shared with other religious symbols—*e.g.*, icons (images), statues, drama, music, and ritual. According to the richness of the liturgical or ritual vocabulary employed, the more feasibly can a symbology of vesture be attempted. This is especially the case with Eastern Orthodoxy, whose predilection for symbolical theology has spread from sacraments to sacramentals and everything associated with worship, including dress. With allegory paramount in the Middle Ages, the Western Church could not escape attributing symbolical values to garments whose origin may have owed little to symbolism. From the liturgical writer Amalarius of Metz in the 9th century to the theologian Durandus of Saint-Pourçain in the 13th–14th century sacerdotal vestments, in particular the stole and the chasuble, were viewed as symbols and indeed operated as such in a way that still influences current usage. Thus, because the stole is a yoke around the neck of the priest and he should rejoice in his servitude, on donning or doffing it he kisses the emblem of his servile status.

The notion of dress as a substitute skin and, hence, as an acquired personality temporarily assumed has been widespread in primitive religion; such practices in shamanism have been widely observed in Arctic and Siberian regions. The use of a substitute skin in religious ritual is also explicit in the cultic actions of some advanced cultures, such as in the rite of the Aztec maize goddess Chicomecoátl. A virgin chosen to represent Chicomecoátl, after having danced for 24 hours, was then sacrificed and flayed; and the celebrant, dressed in her skin, re-enacted the same ritual dance to identify with the victim, who was viewed as the goddess.

Religious dress may also serve a memorial function, as in the case of the religious leaders (mullahs) of the Shi'ites (Muslim members of the party of 'Ali), whose black gowns allude to the sufferings of Husayn ('Ali's son by Fāṭimah, Muhammad's only surviving daughter), who was martyred at Karbalā', in modern Iraq, in AD 680. In the Eucharist, which is both a thanksgiving and a re-enactment of the sacrifice of Christ on Golgotha, the chasuble (outer garment) worn by the celebrant depicts scenes from the Passion on the orphrey, the name given to the elaborately embroidered strips stitched on the chasuble. The fringes on the Jewish prayer shawl witness to "the commandments of the Lord" in Numbers, chapter 15, and remind the worshipper that he has covenanted to observe them.

TYPES OF DRESS AND VESTMENTS IN WESTERN RELIGIONS

Judaism. Early sacerdotal dress. Jewish vesture is an amalgam of very ancient and extremely modern religious dress, Originally, sacerdotal dress was probably varied and complex, but, after the destruction of the Second Temple in AD 70 and the subsequent disappearance of the Temple offices, many garments associated with priestly functions passed into oblivion. Chief among these offices was that of the high priest. In addition to the usual Levitical garments (those of the priestly class), the high priest, while officiating, wore the me'il (mantle), the ephod (an upper garment), a breastplate, and a headdress. The *meʿil* was a sleeveless robe of purple the lower hem of which had a fringe of small gold bells alternating with pomegranate tassels in red, scarlet, purple, and violet. The ephod—an object of much controversy—probably consisted of a wide band of material with a belt to secure it to the body, and it was worn over the other priestly garments. Most important was the breastplate (hoshen), which was square in outline and probably served as a pouch in which the divinatory devices of Urim and Thummim were kept. Exodus, chapter 28, verse 15, specifies that it was to be woven of golden and linen threads dyed blue, purple, and scarlet. Because of its oracular function, it was called the "breastpiece of judgment." On the face of the breastplate were set 12 gems in four rows, symbolizing the 12 tribes of Israel. These stones were a sardius, a topaz, and a carbuncle in the first row; an emerald, a sapphire, and a diamond in the second; a jacinth, an agate, and an amethyst in the third; and a beryl, an onyx, and a jasper in the fourth. The identity, sequence, and objects of representation of these stones are matters of controversy. Worn over the ephod, the breastplate was slung from the shoulders of the wearer by golden attachments. On his head the high priest usually wore a mitzenfet (either a tiara or a turban), except on Yom Kippur ("Day of Atonement"), when he wore nothing but white linen garments upon entering the Holy of Holies (the inner sanctuary).

High priestly garments

Later religious dress. Later religious dress of Judaism after the fall of the Temple in AD 70 reflects usages that predate that event but were continued in Judaism at the synagogue. Included among such garments are *tefillin* (phylacteries) and tzitzit (fringes), which have certain features in common. The name phylacteries is sometimes thought to point to a prophylactic origin, but the term is actually a translation of the Hebrew word for "frontlets" (*ṭoṭafot*). Phylacteries are worn in obedience to the commandment found in Deuteronomy, chapter 11, verse 18, and Exodus, chapter 13, verses 9 and 16: "And you shall bind them [*i.e.*, the words of God] as a sign upon your hand, and they shall be as frontlets between your eyes." This implies that there should be two phylacteries: one to be worn on the arm, the other on the head. Both kinds consist of a small black box of hide containing a manuscript and are secured to the respective parts of the body by leather thongs. On the sides of the head *tefilla* is the Hebrew letter ש, the first letter of Shaddai (Almighty). Both boxes are secured by leather thongs. The practice can be dated at least as far back as the 3rd century BC. The knotted thongs indicate a prophylactic purpose—*i.e.*, to protect the wearer against demons. Likewise, the wearer of these objects was, for the prayer's duration, under the protection of the Almighty, whose name he bore. The importance of knots in Semitic magic is also alluded to in the Qur'ān (the Islāmic scripture).

The use of phylacteries and fringes

Something similar obtains in the case of the tzitzit (fringes), or "twisted cords." The wearing of fringes is in obedience to a commandment in Numbers, chapter 15, verses 38–40: "It shall be to you a tassel to look upon and remember all the commandments of the Lord, [and] to do them." The fringes were attached to the outer garment with no attempt at or reason for concealment. Later, because of persecution, they became an inner garment, enabling the wearer to observe the Law clandestinely. This garment, which is not entirely obsolete, is styled *arbaʿkanfot* ("four corners") in allusion to Deuteronomy, chapter 22, verse 12 ("you shall make yourself tassels on the four corners of your cloak with which you cover yourself"), although no literary reference to its use can be traced further back than the 14th century.

The tallith, or prayer shawl, has the four fringes also, but it is confined to synagogal use and, even there, is limited to the morning service, whereas the arba'kanfot is worn all day. Both silk and wool are used, but the woollen tallith is preferable, with white as its ground colour. In the 20th century the tallith is worn like a scarf and is sometimes pulled over the head to aid in concentrating during prayer. Formerly, however, it was always wrapped around the head. In orthodox Judaism, the head is invariably covered during worship, usually by a skullcap known as a yarmulka or *kappel.*

Because a Jewish male is not supposed to walk more than four cubits (six feet) with his head uncovered, a religious Jew will wear the skullcap clipped to his hair, and indeed he may wear it all day because he believes himself to be in the presence of God at all times.

The dress of rabbis never conformed to precise standards. Current practice approximates modern Genevan (Protestant) conventions (gown and bands). The Jewish Reform movement, which began in Germany, further emphasized the Protestant character of rabbinical dress, and Reform rabbis differ little in this respect from ministers of various Protestant churches. Both cantor (*ḥaz*zan) and rabbi now use the black gown and round black hat, which came into use during the 19th century.

Garments used on Yorn Kippur

On Yom Kippur, it was the custom for participants to wear a sargenes, or white garment, emphasizing that Yom Kippur was an occasion not only of repentance but also of grace, for which festal wear was appropriate. Emphasis on the atoning aspect of the occasion, however, led to the sargenes being interpreted as takhrikhim, or graveclothes, which are worn to aid the worshipper toward a mood of repentance, a practice also adopted by the hazzan on two other occasions and by the host at the seder (meal) on Passover (a feast celebrating the Exodus of the Hebrews from Egypt in the 13th century BC). Officiants at the Yom Kippur service still dress in white robes. Shrouds are normally of unadorned white linen, following the sumptuary ruling of the 1st-century-ADrabbi Gamaliel the Elder. To the shroud may be added the tallith used by the deceased, but with the fringes removed or cut, because the prescription governing their use applies only to the living. Both liturgical vesture and everyday clothing must conform to the Mosaic requirement that forbids the combination of linen and wool in the same garment (see also JUDAISM).

Christianity. In the pre-Constantinian church (before the early 4th century), no distinctive liturgical dress was worn, and the Eucharist (Holy Communion) was celebrated by priests whose dress did not differ from that worn by lay members of their congregations. Present liturgical vestments in Roman Catholic and Orthodox churches derive from a common origin—*i.e.*, the garments that were fashionable in the late Roman Empire. After the Schism of 1054, however, they each followed separate courses (see also CHRISTIANITY).

Roman Catholic religious dress. A distinction is made between the insignia of ecclesiastical and sacerdotal office in the hierarchy and the functionally and symbolically significant liturgical robes. After the barbarian invasions of the Roman Empire from the 4th century on, fashions in secular dress changed, and thus the clergy became distinct in matters of dress from the laity. Certain robes indicate a position in the hierarchy; others correspond to function and may be worn by the same individual at different times. The most important vestment among the insignia is the stole, the emblem of sacerdotal status, the origin of which is the ancient *pallium.* The stole originally was a draped garment, then a folded one with the appearance of a scarf, and, finally, in the 4th century, a scarf. As a symbol of jurisdiction in the Roman Empire, the supreme pontiff (the pope, or bishop of Rome) conferred it upon archbishops and, later, upon bishops, as emblematic of their sharing in the papal authority.

Liturgical garb

The distinctive garb of the liturgical celebrant is the chasuble, a vestment that goes back to the Roman *paenula.* The *paenula* also was the Orthodox equivalent of the chasuble, the *phelonion*, and perhaps also the cope (a long mantle-like vestment). In its primitive form the *paenula* was a cone-shaped dress with an opening at the apex to admit the head. Because ancient looms were not wide enough to make the complete garment, it was made in several parts sewn together with strips covering the seams. These strips, of contrasting material, developed into the orphrey (embroidery), on which much attention was later lavished. Next in the hierarchical order after the priesthood were the diaconate and subdiaconate, whose characteristic vestments were, respectively, the dalmatic (*dalmatica*), a loose-fitting robe with open sides and wide sleeves, and the tunic (*tunica*), a loose gown. A priest wore all three, one over another. Under these he wore the alb (a long white vestment), held round the waist by a girdle, and around the neck the amice (a square or oblong, white linen cloth), with the maniple (originally a handkerchief) on the left arm. Although the deacon used a stole, the subdeacon did not. In the formative period of liturgical dress, these practices were in the process of becoming normative. During the 9th to the 13th century the norms now familiar were established. The chasuble became an exclusively eucharistic garment; the cope, excluded from the Eucharist, became an all-purpose festive garment.

Next in importance to the chasuble is the cope, a garment not worn during the celebration of the mass but rather a processional vestment. It is worn by the celebrant for rites of a non-eucharistic character, such as the Asperges, a rite of sprinkling water on the faithful preceding the mass. The origins of the cope are not known for certain by liturgical scholars. According to one theory, it derives from the open-fronted *paenula,* just as the chasuble derives from the closed version of the same garment. (The subsequent wide divergence between the two vestments need not preclude a common origin.) Unlike the chasuble, the form of which has never stopped changing, the evolution of the cope was complete before the end of the Middle Ages. Cope chests, based on the quadrant of a circle and designed to preserve the embroidered surfaces by keeping the copes flat, were a common feature of medieval cathedrals. When it is worn, the two sides of the garment are held together by a morse (a metal clasp). The cope occupied an intermediate position between liturgical and nonliturgical vestments, the most important of which was the cassock, the normal dress of the priesthood outside church ceremonies. When engaged in religious ceremonies, the officiant would wear the liturgical vestments over his cassock.

The tiara, the papal diadem or crown apostolic, emerged in the early medieval period; and the mitre (the liturgical headdress of bishops and abbots), the most conspicuous of the episcopal insignia, began as a mark of favour accorded to certain bishops by the supreme pontiff at a somewhat later date.

Like the cope, the surplice (a white outer robe) entered liturgical usage in the Middle Ages as a late modification of the alb. By the 14th century its present role as a choral or processional garment was established. With the passage of time, the length of the garment grew progressively shorter.

Monastic garb

The surplice was also associated with the monastic orders, but vesture distinguished only the order and not the kind of order. Eremitical (hermitic) monasticism allowed no standard form of dress to develop, and only communal monasticism, beginning with the Rule of St. Benedict of Nursia in the 6th century, enabled standardization to become possible. Monastic dress included habit, girdle or belt, hood or cowl, and scapular (a long narrow cloth worn over the tunic). The salient characteristics of monastic dress have always been sobriety and conservatism. The orders proved even more retentive of archaic fashions than the hierarchy, and, in contrast to the deliberate splendour of ecclesiastical vestments, monastic dress was expressive of a renunciation of luxury. The contrast was functional in origin: the menial tasks of the monk related him sartorially to the peasant, whose humble avocations he often duplicated, rather than to the princes and prelates of the church, whose dress reflected the splendour of the ceremonies in which they engaged.

Because of the diversity of the monastic orders, only a summary account of their vesture may be given. The Benedictine mantle was black, fastened with a leather belt; but the Cistercians—reformed Benedictines—eschewed any dyed material and instead dressed in undyed woollen material, which w.ıs off-white in colour. In the course of time this became white, a tacit relaxation of the primitive austerity adopted as a protest against "luxury." Carthusians, a contemplative order founded in the 11th century, likewise wore white. In the 13th century the mendicant orders (friars) emerged. The Franciscans, founded by St. Francis of Assisi, first used a gray habit, which in the 15th century was exchanged for a brown one; in spite of this change they continued to be known as the Grey Friars. The Carmelites, an order founded in the 12th century, became known as White Friars. Dominicans, founded by St. Dominic from Spain, adhered from the beginning to a black robe over a white gown. Canons regular (communal religious persons living under vows), although ordained, lived like the orders under a rule, and the Augustinians (several orders following the Rule of St. Augustine) are styled Black Canons in contradistinction to the Premonstratensians, oı White Canons, an order founded by St. Norbert in the 12th century. Because the office (prescribed prayers) took up so much of a monk's time, his choir robes were almost as important as his day clothes. Surplices were worn in choir with an almuce over; this last was a lined shoulder cape designed to help the wearer resist the cold of medieval churches.

Nuns' costumes were similar to those of monks, the chief difference consisting in the replacement of the hood by a wimple (collar and bib) and head veil. Habits are white or black or mixed, and this remained unaltered till the 17th century, when the Sisters of St. Vincent de Paul introduced blue. This exception remained unique; nuns' habits retained a markedly medieval aspect until reformed by the second Vatican Council (1962–65).

The cassock has its origin in the barbarian *caracalla,* a robe favoured by the Roman emperor Bassianus (reigned 211–217), who came to be known as Caracalla because of the garment he habitually wore. Worn by the clergy as early as the 5th century, it became in time the standard day wear for prelates and priests, hierarchical rank being indicated by colour: bishops, archbishops, and other prelates wore purple; cardinals, red; the pope, white; and ordinary clergy, black (see also ROMAN CATHOLICISM).

Garb of priests, bishops, and monk

Eastern Orthodox religious dress. The Middle Ages also witnessed the evolution of Eastern Orthodox vestments into approximately their present form. The eucharistic garment corresponding to the chasuble was the *phelonion*, with variant forms in the Greek and Russian churches. The *sticharion,* which is held by the *zōnē*, or girdle, corresponds to the alb. The cuffs, or *epimanikia,* which fit over the *sticharion,* bear little or no resemblance to the maniple. The *epitrachēlion* is the Orthodox equivalent of the stole, but it hangs straight instead of being crossed over the chest, as is the case with the stole in Western churches. On the deacon, the *epitrachēlion* is pinned to the left shoulder and hangs in front and behind; with this exception, the deacon's vesture is identical with the priest's. The bishop wears an *omophorion*, whose shape and manner of wearing are closer to the original *pallium* than either the stole or the *epitrachēlion*. In place of the *phelonion*, since the 16th century, the bishop uses a dalmatic known as the *sakkos*. The *epigonation*, or rhombus-shaped portion of silk hanging to below the right knee, is common both to bishops and archimandrites (head abbots).

The monastic habit of the monk differs according to which of the three grades he occupies. The fully professed monk wears the great, or angelical, habit, which consists of the inner and outer rhasons, girdle, cowl (with veil), *analvos*, and *mandyas* (mantle). The inner rhason corresponds to the cassock and, like it, is used by the secular clergy. The outer rhason, a wide-sleeved garment, is black in the Greek Church but variable in colour in the Russian Church among the secular clergy (*i.e.*, those who minister in parishes). The *analvos* (shaped like the Western scapular, although historically unconnected with it) differentiates the full, or perfect, monk from the other grades, and its substance must be of animal, nonvegetable origin to remind the wearer constantly of death. The *mandyas* is the bishop's cloak (for non-eucharistic occasions), and in the Russian Church its use is granted to monks of the intermediate grade, although this license does not obtain in the Greek Church. In neither church may the *mandyas* or *analvos* be worn by monks of the lowest grade. Unlike Western orders, Orthodox monks dress only in black, but they share the same sartorial conservatism, their habits having remained unchanged in essentials from medieval times to the present (see also EASTERN ORTHODOXY).

Anglican and Lutheran liturgical vesture

Protestant religious dress. The Reformation of the 16th century varied in intensity from one country to another, and the fate of liturgical vesture suffered accordingly. With the rejection of the dogma of transubstantiation (the Roman Catholic teaching that in the Eucharist the substance of the bread and wine is changed into the body and blood of Christ, with the properties of the bread and wine remaining the same), the use of the mass garments might have been expected to be eliminated, but, wherever an altered eucharistic doctrine survived, an attenuated liturgical vesture contrived to survive with it. In the case of the Anglican and Lutheran churches, a paradoxical situation emerged whereby, in the latter, pre-Reformation practices (*e.g.*, use of crucifixes) survived alongside a Reformation theology, whereas, in Anglicanism, a Catholic theology survived along with a repudiation of Catholic rites. The Lutherans rejected the insignia of a celibate clergy but retained the chasuble for Communion services and the surplice and alb for other services.

Bishops in both Lutheran and Anglican communions retained the cope. The different editions of *The Book of Common Pmyer* (the Anglican liturgical book) attest to 16th-century reforms and the rising power of Puritanism, a 17th-century reform movement; the use of vestments declined in consequence. The cathedrals, however, maintained liturgical vestment standards to a certain degree, even when the last vestiges of liturgical propriety had been extinguished in the parishes in the 18th century. The cope became the High Church (liturgically oriented) vestment *par excellence,* worn by bishops not only processionally but even during Communion. Many views about the ceremonial revival of the 19th century have not in all respects been accurate; and followers of Edward Pusey, a leader of the Catholic revival known as the Oxford Movement, and ritualists sometimes blundered not from excess of archaeological zeal as has been commonly supposed but rather because they were inordinately influenced by their sociocultural environment. This may be less immediately obvious in the case of vesture than in architecture, but one result of overreacting was the loss, in the 19th century, of the customary dress of the clergy. The gown and cassock, as street attire, were allowed to fall into desuetude because in Puseyite views the gown was Genevan, whereas in reality it was the reverse. Another instance lay in the adoption of the (local) Roman biretta, introducing an Italian fashion even though adequate indigenous precedents were not lacking.

The gown, now inseparably associated in the popular mind with Genevan (Reformed) divines, was in fact opposed by these same divines in England and Scotland in the 17th century. In spite of this, standard vesture in Presbyterian churches is now the black gown and white linen bands over cassock and cincture, with the academic hood added for preaching services as a mark of learning appropriate to the pulpit, and a stole or scarf (see also PROTESTANTISM).

Moderrz changes in religious dress and vestments. With a change in emphasis, chiefly expressed in the episcopal use of the cope, Episcopalian usage in the first half of the 20th century differed little from Catholic rules except in Anglo-Catholicism, in which deliberate archaism imposed an adhesion to Baroque (17th to early 1Sth century) models, themselves superseded within Roman Catholicism. The Liturgical Movement of the 20th century has exercised an influence beyond the boundaries of the church in which it originated, and modern clerics of different denominations increasingly resemble one another sartorially because all have had recourse to the same sources of liturgical inspiration.

In Roman Catholicism, the formative period of religious dress was over before the Reformation, and Reformation influence was indirect—via the impetus supplied by the Counter-Reformation, which made Baroque its official art style. The emphasis on richness of material, excessive decoration, and preoccupation with surface set in motion a process of decline that was not arrested till the 20th century. The degeneration of the Gothic chasuble with its pointed folds into a stiff fiddle-backed, overembroidered vestment had begun as early as the 13th century with the practice of elevating the Host (sacrificial elements) in the mass. The elevation of the Host entailed the folding back on the celebrant's shoulders of the sides of the chasuble. The flexibility of the early chasuble permitted this, but, to facilitate the elevation, more and more material was removed from the sides until the garment became a caricature of its primitive form, distorted beyond recognition and its vestigial portions—dorsal (back) and pectoral (front)—came to be viewed simply as canvases for the display of virtuoso embroidery. Undergarments also became what is now viewed as effeminate with the addition of lace, and, although the Liturgical Movement began with a new theology of the Eucharist, its repercussions forced a decline of the Baroque style in dress.

The influence of the second Vatican Council

From the late Middle Ages to the 20th century, the history of religious dress in the Roman Catholic Church has been the history of its rubrical evolution: the regional variants of patristic (early church) and early medieval times were eliminated in the interest of ultramontanism (a theory that advocated a greater authority for the papacy), until the second Vatican Council reversed the process of eight centuries, again sanctioning regional divergences. Council rulings also simplified the use of the mitre and suppressed the use of the maniple altogether. Increased lay participation in the liturgy has led to an extension of lay religious dress in more than one communion. To lay offices such as the verger, who wears a gown over cassock, and chorister, who wears a surplice, Anglicans have added that of the lay reader, who vests in cassock and surplice, with a scarf as his ensign.

The upheavals of the 16th, 19th, and 20th centuries have not had much effect on Eastern Orthodox vesture, and the same canons (rules) prevail today in Orthodoxy as obtained prior to the fall of Constantinople in the 15th century. To ascribe this condition in Eastern Orthodoxy solely to the effects of cultural isolation would be an oversimplification. Suppression of vestments or their alteration is less likely to occur in a church in which such vestments have higher symbolic value attributed to them than in other traditions.

Islām. Islām attaches less importance to liturgical vestments than do most religions, but the social emphasis of the Islāmic faith finds expression in the universal application of the regulations governing dress; *e.g.*, all who enter the mosque remove their footwear, and all going on pilgrimage must wear the same habit, the ihrdm, and thus appear in the holy places in the guise of a beggar.

Because Islam recognizes no priesthood in a sense of a class sacramentally set apart, "clerical" functions are discharged by the *ʿulamāʾ*, or "the learned (in the Law)," whose insignia is the 'imdmah (a scarf or turban). The garb of the *ʿulamāʾ* exhibits geographical variations, but the *ʿimāmah* is found everywhere. Two broad regional distributions obtain, with Iraq as the area of confluence between the two. In the western part of the Muslim world, "clerical" dress has tended to become standardized according to the Azhar (Egyptian) pattern: a long, wide-sleeved gown (jubbah) reaching to the feet and buttoned halfway down its total length over a striped garment (caftan); and the headgear consists of a soft collapsible cap (*qalansūwah*) of red felt around which is wound a white muslin 'imdmah. In Syria a hard *ṭarbūsh* of the same red shade replaces the *qalansūwah.* Both the *qalansūwah* and the *ṭarbūsh* are provided with a blue tassel. The jubbah is usually a sober shade of blue, gray, or brown, and seldom black. Among the Sunnites—from Iraq eastward—the jubbah is worn in association with an *ʿabāʾ* (a long, full garment), traditionally of camel's hair and brown or black in colour. This is sometimes secured by a *ḥijām,* or cummerbund. In this second regional variant, the 'imdmah becomes a full turban replacing the cap, or fez. A green turban usually denotes a *sharīf*, or descendant of the Prophet Muhammad; and among the Shi'ites (the party of 'Ali) the entire "clerical garb" is black, as a symbol of mourning for the death of Husayn at Karbalā'.

The Ottoman Turks, as strict Sunnites, preferred turbans of other colours, which, elaborately wound, served to distinguish the wearer from a non-Muslim. On conquering Constantinople in 1453, they adopted the Byzantine cap and wound the turban around it in demonstration of conquest. The elaborately wound turbans of Persia and India also have a skullcap as a foundation for their folds. The art of winding a turban required no small degree of skill, the wearer fitting the cap over his knee and winding it in that position, whereafter the cap kept the folds in place. To the Prophet Muhammad is attributed the saying "What differentiates us [in appearance] from the polytheists is the turban." In India the turban has also been worn by non-Muslims, but the Muslim turban has remained distinguishable from the Hindu by the use of the skullcap as its foundation.

Religious dress **prohibitions**

For all Muslim males, whether Sunnite or Shi'ite, clerical or lay, the wearing of gold or silk is forbidden in consequence of a prescription (Hadith) of the Prophet, whereby the wearing of either was rendered "*ḥarām* [forbidden] to the males of my nation." Footwear must be removed on entering a mosque for fear of defiling the interior with ritually impure substances that may have adhered to the sole of the shoe. This rule applies also to entering a grave; thus, gravediggers and stonemasons must be unshod on such occasions. Because covering the head is a Near Eastern way of showing respect, a head covering should properly be worn in the mosque and even when praying outside the mosque.

When a Muslim purposes to visit the holy city of Mecca at the time of the major pilgrimage (hajj), he enters on a state of consecration and robes himself in two white seamless garments (*iḥrām*), which may not be exchanged for normal dress until he deconsecrates himself after the conclusion of the pilgrimage ceremonies. To these two garments women may add a veil.

Many of the mystical dervish orders (*ṭuruq*) wear distinctive robes, frequently with hierarchical differences. In Turkey, headstones are carved in the shape of the headdress distinctive to the order to which the deceased belonged and are tinctured in the appropriate colours. Particularly interesting are the ceremonial robes of the Mawlawiyah order (popularly known in the West as the Whirling or Dancing Dervishes), in which the symbolism of the robes is central to the mysteries of the order. The dervishes wear over all other garments a black robe (*khirqah*), which symbolizes the grave, and the tall camel's hair hat (sikke) represent3 the headstone. Underneath are the white "dancing" robes consisting of a very wide, pleated frock (*tannūr*), over which fits a short jacket (*destegül*). On arising to participate in the ritual dance, the dervish casts off the blackness of the grave and appears radiant in the white shroud of resurrection. The head of the order wears a green scarf of office wound around the base of his sikke.

For all Muslims of whatever sect the standard graveclothes are the threefold linen shroud, or kafan: the izdr, or lower garment; the *ridāʾ*, or upper garment; and the *lifāfah*, or overall shroud. Martyrs, however, are buried in the clothes in which they die, without their bodies or their garments being washed, because the blood and the dirt are viewed as evidences of their state of glory (see also ISLAM).

TYPES OF DRESS AND VESTMENTS IN EASTERN RELIGIONS

Indian religions. The distinction between ordinary dress and religious dress is difficult to delineate in India because the ordinary members of the various socioreligious groups may often be distinguished by their costumes. For example, Parsi (Zoroastrian) women wear the *sārī* (robe) on the right shoulder, not the left.

Hindu men frequently wear short coats (*angarkhā*), and the women wear a long scarf, or robe (sari), whereas typical Muslim attire for men and women is a long white cotton shirt (kurtah) and trousers (pd'ijamah). Muslim women also wear a veil called the burqah, which not only hides the face but also envelops the entire body. Traditional Sikh (a religion combining Hindu and Muslim elements) dress is an ordinary kurtah and cotton trousers, covered by a long hanging coat (choghah). The male Sikh is recognized especially by his practice of wearing his hair and beard uncut, the former being covered by a particularly large turban and the latter often restrained by a net.

Hindu religious garments

The Brahmin (Hindu priest) is distinguished primarily by the sacred thread ʿ*upavīta*), which is bestowed on him during his boyhood investiture and worn diagonally across the body, over the left shoulder, at all times. During the water offering to saints, it is worn suspended around the neck and, during ancestor rites, over the right shoulder. Devotees may also wear a tonsure that leaves a tuft of hair longer than the rest (*śikhā*). The *pravrajyā* ("going forth") associated with some *Upaniṣads* (Hindu philosophical treatises) involved a ritual rejection not only of homelife but also of the upavita and *śikhā*. Ascetics usually wear the ordinary loincloth, or dhoti, for meditation or yoga (a physical and psychological meditation system), but there is also a tradition of naked asceticism. A teacher (*swāmī*) traditionally wears a yellow robe (see also HINDUISM; SIKHISM; ZOROASTRIANISM AND PARSIISM).

Buddhism. Buddhism became more widespread in Asia than other ascetic and meditational movements, partly because of the strong organization of its monastic communities (sarigha). One of the main outward signs of the sarigha, along with the tonsure and the begging bowl, has always been the monk's robe; "taking the robe" became a regular expression for entering the sarigha. The sarigha was organized in accordance with the traditional code of discipline (vinaya), which includes the basic rules regarding robes in all Buddhist countries. These rules are all linked to the authority of the Buddha himself, but at the same time they allow considerable flexibility to cater to changing circumstances.

The Buddhist robe

The robe (civara) illustrates two main types of religious action, each symbolized by the character of the materials used. First, the wearing of "cast-off rags" was one of the "four resources" of a monk, being an exercise in ascetic humility similar to the other three, which are living on alms, dwelling at the foot of a tree, and using only cow's urine as medicine. The use of rags was later formalized into making the robes out of separate strips or pieces of cloth, but the rough patchwork tradition was carried over into China, where hermit monks in modern times wore robes made of old rags. In Japan, robes have been preserved with designs imitating the effect of patchwork, and robes sewn from square pieces of cloth were nicknamed "paddy-field robe" (*densōe*). This latter term is reminiscent of an old Indian Buddhist tradition according to which the Buddha instructed his disciple Ānanda to provide robes for the monks made like a field in Magadha (in India). which was laid out in "strips, lines, embankments, and squares." In general, whatever the degree of formalization, the rag motif ensured that the robe was to be "suitable for recluses and not coveted by opponents." The second type of religious action associated with the robe stemmed from the permission granted to monks to receive robes or the materials for making them from the laity. This meant that the laity "became joyful, elated, thinking: 'Now we will give gifts, we will work merit. . .' " (*Mahāvagga* VIII, 1, 36). Thus, the presentation of materials for robes was thought to have the same beneficial karmic effects (toward a better birth in the future) as the offering of food. The practice meant that various good materials were offered as well as rags, and in due course six types were allowed on the authority of the Buddha, namely, linen, cotton, silk, wool, coarse hempen cloth, and canvas.

There are three types of civara (*i.e.*, tricivara): the inner robe (Pāli, antaravdsaka), made of five strips of cloth; the outer robe (*uttarāsaṅga*), made of seven strips; and the great robe, or cloak (*saṃghāṭi*), made of nine, 15, or 25 strips.

In order to avoid the primary colours, Buddhist robes are of mixed colours, such as orange or brown. Another common term for the robe, *kasāya*, originally referred to the colour saffron, though this meaning is lost in the Chinese and Japanese derivatives, *chia-sa* and kesa. The robe is normally hung from the left shoulder, leaving the right shoulder bare, though some ancient texts speak of disciples arranging their robes on the right shoulder before approaching the Buddha with a question. In cooler climates, both shoulders may be covered with an inner robe, and the outer robe is hung from the left shoulder, as ın China.

Sandals are allowed if they are simple and have one lining only, or they may have many linings if they are cast-off sandals. The rules for nuns' robes are similar, but they also wear a belt and skirt. Some special vestments are worn by Tibetan Buddhists, including various hats characteristic of the different sects (see also BUDDHISM).

Chinese religions. Court dress, sacrificial dress, and ordinary dress were all influenced in ancient China by the Confucian-inspired civil religion. The classical text for the Confucian ideal of deportment and dress is Book X of the Analects, in which the emphasis is on propriety in every detail, whether at home or in affairs of state or ceremony. The undergarment, for example, was normally cut wide at the bottom and narrow at the top to save cloth, but it had to be made full width throughout for court and sacrificial purposes.

The Influence of the Confucian ideal of propriety

Confucius was also said to have insisted on the primary, or "correct," colours — blue, yellow, red, white, and black — rather than "intermediate" colours, such as purple or puce, and to have avoided red for himself because it was more appropriate for women.

Garments used in sacrifices to former kings and dukes were prepared from silk grown in a special silkworm house. According to the "Doctrine of the Mean," the clothes used by ordinary people at sacrifices were "their richest dresses." The fully developed Imperial costume for sacrifices was a broad-sleeved jacket and a pleated apron around the waist. Decorative symbols represented the universe in microcosm and thus the universal sovereignty of the emperor.

Funeral dress was generally white, although the *Shu Ching* ("Classic of History") refers to a funeral at which those who officiated wore hempen caps and variously coloured skirts. According to the I *Li*, mourning dress consists of "an untrimmed sackcloth coat and skirt, fillets of the female nettle hemp, a staff, a twisted girdle, a hat whose hat string is of cord, and rush shoes." For Mencius, a 4th–3rd-century-BC philosopher, the wearing of a coarse cloth mourning garment was an important aspect of traditional filial piety.

Buddhist robes in China followed Indian tradition fairly closely, though they were noted under the T'ang dynasty (AD 618–907) for being black in colour. Taoist robes, in contrast, were yellow. That this is an old tradition may be seen from the example of the 2nd-century-AD Yellow Turban movement, in which the missionaries and priests wore yellow robes and the followers yellow headdresses (see also CHINESE RELIGION).

Japanese religions. The priestly robes of Shintō are an example of the way in which rather normal garments of a formative age became the specialized religious vestments of later times.

Shintō religious garments

They consist of an ankle-length divided skirt (hakama) in white, light blue, or purple, depending on rank; a kimono in white, symbolizing purity, and of which there are various types; and a large-sleeved outer robe of various colours that is frequently a kariginu, or hunting garment, as used in the Heian period (794–1185). The headgear is a rounded black hat (eboshi). The more elaborate "crown" (kammuri) has a flat base, a protuberance rising forward from the back of the head, and a flat band curving down to the rear. Within a shrine, stiff white socks with a divided toe (*i.e.*, tabi) are worn, and, when proceeding to or from a shrine, officiants wear special black lacquered clogs (asagutsu) of paulownia wood. Shintō priests carry a flat, slightly tapered wooden mace (shaku), which symbolizes their office but otherwise has no precisely agreed upon significance. The dress of *miko* (girl attendants at shrines), whose main function is ceremonial dance, also typically consists of a divided skirt and a white kimono. They carry a fan of cypress wood. Young male parishioners bearing a portable shrine through the streets may wear a kimono marked with the crest of the shrine and a simple eboshi.

Buddhist robes continued the general Buddhist tradition, but of particular interest are the ornate ceremonial robes of high-ranking monks, especially in the Shingon and Nichirenite sects; the white robes worn by devotees in the syncretistic Shugen-db tradition (famous for its yamabushi, or mountain priests) during lustrations and similar rituals, symbolizing purity, as in Shintō; and the deep, inverted bowl-shaped hats of woven straw (*ajiro*-gasa) worn by Zen monks during begging tours.

Many new religions in Japan have carefully manufactured ceremonial vestments based on Shinto or Buddhist models or of mixed or original design. A common feature is the use of fairly simple uniform clothing for all believers during dedicated labour, mass rallies, or acts of worship. In Tenri-kyō, a religion founded in the 19th century by Nakayama Miki, the name of the religion figures prominently on the back of the garment, and, in Nichiren movements, the central symbol *namu Myōhō renge-kyō* ("Homage to the Lotus of the Good Law") may be displayed on a stole hanging from the left shoulder (see also JAPANESE RELIGION).

BIBLIOGRAPHY. HILAIRE and MEYER HILER, *A* Bibliography of Costume (1939, reprinted 1967), furnishes the widest bibliographical account. In the field of Christian dress the best recent works are HERBERT NORRIS, Church *Vestments:* Their Origin and Development (1949), a detailed and reliable historical coverage; and CYRIL E. POCKNEE, Liturgical Vesture: *Its* Origins and *Development* (1960), a succinct, well-illustrated account. All previous research was superseded by JOSEPH BRAUN, Die liturgische Gewandung im *Occident und* Orient...(1907), a publication that marked a turning point in liturgiological studies. JOHN B. O'CONNELL, The *Celebration* of Mass, new ed. (1956), a handbook intended for the priest, is a study of the rubrics and is supplemented by ADRIAN FORTESCUE's invaluable Ceremonies of the *Roman* Rite Described, 12th rev. ed. (1962), recognized as the standard work on Roman Catholic ritual, directing which garments should be used in the appropriate circumstances. Concerning Orthodox vesture, N.F. ROBINSON, Monasticism in the Orthodox Churches (1916, reprinted 1971); and Алексей **Николаевич Свирин,** *Древнерусское шитье* (1963), are descriptive rather than historical or analytical in methodology. Protestant vesture has not attracted the attention of liturgists, but PERCY DEARMER, The Parson's *Handbook*, 13th rev. ed. (1965); and

CHARLES WALKER, ***The Ritual Reason Why,*** new ed. (1931), treat of the subject from a High Anglican standpoint. On religious dress in Judaism, ***The Universal Jewish Encyclopaedia,*** 10 vol. (1939–43); and WILLIAM OESTERLEY and G.H. BOX, ***The Religion and Worship of the Synagogue*** (1907), are useful. On Islāmic dress, nothing has appeared in European languages, but in Persian there is MOHAMMAD BAQIR OL-MAJLISI, *Helyet ol-motaqqīn* (1952). On Far Eastern Religions there are a few relevant works. JEAN HERBERT, ***Shintô*** (1967), is the standard study of the subject and incorporates drawings of priestly attire as does S. ONO, ***Shinto: The Kami Way*** (1960). On Buddhist religious dress and vestments, ***The Book of the Discipline,*** vol. *4*, trans. by I.B. HORNER (1951), gives the early Buddhist traditions about "the robe." HOLMES WELCH, ***The Practice of Chinese Buddhism,*** vol. 1, ***1900–1950*** (1967), contains much incidental material about monastic vestments, with photographs.

(J.Di./M.Py.)

Religious Education

Religious education, the teaching or inculcation of spiritual, cultic, and moral precepts and practices, has been of significance to human culture in various ways. Where religion has been the primary shaper of a particular culture (*e.g.*, Hinduism of Indian culture), education in that religion has been essential to the well-being of the culture.

NATURE AND SIGNIFICANCE

Universal religious traditions—such as Buddhism and Christianity—have played significant roles as bridges between cultures. Where a particular religion is but one among others in a culture, religious education is important chiefly in a narrower sense; *i.e.*, for the survival of that particular religion. The significance of religious education to personal development is much debated in modern society, but there can be little doubt that in most societies of the past it has been important in instilling and illustrating the fundamental life models.

Religious education is concerned with morality, cultic practices, world views, myth, and religious experience. It is crucial in the process of socialization in most cultures. In those societies in which an idea of the sacred suffuses all life, all education may be said to be religious, and even the learning of such skills as hunting or planting has religious connotations. In traditional societies, in which a distinction is made between the sacred and the profane, and where formalized education develops, that education is likely to be dominated by religious ends.

Relationship between initiates and noninitiates

The development of a distinction between initiates and noninitiates in the religious community has a number of implications for religious education. Potential initiates must be properly identified and trained so that they may become full-fledged members of the community and may, in turn, play the role of initiators. There is also the question of stance toward noninitiates. Where status is determined solely by heredity that stance is likely to be one of rejection. But where status is dependent upon some criterion other than heredity the relationship to noninitiates is potentially a more positive one. Converts may in fact be openly sought—as in Buddhism and Christianity. Such a procedure will likely entail the elaboration and systematization of teachings and techniques designed to win converts, to educate them in the ways of the faith, and to meet the challenges of nonbelievers.

Functional differentiation within the religious community also has had considerable significance for religious education. Specialists have assumed the central role in religious education, both in shaping the content and techniques and in the actual process of preparing the initiate and educating the faithful. Education within Buddhism, for example, has been almost entirely carried on by monks. There developed in Judaism, on the other hand, a nonpriestly group (rabbis) whose primary function was study, interpretation, and teaching of the Torah.

TYPES OF RELIGIOUS EDUCATION

Religious education centred on the study of sacred scriptures. Many systems of religious education concentrate primarily upon the memorization or study of sacred scriptures or both. The importance of scriptural study in the Vedic (from the scriptures called Vedas) tradition of India is suggested, for instance, in the affirmation that "whatever may be the toils here between heaven and earth, the study of the scriptures is their last stage, their goal, . . ." In the West, scriptural tradition has roots in ancient Israel, from which emerged the Hebrew Bible (Old Testament). For more than 2,000 years this literature has constituted the core content of Jewish education. Christianity added the New Testament to this literature, and the resulting Christian Bible has been the foundation of religious education in that tradition. In some instances, in fact, the Bible has been almost the sole content of religious education. Likewise in Islām, the Qur'ān (sacred scripture) has constituted the core content of religious education in mosque schools.

Elaboration upon sacred scripture has occurred for purposes of explanation, summary, propaganda, and education. To the Hebrew Bible, for example, the Jewish tradition added the Talmud (interpretations of the Torah, or Law), which became the guide to content, method, and goal in Jewish education. Very early in the history of Christianity, creeds and confessions were developed for purposes of summarizing and explaining the faith. Some early Christians prepared special materials (catechisms) for use in the religious education of converts. In Isliim there emerged what is called *Ḥadīth*, that is, authoritative sayings of the prophet concerning revealed matters. These were often expounded in the mosque in connection with tha instruction of the faithful.

The importance of oral tradition

Preliterate peoples, of necessity, depended upon memorization for conveying culture from one generation to the next. Oral traditions also played a crucial role in the formation of sacred scriptures. Memorization of scriptures has been fundamental in certain forms of religious education. Phenomenal feats of memorization generally have been honoured as signs of piety and learning.

Religious education centred on theological issues. Most scripturally based traditions have valued analytical study and discussion as a logical step beyond memorization. Such analysis has tended to produce particular methods and schools of scriptural interpretation. Religious education reached a sophisticated intellectual level in the medieval schools of Europe that focussed attention on discussions of theological issues. These institutions utilized the dialectical method: the presentation of opposing points of view with the conviction that truth would emerge from the skillful use of logic and superior appeal to authority, especially scriptural authority. Religious education focussed upon detached discussions of profound religious-philosophical questions is evident in the Hindu *Upaniṣads* (composed c. 9th–5th century BC).

The role of the Catechetical school of Alexandria

Religious education centred on progressing from "lower" to "higher" sciences. Exposure to the thought forms and educational patterns of alien cultures has constituted a special challenge to religious traditions. An effort may be made to adapt the educational ideologies and methods of the alien culture to the ends of the religious tradition, as Buddhism did with indigenous patterns in China, and as Christianity did with Greek philosophy and practice. This process of adaptation is likely to work both ways. For example, the Christian Catechetical school of Alexandria, Egypt, in the 2nd and 3rd centuries, developed a kind of religious education that progressed from the "lower sciences" (mathematics, music, literature, biological sciences, philosophy, etc.) to the "higher (sacred or theological) sciences"—*i.e.*, those available only through study of the Christian Scriptures.

Religious education centred on experience. The ultimate goal of much text-centred religious education is some type of experience—such as union (with Christ, or God, or the Ineffable), deliverance (from illusion, temporality, etc.), or enlightenment. Some traditions in religious education, however, have more directly stressed experience and have given relatively less attention to sacred texts. Perhaps the most notable in this regard is Zen Buddhism (literally, meditation Buddhism) in which word-centred methods play a subordinate role in the achievement of enlightenment. Correct meditation proceeds through increasing withdrawal from surface realities and intellection to an ineffable experience. This meditation (*zazen*) is facilitated by a cryptic style of teaching that includes directed concentration upon enigmatic sayings (koans). The master, or *roshi* (venerated teacher), does not explain anything in a rational fashion, for it is contrary to his approach to appeal to the intellect.

Some traditions that stress experience speak in terms of stages—some of which may be more akin to psychophysical states than to levels of conceptual knowledge, such as the stages of mysticism: preparation, purgation, illumination, and union. Many religious traditions specifically require the individual to undergo certain physical disciplines to attain some religious knowledge. Fasting is a common technique. The controlled use of chemical substances, such as the Vedic *soma* or the peyote of certain American Indians, is another example of the attempt to facilitate religious knowledge through some prescribed course of action (see also PHARMACOLOGICAL CULTS).

The role of the teacher or master is especially significant in those traditions that emphasize progression in personal experience. Formalization and institutionalization tend to occur around especially adept masters. Since the desired experience is intensely personal in nature, however, the teacher cannot produce it. He is but a skilled midwife.

Religious education centred on myth and ritual. Myth and ritual perform important educational functions for common men in most religious traditions. By story myth communicates and evokes an experience of sacred reality that cannot be expressed by a simple, straightforward prose account. Similarly, ritual may be a method of communicating, hence, of educating. Many forms of expression may be used in ritual practice, including music, dance, drama, drawing, sculpture, and even architecture. Such forms may also be used as aids in formal instruction.

Religious education relies considerably on symbolic forms as teaching devices. Myth and ritual employ symbols that suggest familiar images and concepts. Thus, in Christianity, Baptism can suggest images of death and rebirth, washing and purification, the water that sustains life (drinking water), the water of eternal life, life-giving rain, etc. Through the use of such symbols religious education provides for a deepening understanding of a tradition based on the expansion of individual experience.

The use of verbal and nonverbal methods

Among the rites of passage commonly practiced by preliterate peoples, the rites associated with puberty, which are usually also initiatory rites, combine the verbal and the nonverbal in an especially significant way. "Explanations," chiefly of a mythological character, set the stage for these rites and give them meaning. The rites themselves may be highly dramatic and, in some instances, even painful, involving ordeals of an extraordinary sort. Upon passing through the experience the young man (or woman) is generally more fully exposed to rituals and teachings associated with sacred power and meaning.

Seasonal festivals and holy days, which characterize most religious traditions, are generally both celebrative and educative in nature. For example, the annual winter festival (chalako) among the Zuni Indians of the southwestern United States is an elaborate affair that conveys in dramatic form the basic tribal world view. Ritual and teaching are brought together in a potentially evocative way in the Jewish festival of Passover, which recalls to the child, both through his senses (taste especially) and his intellect, a crucial episode—the Exodus from Egypt in the 13th century BC—in the life of the people with which he is identified.

THE DEVELOPMENT OF RELIGIOUS EDUCATION IN THE RELIGIONS OF THE WORLD

Primitive religions Among primitive or preliterate peoples all activities that relate to survival are believed to have their origins with the gods. All education is related in some way to this sacred context and hence is religious. The world view is communicated through myth, ritual, and other ways. Its importance to group survival is made evident by various devices. In initiation the individual may be told what calamities will befall the group if certain taboos are violated. Every activity is communicated to children by means that tend to emphasize its proper performance and to insure its continuation.

Use of myth and ritual in primitive religion

Perhaps the best known aspects of primitive religions are their creation rites and seasonal festivals, often associated with fertility cults. Through these dramatic and colourful ceremonies, significant episodes in the religiocultural history of the tribe or people are re-enacted. These have both religious and educational value. The young may observe or participate in them. They may also imitate these adult religious activities in their own play activities. Seasonal festivals often utilize or cooperate with events in nature to communicate and reinforce the individual's understanding of the natural cycle of death and rebirth that is basic to his survival and his religious consciousness.

Rites of passage may also perform a religious educational function. In initiation, for example, the sacred secrets of the group are disclosed. The secret nature of this rite may generate curiosity, fear, and a desire for acceptance. all of which heighten the impact of the material presented. As knowledge of sacred things is presented, the novice may be told about exploits and travails of the gods and cultural heroes who provided this knowledge. Thus at the same time that respect and veneration for the sacred objects are inculcated, exemplary heroes may be held up as models of good behaviour and antiheroes may be portrayed to illustrate behaviour to be avoided. An initiatory rite of the aborigines of Arnhem Land in Australia illustrates the techniques used to educate in a primitive setting. The novices are shown the animal dances, and the older men explain their meaning. The novices are also shown the sacred trumpet.

> Each initiate is asked to try to blow the trumpet. Then the old men command all of them to "respect their fathers and mothers," "never to tell lies," "not to run after women who do not belong to them," "not to divulge any of the secrets of the men to the women, men who belong to a lower division of the association, or uninitiated boys," and all in all to live up to the tribal code.

Eventually the young men are circumcised. Then their wounds are steamed over a fire while they are instructed to avoid obscenity and adultery and admonished always to tell the truth.

Graded initiations are common among primitive people. The youth of the tribe may be assembled into "bush schools," where they are progressively initiated into higher levels of secret knowledge. Regular instructions in the form of lectures or memorized dialogues may be given. Advancement often is based on accomplishment, and recognizable "teaching" takes place. A tutor–tutee relationship may be fostered in which one member of the tribe or clan will have a special responsibility for a particular youth. "Wise men" may educate the group in the moral aspects of their religion through communication of "campfire lore" or through the recitation of proverbs. Simple stories for children serve to inculcate moral virtues and common values.

Though religious education among primitives can correctly be classed as being conservative in nature, this is not to say that such systems are necessarily repressive to creativity and innovation.

Religions of ancient peoples and civilizations. Religion was indistinguishable from the state in Egyptian, Mesopotamian, Aztec, and Mayan civilizations, and the cult of their nation-churches was geared to the perpetuation of the cosmic order of nature and society. This state cult itself was an impressive educational device, as can be noted in the annual new year's festival of cosmic renewal in Mesopotamia at which the great epic of creation (*Enuma elish*) was recited in its entirety and much of it was acted out, or in the Egyptian so-called Mystery Play of Succession marking the transition from one Egyptian king to another.

Education under the concepts of sacred kingship

These cultures were the products of significant developments in the technology of agriculture. There developed around this technology a pattern known today as sacred or divine kingship in which the king figure embodied the essential aspects of the civilization by combining in one functionary a responsibility for agriculture, the cosmos, and the political survival of the state. Effective educational transmission was essential to the ongoing life of these civilizations. This education was usually conducted by priests, and the schools were thus most often associated with temples. In the temple schools of Babylonia, for example, youths were taught to read and write the complicated script (cuneiform) and more advanced students pursued the higher branches of priestly learning; *i.e.*, the myths, the arts of divination, the liturgies and incantations, and other aspects of religious knowledge. It is likely that young men selected for official civic posts also received their education in temple schools. Whether designed for the priesthood or for more mundane pursuits, all education was of cosmic import in these ancient civilizations, and all was essential to the survival of the state.

In Egypt the sons of the priests went to the House of Books or to the school located in the "house of life" (temple) of some deity. The sons of nobility and of royalty also attended such schools to learn how to fulfill the religious duties of their particular offices. At lower levels, boys were educated in the many scribal duties that would make them a part of the functioning state. At a still lower level were the "trade schools," but even these instilled moral precepts and had their place in the overall schema of the religious destiny of the state. Life in these schools was not likely to be easy. Precept was not only drummed into the boys' ears but the master's principle was, "A boy's ears are on his back, and he hearkens when he is beaten."

Religious education in ancient American civilizations

In ancient Mexico systematic education took place in two types of institutions, both of which were essential to the religious functions of the state. Instruction in necessary priestly duties was given in *calmecac* schools where little boys chanted strophes from a ritualistic chant while sitting, according to the archaeologist George Vaillant (1901–45), with "their little legs and faces lacerated by maguey spines, their bodies thin from fasts and penance and their eyes dulled by the monotony of self-denial, . . ." The *telpuchalli* (house of youth) schools, where young men were trained in citizenship and the martial arts, offered no less serious an image. There the youths were prepared for ritual combat between the "sons of light" and the "sons of darkness," the object of which was to capture victims for the blood sacrifice that was believed to be required in order that the sun god might continue in his appointed rounds. Other schools trained women to be priestesses.

Glyph systems, alphabets, and the art of writing were introduced in the ancient civilizations discussed above, and this had important implications for religious education. Copying and memorization of texts became important methods in education. Rites and beliefs could be standardized and communicated through means other than human voice and memory, and cultural wisdom could be more surely preserved.

Though the sacred state produced a massive technoreligious culture, it also sowed the seeds of secularization. Class and caste distinctions became determinative of the type of education an individual might receive. Formal education increasingly prepared specialists to fill set roles. Among the increasingly intellectualized priestly classes, speculative and theological reflection became commonplace, but the common man retained the older forms of religion. The great cosmic rites of the sacred king became spectacles to be observed rather than rituals to be joined.

Religions of the East. *Hinduism.* The goal of traditional Hindu education is twofold: to prepare one to do his duty in this life, and to enable one to achieve release from the illusions of this life through realization of the identity of *Ātman* (individual soul) and Brahman (universal soul). The teacher, or guru, is the pivot of the traditional educational system designed to achieve these ends. The importance attached to proper religious education is evident in the fact that the category "student" is the first of the four stages of life according to orthodox Hindu doctrine, the others being the family supporter, the forest hermit, and the religious recluse. Ideally, this student stage is to be experienced by all potentially "twice-born" Hindus; that is, all members of the three upper classes, or castes (Brahmin, Kṣatriya, and Vaiśya, or priestly, warrior, and merchant), who alone may be initiated into the Hindu community. In actuality, Vedic education became the prerogative chiefly of the Brahmin caste and, in all likelihood, of only a minority of that caste in most periods of Indian history.

The role of the guru, or teacher

According to Vedic standards, the Brahmin youth is ready for initiation into the student stage at the age of eight. The initiation ceremony symbolizes the "second birth" of the boy. The initiate is invested with the sacred thread, which he wears for the remainder of his life; his hair is cut in a certain fashion; his chosen guru formally assumes responsibility for his welfare; and the boy, by ritual and vow, puts himself under the care of the guru. This initiation ceremony is not technically a puberty rite; the boy is not admitted to adulthood but to a crucial stage of study in preparation for the adult roles he later assumes. Girls have seldom been encouraged in formal Vedic study and, indeed, were sometimes even prohibited from such study.

The principal duties of the student are study of the Vedas, service to the guru, and chastity. The teacher transmits orally sacred texts and truths. In this learning situation, concentrated, pithy, and sometimes cryptic combinations of words have been devised to compress a maximum of meaning within a minimum of words. The crowning example of this is the sacred syllable *Om*, which is understood to contain within itself a world of meaning. The student assimilates what the guru has communicated and, ideally, through meditation realizes the profounder truths. Study is aided or accompanied by the performance of certain prescribed rites under the guidance of the guru. The teacher is served in a variety of ways in keeping with the dignity and importance of his position. This kind of Brahminic discipleship is given cosmic significance in one of the hymns of the Atharvaveda, which suggests that by the faithful performance of his duties the twice-born student actualizes and harmonizes the gods.

The guru tradition is one of the most long lasting and influential of traditions in a culture that itself has had a relatively long continuous history. As most commonly practiced, it involves one teacher and a few students living and learning together in the guru's home. On occasion, however, groups of learned men may come together for mutual instruction and to teach others, either in withdrawal from society as in the *āśrama* (spiritual or intellectual retreats) or as travelling bands within the world. The close relationship between pupil and guru also continues to exist in the monastic traditions in India.

Buddhism. According to commonly accepted tradition, after having achieved enlightenment, Gautama, the Buddha, out of compassion, taught others before he entered Nirvāṇa (state of enlightenment). First he taught the dedicated seekers after enlightenment—the early monastics. Then he urged his followers to teach other ascetics and those at a greater distance, the common people who did not have the inclination to become monastics but were involved in the daily round of family, farm, and village.

Importance of Buddhist monasticism

The history of Buddhist education is essentially an aspect of the history of Buddhist monasticism. Entrance into monastic life involved from earliest times a regimen of chastity, poverty, and other moral standards, and also submission to discipline and learning under the direction of a teacher. Meditation and study were the two essential methods in Buddhist formation. Principles such as those known as the four noble truths, the eightfold path, and details of the *dharma*, or law, were indispensable elements of study (see also BUDDHISM). In the early years the sayings, stories, and legends of the Buddha were gradually collected together into the canonical scriptures. Frequent repetition and chanting aided in remembering the words of and about the enlightened founder. Debate and discussion on doctrine also seem to have been an important aspect of monastic education.

A famous Indian Buddhist convert, King Aśoka (*c.* 265–238 BC), played a significant role in transforming Buddhism from an Indian sect into a world religion by encouraging the movement of monastic missionaries into all areas of India, and into Ceylon and Kashmir. Aśoka also prepared a number of Buddhist-informed edicts that he commanded to be engraved on rocks and polished stone pillars and to be read to the public on festival days.

Of the major sects of Buddhism, Theravāda (the religion of the Elders) claimed the PHli canon (sacred texts written in the PHli language) as its sole source of authority. The development of Mahāyāna (the Greater Vehicle) was accompanied by the production of additional siitras, or scriptures, that contained advanced doctrines attributed to the Buddha. These scriptures assumed a significant role in the development of Buddhist teaching and practice in China (and through China to Korea, Japan, and Vietnam) and in Tibet.

As the scriptural canons were established, scholarly monastics extended the scope of their learning in their efforts to explain and defend sacred truths while seeking converts. With the monastic community as a base, this expansionist impulse found expression in Buddhist monastic universities beginning in about the 6th century. The most famous of these was Nālandā in northern India, which, according to the accounts of a 7th-century Chinese writer, attracted thousands of monks of various Buddhist sects who engaged in study under the guidance of monks distinguished for their learning. Laymen also received instruction at Nālandā.

The monastic order (*saṅgha*) was the primary bearer of literate culture in Ceylon, Burma, and other regions of Southeast Asia into which Theravāda Buddhism penetrated. In Ceylon, for example, there appeals to have developed, following the introduction of Buddhism in the 3rd century BC, a continuous history of learning that centred especially on Buddhist literature. Schools associated with temples apparently were the only educational institutions until the establishment of Christian missionary schools. Before the Western incursions of recent centuries nearly all Burmese males spent some time in the pagoda schools. Girls received little formal education, although nunneries were an established aspect of Theravāda Buddhism and did attract some followers.

The empirical approach of Ch'an, or Zen, Buddhism

In China, Buddhism encountered a highly developed culture. Buddhist texts were translated and taught by Chinese Buddhist monks who often used the technical terminology of the classical Chinese religiomagical system known as Taoism in rendering the Sanskrit or PHli texts into Chinese. This religiocultural intermixture of Mahāyāna Buddhism and Taoism found institutional expression in the development of Ch'an (or meditation) Bnddhism in China. Ch'an (Japanese Zen) was radically empiricist in approach; it generally eschewed metaphysics (studies of supersensual reality), theory, and abstract reasoning in preference for direct insight produced by primal experience. Its teachers employed unconventional methods in an effort to elicit by indirection the coveted enlightenment. Another major Buddhist sect in China, Pure Land, or Amidism, stressed a universal salvation and utilized a more conventional approach to education than that employed in Ch'an. Scholarship and the study of scriptures were valued along with meditation.

For nearly eight centuries (c. 3rd–10th centuries AD), while Confucianism remained the primary influence upon domestic and civil life, Buddhism dominated the higher culture in China. Some Buddhist temples became great centres of learning, chiefly for monks and members of the upper classes. After the 10th century, however, Buddhism went into decline in China and monastic education tended to deteriorate.

As Chinese influence spread into Korea and Japan it carried the Buddhist imprint upon it. Though the sects that developed in Japan were similar to those that had emerged in China, Japanese Buddhism developed a character of its own and also exerted a much greater influence upon general educational patterns than Buddhism did in China. The village Buddhist temple often became the village school and the priest the teacher. The Zen stress upon effective mind and body control through a life of disciplined simplicity and earnest labour appealed to the military caste of Japan (the samurai), and Zen became influential in the training of warrior statesmen.

By the 9th century, Buddhism had achieved such influence in Tibet that under royal patronage a massive work of translating scriptures from Sanskrit into Tibetan was carried on. This Tibetan canon remains as one of the most significant sources of Buddhist history and thought. For centuries Buddhist monasteries, or lamaseries, were the sole centres of education in Tibet.

In the past century or two Buddhist religious education has been challenged variously by Christian missionary schools, the growth of state schools, and increasing secularization. At the same time, a Buddhist renaissance has occurred, with the encouragement of such Westerners as T.W. Rhys Davids (1843–1922), founder of the PHli Text Society, and Henry Olcott (1832–1907), an American theosophist who did much to promote the revival of Buddhist culture and education in Ceylon, Burma, and India. In recent years the *saṅgha* (monastic order) has been given increasing educational responsibilities, in some places in supervising modern-style private schools and universities and religious education in public school systems. Buddhist educational practices, especially those centred upon meditation, have also exerted an increasing influence in certain Western circles.

Chinese religions. China's educational system was dominated for two millennia by Confucianism, which, following its founder, stressed formal, precept-oriented, and popular education. For the founder of Taoism (Lao-tzu), on the other hand, popular education was a great mistake since it leads men farther and farther from the state of nature that is their only source of salvation. Hence Taoism had little significant influence upon Chinese education, aside from the appeal of its simple and yet paradoxical ideas. Even in its own monastic schools Taoism emphasized the way of direct insight. Though Taoist texts were studied in Chinese schools, they were interpreted in terms of Confucian methods.

Purpose of Confucian education

The system of schools linked the smallest village and its students through a chain of higher institutions and competitive examinations that culminated in the imperial academy. Education maintained a deceptively secular appearance in that its obvious purpose was the preparation of individuals for entrance into government civil service. Government itself, however, was intended to be the vehicle wherein certain basically religious functions might be performed.

Confucian education sought to maintain harmony throughout the cosmos by giving each individual moral training so that he might behave justly and by preparing and selecting those who were most learned in right conduct as most suited to govern. The classic textual guides for proper behaviour were linked to religion, both in terms of their origins and in their function of preserving the state's relationship to heaven through the Son of Heaven; *i.e.*, the emperor.

Japanese religions. The ease with which Japan has assimilated foreign cultural imports and remolded them into distinctively Japanese systems is perhaps nowhere more apparent than in the history of Japanese religions and religious education. At the base of Japanese religion is the ancient and pervasive practice of Shintb, the kami way (the way of the gods), an essentially primitive form of religious experience that has left its mark on all succeeding systems. To this base were added Confucian, Taoist, Buddhist, Christian, and Western secular ideas and practices. Through this eclectic fusion, a real sense of unity—a unique Japanese piety—was maintained. This process is clearly evident in two of the most influential educational systems in Japan, Bushido (Code of Warriors) and that of the Meiji restoration of the 19th century. Bushido was a unique synthesis of various religious and other elements that stresed the martial arts in the training of the samurai (warrior class). The system produced a remarkable combination of learning, piety, ceremonial propriety, military skill, loyalty, self-control, and selflessness. Meiji education fused elements of Shintb mythology, kami faith, Confucian familial ethics, authoritarian statism, and modern educational methods into an impressively effective instrument for the inculcation of devotion to the state.

Shintb traditionally provided little or no formal religious instruction. It was passed on chiefly through ritual participation in home, temple, and community. Confucianism, on the other hand, was a strong force for formal education. Confucian learning was developed most extensively under the Tokugawa shogunate (17th–19th centuries). The traditional curriculum extended through Japan to the courts of the provinces. The temple or parish schools (terakoya) evolved into a semisecular system of popular education. Taught by a priest (either Buddhist, Shintō, a Confucian scholar, or an unattached Samurai), pupils received a watered-down version of samurai education. Morals were inculcated through typical copybook methods for learning to read and write. Traditional texts were memorized and reproduced.

With the demise of the imperial cult of Shintō in 1945, the state schools discontinued their direct inculcation of a particular ideology. Religious groups, including even traditional folk Shintō, became more self-consciously concerned about religious education. Western, especially Christian, institutions of religious education were emulated. Some religious groups even established their own school systems.

Religions of the West. Judaism. Religious education is stressed repeatedly in the Hebrew Bible. The Israelite is commanded to teach the commandments to his children, to recite them faithfully, to wear them upon his person, and to inscribe them on his doorpost. The household is, then, the primary locus of this sacred duty.

Importance of religious teachers in Judaism

Relatively eaily in Israelite history, the community looked to special teachers to reinforce and extend teaching in the family. Levites are referred to as those "who taught all Israel. . . ." (II Chron. 35:3.) Following the Babylonian Exile (6th to 5th centuries BC) there emerged a group of scholars ("scribes") who were especially qualified in the interpretation and transmission of the divine teaching (Torah). Scholar-teachers (rabbis) in this scribal tradition after the destruction of the Second Temple (AD 70) played a decisive role in assuring the survival of Judaism and in the shaping of that religion down to modern times. Among the better known of the early rabbis were Hillel (1st century BC to 1st century AD), and his student, Johanan ben Zakkai (1st century AD), under whom the rabbinic academy at Jabneh (Jamnia) in Palestine became the chief religious-intellectual centre of Jewry after the destruction of the Second Temple; Gamaliel the Elder (1st century AD), who is mentioned in the New Testament as the teacher of St. Paul; and Akiba (1st–2nd centuries AD), who played a significant role in preparing for the codification of the oral law in the Mishna (a part of the Talmud, a commentary on the Torah, or Law). The compilation of the oral law as developed by rabbinic interpreters over four or five centuries was completed around AD 220. There then ensued among the rabbis a period of detailed discussion of the Mishna, a work that entailed an intensive use of a question and answer method and that produced the Gemara (the commentary on the Mishna). The Mishna and the Gemara were put together in the compilation of the Talmuds (commentaries on the Torah) of Palestine (c. AD 400) and Babylonia (*c.* 500 AD). Subsequently the Talmuds played a primary role in Jewish religious education. Each of the Talmuds emerged from the major rabbinic academies of the day—institutions that, along with others of like nature, became authoritative centres of interpretation and practice in the Jewish religion.

The role of the synagogue in Jewish education

The educational work of the rabbi and the family met and merged in the synagogue. This remarkable institution was both a congregational house of prayer and a place of religious instruction. (The Yiddish term for synagogue is *shul* or "school.") It also was of considerable significance in the survival of Judaism beyond the destruction of the Second Temple in AD 70. Its educational work was supplemented from relatively early times by formally established community schools.

The importance attached to formal religious education is attested to in various ways in the history of Judaism. The initiation of the child into formal education outside the family circle became a solemn rite. It took place relatively early in the child's life and preferably on Shavuot, the festival commemorating the giving of the Law to Moses. The child was brought to the synagogue in special dress and there heard the reading of the Decalogue (the Ten Commandments). Then the teacher began to teach him the Hebrew alphabet from a tablet smeared with honey that the child ate as he pronounced the letters in order to encourage the view that learning is sweet. From then on religious education was a concentrated business, with stress upon close study and memorization of the various texts. Public evidence of learning in the Torah was given upon the ceremony of the Bar Mitzwa (literally, "son of the commandment"), at the age of 13. Formal religious education was generally limited to boys and men. Occasionally a woman became an accomplished student of the Torah, but the primary educational role of women was that of example of motherly faithfulness.

The basic object of all religious education was "labouring in the Torah for its own sake," and this life of study has been considered to be the most laudable kind of activity for any pious Jew to follow. Rabbinical guidance of this process was given not only in the synagogue and the attached school but also through the academies of higher learning (called yeshivas—plural of yeshiva). These academies were centres of authoritative interpretation and advanced study of the Talmud. The term Talmud Torah, "Study of the Talmud," became the title of various types of educational institutions in medieval and modem times.

Emancipation from the ghettos

Emancipation from ghettos in the late 18th century brought about profound changes in Jewish religious education. The long-established principles and practices of "labouring in the Torah" had become so systematized and so pervasive in the Jewish community that the introduction of the radically different ideas and practices of the modern, secularized West threatened to undermine the Jewish religion. Though the more orthodox defenders of Jewish belief and practice bitterly resisted the encroachment of secularization, reformers, under the leadership of such men as the great Jewish philosopher of the 18th-century Enlightenment Moses Mendelssohn, endeavoured to adapt Jewish education, and the whole Jewish world view, to the intellectual and institutional realities of the secularized world. Theological schools were founded for the training of rabbis, and efforts were made to strike a mean in these schools between ancient and modern methods and views. After emancipation, which brought civil rights and political equality, many Jews eagerly embraced secularized education as the chief pathway to successful life in the modern world. But disappointment sometimes followed this eagerness, as state schools were used to convert Jews to Christianity or to inculcate a nationalism that also threatened to destroy Jewish identity or both. Since the Nazi virtual extermination of European Jewry in the 1930s and '40s, efforts have beenmade, especially in the United States and Israel, to revive many of the ancient educational principles and practices. At the same time, many rabbis and educators continue the attempt to develop new and viable methods for teaching Hebrew and the Jewish heritage in Western secularized society.

Zoroastrianism and *the Parsis.* The fifth most important duty of a Zoroastrian, a follower of the religion founded by the 6th century BC Iranian Zarathushtra, or Zoroaster, was "to spend a third of [his] days and nights in attending the seminary and consulting the wisdom of holy men, . . ." Schools among the early followers of Zoroaster were most often located on the premises of the temples in which the fire dedicated to the god Ahura Mazdā (the Good Lord) was maintained. Priests served as teachers. Rote learning of sacred texts, training in ritual observances, transmission of arts of reading and writing, and moral instruction by means of maxims and proverbs occurred at an elementary and fairly universal level in ancient Persia. Opportunities for education were apparently extended to all classes and both sexes. Higher education was concerned primarily with religion, medicine, and law. In western Persia a system of military religious education prepared men to engage in what can be termed "holy wars." Zoroastrianism as an educational system was intricately linked with the needs of the state until AD 651 when the last Zoroastrian dynasty was overthrown by the Muslims. Since the beginning of the British Colonial period the Parsis (the term for the Persian Zoroastrians who settled in India) have shown persistent interest both in systems of Western secular education and in the recovery of their own sources.

Christianity. The word teacher is often applied to Jesus in the Gospels. Much early Christian literature is directed toward questions of belief and practice. The most systematic work of this sort is the *Didachē* (or "The Lord's Teaching to the Gentiles by the Twelve Apostles," late 1st century AD), a manual of discipline and instruction for believers and converts.

The importance of catechetical instruction

The most significant educational device developed in early Christian history was the catechumenate, the systematic instruction of converts preparatory to Baptism and full participation in the fellowship and rites of the church. The convert was prepared through formal instruction in morals and doctrine and through liturgical practice, as in the mass of the catechumens that preceded the mass of the faithful. He was admitted to the latter only after his instruction had been satisfactorily completed, his moral character had been established, and he had been baptized. Trained catechists used special catechetical materials prepared by some of the best thinkers in the early church, such as Gregory of Nyssa (c. 335–394). Formal religious education was concerned, then, primarily with the adult convert; the religious education of children was understood to be the responsibility of the family or household.

Catechetical instruction achieved a high level of sophistication in the "catechetical schools," institutions of higher education that evolved in some urban intellectual circles of Christians in the late 2nd century. In their heyday the schools of Alexandria and Antioch (and later Edessa) were rivals, the former being noted as the champion of Neoplatonism, which emphasized idealism and was marked by the extensive use of allegory in scriptural interpretation, while the latter followed the Aristotelian emphasis on matter and form and more literal interpretation of the Bible.

The catechumenate declined in the 5th and 6th centuries. The sacramental focus of instruction shifted from Baptism to penance in the church of the Middle Ages, and symbolism was employed extensively to convey the truths of the faith. The church of the Middle Ages developed a variety of means for conveying and reinforcing the faith among illiterate and semiliterate peoples. Embellishments upon the ancient liturgy, elaboration of festivals and holy days, and use of the arts—including painting, architecture and drama—contributed to this end.

In the early Middle Ages formal religious education was carried on chiefly in monastic and cathedral schools. Under the leadership of Charlemagne (crowned emperor of the Holy Roman Empire in 800) and his educational aide, Alcuin (c. 732–804), a court school system was developed that exerted significant influence in the 8th and 9th centuries. Education within these schools was concerned with the essentials of the faith. But subjects of study ranged beyond scripture and Christian doctrine to include in some instances all or several of the seven liberal arts; *i.e.*, grammar, rhetoric, dialectic, arithmetic, geometry, astronomy, and the theory of music.

Scholasticism, a system of Western Christian philosophy flourishing from the 11th to 15th centuries, and the rise of the universities brought a significant flowering of critical and systematic thought. Under the impact of the rediscovery of the teachings of Aristotle, Scholastics, or the Schoolmen, led by St. Thomas Aquinas (1225/1226–1274), elevated theology to new heights as "queen of the sciences," and hence as significant to all learning.

Reformation emphases

The religious reforms of the 16th and 17th centuries were characterized by a stress upon both moral and common education within a Christian context. Martin Luther (1483–1546) urged that all people be educated in the faith, and he himself contributed significantly to this end by translating the Bible into the vernacular, preparing catechisms that were widely used, and encouraging the extensive use of music in facilitating both worship and education. John Calvin (1509–1564) was equally convinced of the importance of educating all of the faithful, and he too wrote a catechism and promoted the establishment of schools. The followers of these men stressed religiously informed education wherever they went; for example, the early Puritan (Calvinist) settlers of North America who, within less than 20 years of their settlement (1630) in New England, had founded a college (Harvard) and enacted legislation requiring publicly supported education in all communities of 50 householders or more.

The most significant educational development in the Catholic Reformation sprang from the work of Ignatius of Loyola (1491–1556) and the monastic order of the Society of Jesus (Jesuits) that he founded. This highly disciplined movement sought to bring the total man—mind, body, spirit—under control for the faith and the church. To this end it developed a system of religious and secular education without parallel in Christian history.

Significant developments in Christian education since the Reformation include: (1) the establishment of extensive educational systems by the Roman Catholic Church, with or without state support, a development that was directed ideologically by a series of papal encyclicals that stressed that education of the faithful must be carried on under church control; (2) stress upon education of "the heart" as well as the head among such Protestant groups as the Moravians, the German Pietists, the Quakers, and the Methodists; (3) the founding of liberal arts colleges based upon religious impulses, especially by Protestants in the United States; (4) the gradual separation within both Roman Catholicism and Protestantism of clerical from lay education, a trend that was institutionalized in the establishment of theological seminaries separate from colleges or universities; (5) the flourishing of the Sunday school movement in Great Britain and the United States in the 19th and early 20th centuries, a movement that stressed lay religious education separate from state or public schools; (6) the role of missionary schools in carrying both Christian and modem secular ideas into Asia and Africa; and (7) efforts to adjust method and content in Christian education to the realities of a secularized Western culture, such as in the extensive use of the educational theories of John Dewey (1859–1952) by the Religious Education Movement in the United States in the early part of the 20th century.

The importance of the *kuttāb*

Islām. Muhammad (the founder of Islam in the 7th century AD) is reputed to have admonished his followers to attain education even if it entailed "going to China." Historians record that he allowed certain prisoners of war to gain their freedom through teaching a certain number of Muslims to read and write. The ***kuttdb,*** or elementary school, was established early to promote literacy. These schools employed simple poetry and prose ethical maxims as typical copybook texts for teaching literacy. Sometime after the Qur'ān (Koran) was set down in writing it became the only text employed in a student's early years of ***kuttdb*** instruction. Once the Qur'ān was memorized and a semblance of literacy was achieved (often by the age of 10), *kuttāb* education might extend up to the age of 15. This study included some basic ***Hadith,*** or traditional sayings and actions of the Prophet, these being memorized as guides to proper conduct. Some *kuttāb* seem to have provided also a basic exposure to ***fiqh,*** the law as it was derived from the Qur'ān and the ***Hadith.***

Education developed along many lines beyond ***kuttdb*** instruction. A knowledge of classical Arabic was necessary to study the Qur'ān and the ***Hadith.*** Attempts to clarify the Qur'ān and the ***Hadith,*** as well as to validate many ***Hadith*** sayings, also generated study in linguistics and philology, and the development of deductive logic. The study of ***fiqh,*** or jurisprudence, became a methodology for interrelating the various sources of authority. A knowledge of *fiqh* was necessary even on a village level of administration, and hence such studies became relatively extended through society.

Every mosque served as a school of sorts. It was appropriate for men to meet in the courtyard to discuss tradition and the law. Where one man was recognized as an authority, others grouped around him as their teacher and the situation became somewhat formalized. Students often travelled from mosque to mosque to learn at the feet of respected masters. Many mosques built up libraries that served prominently in public education due to their availability. Indeed, the library was a major aspect of Islāmic society. A number of institutions existed with collections ranging to more than 100,000 volumes. Most of the classical literature that spurred the European Renaissance was obtained from translations of Arabic manuscripts in Muslim libraries.

Formal mosque schools gradually developed around recognized teachers. Instruction in these institutions occurred within the context of the normal Muslim day. The *salāts* (prayers) were observed and many schools had their own *imām* (leader). Some were even identical in physical layout to regular mosques and a number came to offer Friday prayer services.

In AD 1066–1067 (or AH 459 in the Islāmic system of dating of years after the Hegira, or flight of Muhammad from Mecca), Niẓām al-Mulk opened the first *nzadrasah* (higher school) in Baghdad. In a very short time there were large numbers of these schools throughout the Muslim world. The original function of these institutions was to provide an educational base for the teaching of Sunnite (followers of the Prophet's way) beliefs to the then predominantly Shīʿite (followers of Ali ibn Abi Ṭālib, cousin of the Prophet Muhammad) populace. Subsequently, they filled a less controversial role as centres of advanced learning throughout the culture.

Muslim education reached a pinnacle in the al-Azhar Mosque in Cairo. Here instruction has been provided in a great variety of subjects and faculties have been maintained in Arabic and in Muslim law and theology. This university has been especially important to Islam since the 13th century when Mongol invasions destroyed many of the centres of higher learning that had developed in Iraq and Persia.

The religion of Islām generated a constant pressure for education and understanding. Many forms of education developed apart from the more formal schools. But the crowning aspect of Islāmic education was the university. Here the teachings of the Prophet were perpetuated within Islām. Here also aspects of Hellenistic culture were continued and passed on to other, non-Islāmic cultures, such as that of Europe.

RELIGIOUS EDUCATION TODAY

Modern challenges to religious education

Religious education is confronted by a variety of challenges today, including those arising from modern science, technology, mass culture, and the increasing centralization of power in the modern state. Perhaps the last of these poses the most obvious institutional challenge. The political revolutions of recent times have thoroughly undermined traditional religious education and instituted state-controlled systems in its place. Religious groups have developed various ways of dealing with this situation: liaisons with state schools, as in Germany and England; the establishment of separate school systems, such as the Roman Catholic parochial school system in the United States; and fugitive efforts where state control has been most rigorous, as in the Soviet Union and China.

The American Revolution (1775–1781) brought about separation of church and state, and in the 19th century the public school assumed the primary role in inculcating common values, similar to the role that had been played by the established church in earlier times. Such practices as Bible reading, hymn singing, and prayer were utilized in the public schools as much for moral and civic as for religious ends. These and other practices, such as the allowance of time for religious education, were generally encouraged by religious groups in their efforts to carry on the religious education of the young. In recent decisions, the Supreme Court of the United States has declared unconstitutional released time for religious education on public school grounds and school-sponsored prayers and devotional Bible reading. One of the most recent of these decisions (*School District of Abington Tp., Pa.* v. *Schcmpp,* 1963), however, apparently encouraged the study of, as contrasted with the practice of, religion in the public schools.

The study of religion has been encouraged in some state colleges and universities in the United States for more than a generation, beginning with religiously affiliated schools of religion (such as at the universities of Iowa and Montana) and continuing more recently in departments of religious studies (such as at the University of California at Santa Barbara). This development has been characterized by an increasing emphasis upon the scholarly study of religion—as illustrated, for example, in the increasing attention given to the various world religions and to an examination of the nature and function of religion itself. A similar shift is gradually occurring in the public schools in the movement from devotional Bible reading and prayer to curricular programs in the study of religion. This sort of change represents a significant departure from common practices in religious education in the past. It may serve an important function in encouraging knowledge about man's religious faith, but how far it can go in inculcating or sustaining faith is not clear.

In recent years religious teaching has also been excluded from or severely limited in the state-controlled schools of such countries as Turkey, India, and Japan. At the same time, in such recently established nations as Burma, Sri Lanka (Ceylon), Israel, and Pakistan the prevailing religious traditions have been used in varying degrees as vehicles for promoting national unity. A kind of religious education or religiously informed communal education has been encouraged in the state schools of these countries. Efforts have been made to bring together traditional and modern teachings and practices.

Trends in religious education

Trends and experiments illustrating the continuing efforts of religious groups and leaders to meet the various modern challenges to religious education include: the continuing establishment of colleges and universities under religious control or direction or both (*e.g.*, Buddhist universities in Sri Lanka, the International Christian University in Japan); attempts to bring theological education into closer relationship with university education; ongoing attempts to develop viable programs in religious and moral education in state schools; a resurgence of religiously sponsored day schools, especially among certain conservative religious groups; experimentation with such techniques as sensitivity training and affective learning; utilization for purposes of evangelization and education of the various electronic and other means afforded by modern technology and mass culture; and the establishment in the West of Eastern or Oriental types of religious education programs (*e.g.*, Zen meditation centres and monasteries). Generalization is hazardous, considering this variety. In fact, stating that there is the wide variety of approaches may be the only adequate generalization.

BIBLIOGRAPHY. C.J. ADAMS (ed.), *A Reader's Guide to the Great Religions* (1965): for the subject of religious education the volume is most directly helpful on Buddhism (pp. 150–152), and Christianity (pp. 275–276); E.D. MYERS, *Education in the Perspective of History,* with a chapter by ARNOLD J. TOYNBEE (1960), a general history of education primarily as manifested in the religious education of major civilizations; G. DE SANTILLANA and H. VON DECHEND, *Hamlet's Mill: An Essay on Myth and the Frame of Time* (1969), a startling, if not revolutionary, study of the object of religious education in antiquity; R. ULICH, *A History of Religious Education: Documents and Interpretations from the Judaeo-Christian Tradition* (1968), a brief guide to the development of religious education in the West, and (ed.), *Three Thousand Years of Educational Wisdom: Selections from Great Documents* (1954), brief selections from India, China, and Islām, with longer selections from ancient and medieval Christianity and the Judaic tradition and material on Renaissance Humanism and the development of modern educational theory and practice; W.D. HAMBLY, *Origins of Education Among Primitive Peoples: A Comparative Study in Racial Development* (1926), somewhat dated, but still the best single-volume treatment; W.T. DE BARY (ed.), *Sources of Indian Tradition* (1958), short selections on studentship in Hinduism (pp. 228–230), and the English impact on Indian education (pp. 587–601); A.L. BASHAM, *The Wonder That Was India: A Survey of the Culture of the Indian Sub-continent Before the Coming of the Muslims,* pp. 161–165 (1954), a classic work with a brief discussion of education; RADHAKUMUD MOOKERJI, *Ancient Indian Education (Brahmanical and Buddhist),* 2nd ed. (1952), a detailed, extensive use of sources, although it lacks sharpness in analysis and is given at times to unfounded generalizations or assumptions; SUK-

UMAR DUTT, *Buddhist Monks and Monasteries of India: Their History and Their Contribution To Indian Culture* (1963), good selections on "eminent monk-scholars" and "monastic universities" (pt. 4 and 5); H.S. GALT, *History of Chinese Educational Institutions* (1952), difficult reading but deals directly with the subject; F.A. LOMBARD, *Pre-Meiji Education in Japan: A Study of Japanese Education Previous to the Restoration of 1868* (1913), a good, but hard to obtain, general history; R.P. DORE, *Education in Tokugawa Japan* (1965), *a* comprehensive treatment of the Tokugawa period; L.S. DAWIDOWICZ (ed.), *The Golden Tradition: Jewish Life and Thought in Eastern Europe,* sect. 3, "The Quest for Education" (1967), sources showing some of the effects of emancipation; L.J. SHERRILL, *The Rise of Christian Education* (1944), substantial scholarly treatment from the ancient Israelites to the Reformation; K.A. TOTAH, *The Contribution of the Arabs to Education* (1926), a fairly comprehensive treatment and probably most available in the U.S.; AHMAD SHALABY, *History of Muslim Education* (1954), more up-to-date and complete than Totah but hard to obtain; M.N. DHALLA, *Zoroastrian Civilization* (1922), three chapters on education in each of the three major epochs of Persian history, but does not treat the Parsis; CHRISTOPHER DAWSON, *The Crisis of Western Education* (1961), illuminating treatments of the impact of science, the effects of the French Revolution on education, and the relationships between religion, patriotism, and education in the American common school; ROBERT MICHAELSEN, *Piety in the Public School: Trends and Issues in the Relationship Between Religion and the Public School in the United States* (1970), discussions of the efforts of churches to use the public schools for religious education, relationships between Catholic parochial schools and the public schools, and recent U.S. Supreme Court cases.

(R.S.M./G.C.T.)

Religious Experience

Religious experience is taken here to include such specific experiences as wonder at the infinity of the cosmos, the sense of awe and mystery in the presence of the holy, feelings of dependence on a divine ground (a divine power or being as the basis of all existence) or an unseen order, the sense of guilt and anxiety accompanying belief in a divine judgment, and the feeling of peace that follows faith in divine forgiveness. Some thinkers also point to a religious aspect or dimension of experience in general having to do with the purpose of life as a whole and with the ultimate destiny of the individual. In the first sense, religious experience means an encounter or coming into the presence of the divine in a way analogous to encounters with other persons and things in the world. In the second case, reference is made not to an encounter with a divine being but rather to the apprehension of a quality of holiness or rightness in reality or to the fact that all experience can be viewed in relation to the ground from which it springs. In short, religious experience means both special experience of the divine or ultimate and the viewing of any experience as pointing to the divine or ultimate. Religious experience is sometimes equated with mysticism, but many interpreters have held that the experiences in question need not be understood exclusively in terms of the immediacy and ineffability characteristic of mysticism.

THE STUDY AND EVALUATION OF RELIGIOUS EXPERIENCE

"Religious experience" was not widely used as a technical term prior to the publication of *The Varieties of Religious Experience (1902)* by William James, an eminent U.S. psychologist and philospher, but the interpretation of religious concepts and doctrines in terms of individual experience reaches back at least to 16th-century Spanish mystics and to the age of the Protestant Reformers. A special emphasis on the importance of experience in religion is found in the works of such thinkers as Jonathan Edwards, Friedrich Schleiermacher, and Rudolf Otto. Basic to the experiential approach is the belief that it allows for a firsthand understanding of religion as an actual force in human life, in contrast with religion taken either as church membership or as belief in authoritative doctrines. The attempt to interpret such concepts as God, faith, conversion, sin, salvation, and worship through personal experience and its expressions opened up a wealth of material for the investigation of religion by psychologists, historians, anthropologists, and sociologists as well as by theologians and philosophers. A focus on religious experience is especially important for Phenomenologists (thinkers who seek the basic structures of human consciousness) and Existentialist philosophers.

Four basic issues

A number of controversial issues have emerged from these studies, involving not only different conceptions of the nature and structure of religious experience but also different views of the manner in which it is to be evaluated and the sort of evaluation possible from the standpoint of a given discipline. Four such issues are basic: (1) whether religious experience points to special experiences of the divine or whether any experience may be regarded as religious by virtue of becoming related to the divine; (2) the kinds of differentia that can serve to distinguish religion or the religious from both secular life and other forms of spirituality, such as morality and art; (3) whether religious experience can be understood and properly evaluated in terms of its origins and its psychological or sociological conditions or is *sui generis*, calling for interpretation in its own terms; and (4) whether religious experience has cognitive status, involving encounter with a being, beings, or a power transcending human consciousness, or is merely subjective and composed entirely of ideas and feelings that have no reference beyond themselves. The last issue, transposed in accordance with either a Positivist outlook or some types of Empiricism, which restrict assertible reality to the realm of sense experience, would be resolved at once by the claim that the problem cannot be meaningfully discussed, since key terms, such as "God" and "power," are strictly meaningless.

Proponents of mysticism, such as Rudolf Otto, Rufus Jones, and W.T. Stace, have maintained the validity of immediate experience of the divine; theologians such as Emil Brunner have stressed the self-authenticating character of man's encounter with God; naturalistically oriented psychologists, such as Freud and J.H. Leuba, have rejected such claims, explaining religion in psychological and genetic terms as a projection of human wishes and desires. Philosophers such as William James, Josiah Royce, William E. Hocking, and Wilbur M. Urban have represented an idealist tradition in interpreting religion, stressing the concepts of purpose, value, and meaning as essential for understanding the nature of God. Naturalist philosophers, of whom John Dewey was typical, have focussed on the "religious" as a quality of experience and an attitude toward life that is more expressive of the human spirit than of any supernatural reality. Theologians Douglas Clyde Macintosh and Henry N. Wieman sought to build an "empirical theology" on the basis of religious experience understood as involving a direct perception of God. Unlike Macintosh, Wieman held that such a perception is sensory in character. Personalist philosophers, such as Edgar S. Brightman and Peter Bertocci, have regarded the person as the basic category for understanding all experience and have interpreted religious experience as the medium through which God is apprehended as the cosmic person. Existential thinkers, such as Søren Kierkegaard, Gabriel Marcel, and Paul Tillich, have seen God manifested in experience in the form of a power that overcomes estrangement and enables man to fulfill himself as an integrated personality. Process philosophers, such as Alfred North Whitehead and Charles Hartshorne, have held that the idea of God emerges in religious experience but that the nature and reality of God are problems calling for logical argument and metaphysical interpretation, in which emphasis falls on the relation between God and the world being realized in a temporal process. Logical Empiricists, of whom A.J. Ayer has been typical, have held that religious and theological expressions are without literal significance, because there is no way in which they can be either justified or falsified (refuted). On this view, religious experience is entirely emotive, lacking all cognitive value. Analytic philosophers following the lead of Ludwig Wittgenstein, an Austrian–British thinker, approach religious experience through the structure of religious language, attempting to discover exactly how this language functions within the community of believers who use it.

RELIGIOUS EXPERIENCE AND OTHER EXPERIENCE

Sense experience and beyond

Views of experience in general. Religious experience must be understood against the background of a general theory of experience as such. Experience as conceived from the standpoint of a British philosophical tradition stemming from John Locke and David Hume is essentially the reports of the world received through the standard senses. Experience, as a tissue of sensible content, was set in contrast to reason, understood as the domain of logic and mathematics. The human mind was envisaged as a wax tablet on which the sensible world imprints itself; and the one who experiences is the passive recipient of what is given. It is possible to distinguish, compare, and in other ways relate these sensible items by means of understanding, but the data themselves are available only through experience—*i.e.*, the sensation of things and reflection upon thought and mental activities, feelings, and desires. According to this classical empiricist view, all ideas, beliefs, and theories expressed in conceptual form are to be traced back to their origin in sense if they are to be understood and justified.

The above view of experience came under criticism from two sides. Immanuel Kant, an 18th-century German philosopher, who still retained some of the assumptions of the position he criticized, nevertheless declared that experience is not identical with passively received sensible material but must be construed as the joint product of such material and its being grasped by an understanding that thinks in accordance with certain necessary categories not derived from the senses. Kant opened the way for a new understanding of the element of interpretation in all experience, and his successors in the development of German Idealism, Johann Fichte, Friedrich Schelling, and G.W.F. Hegel, came to characterize experience as the many-sided reflection of man's multiple encounters with the world, other men, and himself.

A second attack on the classical conception came from U.S. Pragmatist philosophers, notably Charles Sanders Peirce, William James, and John Dewey, for whom experience was the medium for the disclosure of whatever there is to be encountered; it is far richer and more complex than a passive registry of sensible data. Experience was seen as a human activity related to the purposes and interests of the one who experiences, and it was understood as an interpreted product of multiple transactions between man and the environment. Moreover, stress was placed on the social and funded character of experience in place of the older conception of experience as a private content confined to the mind of an individual. On this view, experience is not confined to its content but includes modes or dimensions that represent frames of meaning—social, moral, aesthetic, political, religious—through which whatever is encountered can be interpreted. James went beyond his associates in developing the broadest theory of experience, known as radical empiricism, according to which the relations and connections between items of experience are given along with these items themselves.

The expression and interpretation of experience

Critics of the classical view of experience, while not concerned exclusively with religious experience, saw, nevertheless, that if experience is confined to the domain of the senses it is then difficult to understand what could be meant by religious experience if the divine is not regarded as one sensible object among others. This consideration prompted attempts to understand experience in broader terms. Cutting across all theories of experience is the basic fact that experience demands expression in language and symbolic forms. To know what has been experienced and how it is to be understood requires the ability to identify things, persons, and events through naming, describing, and interpreting, which involve appropriate concepts and language. No experience can be the subject of analysis while it is being had or undergone; communication and critical inquiry require that experiences be cast into symbolic form that arrests them for further scrutiny. The various uses of language—political, scientific, moral, religious, aesthetic, and others—represent so many purposes through which experience is described and interpreted.

Views of religious experience. Specifically religious experience has been variously identified in the following ways: the awareness of the holy, which evokes awe and reverence; the feeling of absolute dependence.that reveals man's status as a creature; the sense of being at one with the divine; the perception of an unseen order or of a quality of permanent rightness in the cosmic scheme; the direct perception of God; the encounter with a reality "wholly other"; the sense of a transforming power as a presence. Sometimes, as in the striking case of the Old Testament prophets, the experience of God has been seen as a critical judgment on man and as the disclosure of his separation from the holy. Those who identify religion as a dimension or aspect of experience point to man's attitude toward an overarching ideal, to a total reaction to life, to an ultimate concern for the meaning of one's being, or to a quest for a power that integrates human personality. In all these cases, it is the fact that the attitudes and concerns in question are directed to an ultimate object beyond man that justifies their being called religious. All interpreters are agreed that religious experience involves what is final in value for man and concerns belief in what is ultimate in reality.

Because of their intimate relation to one another, the religious and the moral have often been confused. The problem has been intensified by many attempts—beginning with Kant's treatise on religion (1793)—to interpret religion as essentially morality or merely as an incentive for doing one's duty. Religion and morality are, however, usually taken to be distinguishable; religion concerns the being of a person, what he is and what he acknowledges as the worshipful reality, while morality concerns what the person does and the principles governing his relation to others. While it is generally acknowledged that religion must affect man's conduct in the world, some have maintained that there is no morality without religion, while others deny this claim on the ground that morality must remain autonomous and free of divine sanctions. Religious experience may be distinguished from the aesthetic aspect of experience in that the former involves commitment and devotion to the divine, while the latter is focussed on the appreciation and enjoyment of qualities, forms, and patterns in themselves, whether as natural objects or works of art. Anthropological studies have shown that primitive religions gave birth to many forms of art that, in the course of development, won independence as secular forms of expression. The problem of the relation between religion and art is posed in a particularly acute way when reference is made to religious art as a special form of the aesthetic. Since it is'concerned with the holy and the purpose of human life as a whole, most scholars would hold that religious experience should be related in an intelligible way to all other experience and forms of experience. The task of tracing out these relationships belongs to theology and the philosophy of religion.

THE STRUCTURE OF RELIGIOUS EXPERIENCE

The three basic elements: subjective, objective, social

The self and the other. All religious experience can be described in terms of three basic elements: first, the personal concerns, attitudes, feelings, and ideas of the individual who has the experience; second, the religious object disclosed in the experience or the reality to which it is said to refer; third, the social forms that arise from the fact that the experience in question can be shared. Although the first two elements can be distinguished for purposes of analysis, they are not separated within the integral experience itself. Religious experience is always found in connection with a personal concern and quest for the real self, oriented toward the power that makes life holy or a ground and a goal of all existence. A wide variety of individual experiences are thus involved, among which are attitudes of seriousness and solemnity in the face of the mystery of human destiny; feelings of awe and of being unclean evoked by the encounter with the holy; the sense of a power or a person who both loves and judges man; the experience of being converted or of having the course of life directed toward the divine; the feeling of relief stemming from the sense of divine forgiveness; the sense that there is an unseen order or power upon which the value of all life depends; the sense of being at one with the divine and of abandoning the egocentric self.

In all these situations, the experience is realized in the life of an individual who at the same time has his attention focussed on an "other," or divine reality, that is present or encountered. The determination of the nature of this other poses a problem of interpretation which requires the use of symbols, analogies, images, and concepts for expressing the reality that evokes religious experience in an understandable way. Four basic conceptions of the divine may be distinguished: the divine as an impersonal, sacred order (Logos, Tao, *ṛta*, Asha) governing the universe and man's destiny; the divine as power that is holy and must be approached with awe, proper preparation, or ritual cleansing; the divine as all-embracing One, the ultimate Unity and harmony of all finite realities and the goal of the mystical quest; and the divine as an individual or self transcending the world and man and yet standing in relation to both at the same time.

The two most important concepts that have been developed by theologians and philosophers for the interpretation of the divine are transcendence and immanence; each is meant to express the relation between the divine and finite realities. Transcendence means going beyond a limit or surpassing a boundary; immanence means remaining within or existing within the confines of a limit. The divine is said to transcend man and the world when it is viewed as distinct from both and not wholly identical with either; the divine is said to be immanent when it is viewed as wholly or partially identical with some reality within the world, such as man or the cosmic order. The conception of the divine as an impersonal, sacred order represents the extreme of immanence since that order is regarded as entirely within the world and not as imposing itself from without. The conception of the divine as an individual or self represents the extreme of transcendence, since God is taken as not wholly identical with either the world or any finite reality within it. Some thinkers have described the divine as wholly transcendent of or "wholly other" than finite reality, some have maintained the total immanence of the divine, and still others claim that both concepts can be applied and therefore that the two characteristics do not exclude each other.

"Solitariness" and community

Social forms or expressions. Most enduring, historical religious traditions find their roots in the religious experience and insight of charismatic individuals who have served as founders; the sharing of their experience among disciples and followers leads to the establishment of a religious community. Thus, the social dimension of religion is a primary fact, but it need not be seen as opposed to religious experience taken as a wholly individual affair. There has been some difference of opinion on the point; Whitehead, for example, put emphasis on the "solitariness" of religious experience precisely in order to deny the claim of those who, like Émile Durkheim, a French sociologist, characterized religion as essentially a social fact. The social expression of religious experience results in the formation of specifically religious groups distinct from such natural groups as the family, the local society, and the state. Religious communities, including brotherhoods, mystery cults, synagogues, churches, sects, and monastic and missionary orders, serve initially to preserve and interpret their traditions or the body of doctrine, practices, and liturgical forms through which religious experience comes to be expressed. Such communities play a significant role in the shaping of religious experience and in determining its meaning for the individual through the structure of worship and liturgy and the establishment of a sacred calendar. Communities differ in the extent to which they stress the importance of individual experience of the divine, as distinct from adherence to a creed expressing the basic beliefs of the community. The tension between social and individual factors becomes apparent at times when the individual experience of the prophet or reformer conflicts with the norm of experience and interpretation established by the community. Therefore, although the religious community aims at maintaining its historic faith as a framework within which to interpret experience of the divine, every such community must find ways of recognizing both novel experience and fresh insight resulting from individual reflection and contemplation.

Objective "intention," or reference. Religious experience is always understood by those who have it as pointing beyond itself to some reality regarded as divine. For the believer, religious experience discloses something other than itself; this referent is sometimes described as the "intentional" object that is meant or aimed at by the experiencing person. Analysis of religious experience, interpretations placed upon it, and the beliefs to which it gives rise may result in the denial that there is any such reality to be encountered or that the assertion of it is justified by the experience in question. This conclusion, however, does not change the fact that all religious experience, whether that of the mystic who strives for unity with God or of the naturalist who points to a religious quality in life, purports to be experience "of" something other than itself. The question of the cognitive import or the objective validity of religious experience is one of the most difficult problems encountered in the philosophy of religion. In confronting the question, it is necessary to distinguish between various ways of describing the phenomena under consideration and the critical appraisal of truth claims concerning the reality of the divine made on the basis of these phenomena. Even if describing and appraising are not utterly distinct and involve one another, it is generally admitted that the question of validity cannot be settled on the basis of historical or descriptive accounts alone. Validity and cognitive import are matters calling for logical, semantic, epistemological, and metaphysical criteria—of the principles of rational order and coherence, meaning, knowledge, and reality—and this means that the appraisal of religious experience is ultimately a philosophical and theological problem. The anthropologist will seek to identify and describe the religious experience of primitive peoples as part of a general history and theory of man; the sociologist will concentrate on the social expression of religious experience and seek to determine the nature of specifically religious groupings in relation to other groups—associations and organizations that constitute a given society; the psychologist will seek to identify religious experience within the life of the person and attempt to show its relation to the total structure of the self, its behaviour, attitudes, and purposes. In all these cases attention is directed to religious experience as a phenomenon to be described as a factor that performs certain functions in human life and society. As William Warde Fowler, a British historian, showed in his classic ***Religious Experience of the Roman People*** (*1911*), the task of elucidating the role of religion in Roman society can be accomplished without settling the question of the validity or cognitive import of the religious feelings, ideas, and beliefs in question. The empirical investigator, as such, has no special access to the critical question of the validity of religious experience.

Cognitive import or objective validity

The most radical form of the denial that religious experience has cognitive import is advanced by the Logical Positivists, who hold that all assertions or forms of expression involving a term such as "God" are meaningless because there is no way in which they can be verified or falsified.

Others who hold that religious utterance based on experience is without cognitive import regard it either as the expression of emotions or an indication that the person using religious language has certain feelings that are associated with religion. Those who follow the lead of Wittgenstein regard religious utterances as noncognitive but attempt to determine the way in which religious language is actually used within a circle of believers. Some psychologists have denied cognitive status to religious experience on the ground that it represents nothing more than man's projection of his own insecurity in the face of problems posed by life in the world and therefore has no referent beyond itself.

Immediacy and mediation. ***Revelational and mystical immediacy.*** Among defenders of the validity and cognitive import of religious experience, it is necessary to distinguish those who take such experience to be an immediate and self-authenticating encounter with the divine and those who claim that apprehension of the divine is the result of inference from, or interpretation of, religious experience. Two forms of immediacy may be distinguished: the revelational and the mystical. Christian theologians, such as Emil Brunner and H.H. Farmer, speak of a "divine–human encounter," and Martin Buber, a Jewish religious philosopher, describes religious experience as an "I–Thou" relationship; for all three, religious experience means an immediate encounter between persons. The second form of the immediate is the explicitly mystical sort of experience in which the aim is to pass beyond every form of articulation and to attain unity with the divine.

Mediation through analysis and critical interpretation. A number of thinkers have insisted on the validity of religious experience but have denied that it can be understood as wholly immediate and self-supporting, since it stands in need of analysis and critical interpretation. Some, like Paul Tillich, hold that there are certain "boundary experiences," such as having an ultimate concern or experiencing the unconditional character of moral obligation, that become intelligible only when understood as the presence of the holy in experience. Others, such as H.D. Lewis and Charles Hartshorne, find the divine ingredient in the experience of the transcendent and supremely worshipful reality but demand that this experience be coherently articulated and, in the case of Hartshorne, supplemented by rational argument for the reality of the divine. Dewey envisaged a religious quality in experience pointing to God as an ideal that stands in active and creative tension with the actual course of events. Whitehead identified the presence of the divine with an apprehension of a "permanent rightness" in the scheme of things and based the validity of the experience on the claim that an adequate cosmology requires God as a principle of selection aiming at the realization of the good in the world process. James found the justification of religious experience in its consequences for the life of the individual: valid experience is distinguished by its philosophical reasonableness and moral helpfulness. Finally, some have sought to combine experience and interpretation by taking the traditional proofs of God's existence and pointing to their roots in the experience of perfection, of the contingency of one's own existence, and of the reality of purpose in human life. On this view, the arguments for the reality of God are not wholly formal demonstrations but rather the tracing out of intelligible patterns in experience.

Three forms of preparation: rational dialectic, moral purification, drugs

Preparations for experience. Mystics, prophets, and religious thinkers in many traditions, both East and West, have been at one in emphasizing the need for various forms of preparation as a preliminary for gaining religious insight. The basic idea is that ordinary ways of looking at the world, dictated by the demands of everyday life, stand in the way of the understanding of religious truth; man must pass beyond these limitations by the disciplining of his mind and body. Three classic forms of preparation may be distinguished: first, rational dialectic for training the mind to reach insight (this explains why many mystical thinkers from the Pythagoreans to Nicholas of Cusa and Benedict de Spinoza were deeply involved in mathematics); second, moral preparation aiming at purity of heart, which was sometimes conjoined with bodily discipline, as in the Indian Yoga exercises; third, the use of drugs to expand the range of consciousness beyond that required for ordinary life. It is significant that the great mystics invariably regarded such preparation as necessary, but not sufficient, for experience. The self may be prepared, but the vision may not come; being prepared, as it were, establishes no claim on the divine. The experience described by St. John of the Cross, a 16th-century Spanish mystic, as "the dark night of the soul" points precisely to the experience of failure. The soul in this situation is convinced that God has abandoned it, cast it into darkness, perhaps forever. Mystics in the Taoist and Buddhist traditions have often emphasized the spontaneity of insight and the need to seek it through an "effortless striving" that combines the need to search with the awareness that the insight cannot be compelled. Zen Buddhists are fond of pointing to insights that are already possessed but not recognized as such until their holder is shaken loose from ordinary patterns of thought.

SITUATIONAL CONTEXTS AND FORMS OF EXPRESSION

Cultic and devotional. Religious experience receives its initial, practical expression in the forming of the cult that provides an orderly framework for the worship of the religious object. Worship includes expressions of praise, acknowledgments of the excellency of the divine, communion in the form of prayer, and the use of sacraments or visible objects that signify or represent the invisible sacred beyond them, feelings of joy and of peace expressed often in musical form, and sacrifice or the offering of gifts to the divine or in the name of the divine. Worship is ordered by means of liturgy directing the experience of the worshipper in patterns that combine the written word, the spoken word, and sacred music in a unity aimed at bringing him or her into the presence of the divine.

Life crises and rites of passage. Religious experience has to do with the quality and purpose of life as a whole and with the ultimate destiny of the person. Certain special times and events in the course of life present themselves as occasions that are set apart and celebrated, because they direct man's thought to the divine and the sacred with peculiar forcefulness. These occasions, called life crises, are regarded as dangerous because they are transitional from one stage of life to another and open to view the relation of life as a whole to its sacred ground. Pregnancy and birth, the naming of a child, being initiated into the community—sometimes called "puberty rites"—the choice of a vocation, the celebration of marriage, and the time of death are experienced as special events distinct from the routine happenings of secular life. These events represent "crises"—*i.e.*, turning points—when man's relation to the sacred becomes a matter of special concern. As Gerardus van der Leeuw, a Dutch Phenomenologist and historian of religions, points out, these transitional times are occasions for celebration in every culture because they mark the death of one stage and the birth of another in a universal cycle of life.

Religious experience and secular life

Sacred and secular. The marking off of these crisis occasions from the routine events of daily life points to the all-important distinction between the sacred and the secular. As directed toward the sacred, religious experience finds expression in the specifically religious form of the cult and in the cycle of sacred life. There is, however, a secular as well as a sacred life, and, since religious experience concerns the whole of life, the religious meaning must be related to all the dimensions of secular life—political, economic, moral, technological, and other. The relationship is twofold; on the one hand, there is the bearing of the conception of the divine on standards of behaviour, and, on the other, there is the influence that the religious meaning has upon one's general attitude toward life. The sacred, thus, makes its impact on the secular by providing principles that are to govern the relations between persons and by holding before men a vision of the divine that gives purpose to life as a whole. Although the sacred retains its dynamism by becoming related to secular life, there is the constant danger that it will lose itself in the secular, unless specifically religious forms of life are preserved. The existence in every society of secret and mystery cults, of sacred brotherhoods, of groups of disciples devoted to holy men, of monastic orders, and, on the broadest scale, of established churches and denominations, points to the need felt to retain the sacred as a special domain that can neither be merged into nor contained within secular society.

Verbal, conceptual, and symbolic. In all of the world religions, religious experience receives its most enduring expression in the form of sacred scriptures and the body of commentary through which they are interpreted. Mythological and symbolic forms of expression are older than conceptual forms and systems of doctrine. Myth takes the form of a story and represents the imaginative use of materials drawn from sensible experience in order to express a religious meaning surpassing the sensible world. Myths of creation in many religions give ample evidence of this imaginative function. The task of the theologian using conceptual tools is to elucidate the thought content of the myth and other primary forms of religious expression—legend, parable, confession, lamentation, prophetic vision—and thereby reduce the degree of dependence on the sensible and imaginative elements. It is important to distinguish devotional and liturgical expressions from the theological use of language. Creeds, confessions, psalms and hymns of praise, litanies and scriptures containing a record of the lives and experiences of sacred persons, all give immediate expression to the primary experience upon which a religious tradition is founded. Systems of theology and religious philosophy make their appearance when it becomes necessary to conceptualize and express consistently the body of belief about the divine, the world, and man implied in this primary experience. Tension exists between religious experience and theological expression at two points: first, the pietistic and evangelical spirit in religion, as seen, for instance, in some forms of Protestant Christianity, and the *bhakti* devotional movement in Hinduism, seeks to preserve the primacy of experience at the expense of theology; and, second, those who acknowledge the indispensability of theology will also demand that its formulations remain in accord with the experience it is meant to express and interpret.

Tension between religious experience and theological expression

TYPES OF RELIGIOUS EXPERIENCE AND PERSONALITY

The personal character of religious experience makes it essential to understand its varieties as manifested in different types of personality and the functions they perform. The mystic, a reflective and contemplative type, shuts out the world and all distracting influences in order to reach true selfhood through purification and enlightenment. Although mysticism has social implications, the mystic is primarily an individualist, whereas the prophet, a person of intense but intermittent experience, sees himself called to be a spokesman for the divine to the community or all mankind, and regards his own experience as a message that enables him to interpret the past and the future in the light of the divine will. The priest is a mediator between man and the divine, and his main function is the proper ordering of worship through liturgical forms. By contrast with the prophet, whose insight is spontaneous, the priest attains the authority of his office through education and training; as guardian of the tradition, he must assume administrative responsibilities in addition to his role as spiritual adviser; thus he is both active and contemplative. The reformer is a figure who stands within a religious tradition and seeks to transform or revitalize it in the light of his own experience and insight. The reforms intended may be moral, intellectual, or ecclesiastical, depending on the particular genius of the individual. Common to all reformers is the conviction that some valid and essential feature of traditional faith has been ignored or distorted and that these deficiencies must be overcome if the religion is to be purified. It is characteristic of the reformer to be actively engaged in bringing about the reforms indicated by his renewing experience. The monk or member of a religious order is in search of a special or sacred place set apart from secular life within which a religious life can be lived and moral and religious demands fulfilled to a greater degree than is possible in the world. Different orders stress different aspects of experience: some emphasize ascetic practices and self-discipline; others are devoted to the preservation of learning and the development of theology; still others make missionary zeal uppermost, and the members are impelled by their own experience to seek to convert others. The forerunner of the monk, who lives in a community governed by rule, was the hermit or religious recluse, the type for whom solitary existence, preferably in deserts and barren places, is necessary for communion with the divine and self-purification. The saint is a figure venerated by the religious community as one who embodies perfection in some form. The saint may have been a martyr, exhibiting perfection in faith; a person possessed of intensified capacity for experience and communion with the divine; or one who achieves to a supreme degree the moral and spiritual ideals of the beatific life. The theologian has the task of expressing the historic faith of a community concerning the divine *(tkeos)* in rational or conceptual form (*logos*). The content of his thought, though handed on to him in its essentials by the tradition, will depend on his own experience and his insight into the special relevance of that tradition for his time. The theologian both interprets and reinterprets. The founder, as might be expected, surpasses all others in importance. The founder's experience forms the basis of his own authority and the substance of the religion he establishes. The intensity of his experience and the effect it has upon his personality are decisive factors determining the response of his initial followers and disciples. There is reason to believe that the founders of the great religions, such as Moses, Buddha, and Jesus, did not intend to fill this role; the founding of the religion in each case was the result of the impact of their personalities and of the profundity of their experience on those who gathered around them.

The founder's experience: Moses, Buddha, and Jesus

BIBLIOGRAPHY

The nature of religious experience: WILLIAM JAMES, *The Varieties of Religious Experience* (1902), a classic philosophical and psychological study; RUDOLF OTTO, *Das Heilige,* 9th ed. (1922; Eng. trans., *The Idea of the Holy,* 1923; 2nd ed., 1950), a study of the nonrational in religious experience; J.M. MOORE, *Theories of Religious Experience, with Special Reference to James, Otto and Bergson* (1938); JAMES A MARTIN, JR., *Empirical Philosophies of Religion* (1945); and C.C.J. WEBB, *Religious Experience* (1945), contain valuable appraisals and good bibliography; H.E. BRUNNER, *Wahrheit als Begegnung* (1937; Eng. trans., *The Divine-Human Encounter,* 1943); and MARTIN BUBER, *Ich und Du* (1923; Eng. trans., *I and Thou,* 1937 and 1970), express the view that authentic religion is based on personal encounter between man and God; JOHN DEWEY, A *Common Faith* (1934), argues for the "religious" in experience; E.S. BRIGHTMAN, A *Philosophy of Religion* (1940), represents the Personalist view that personhood is the most basic quality of reality; W.T. STACE, *Time and Eternity* (1952); and D.T. SUZUKI, *Mysticisnt: Christian and Bucldhist* (1957), represent the mystical interpretation; W.E. HOCKING, *The Meaning of God in Human Experience* (1912, reprinted 1963); and J.E. sMITH, *Experience and God* (1968), emphasize the experiential basis of the question of God.

Religious experience and other experience: H.D. LEwIs, *Our Experience of God* (1959); and J.E. SMITH, *Religion and Empiricism* (1967), deal with the bearing of different conceptions of experience on religion; JOSIAH ROYCE, *The Sources of Religious Insight* (1912); W.G. DE BURGH, *From Morality to Religion* (1938); and PAUL TILLICH, *Morality and Beyond* (1963), treat the relation between religion and morality; GERARDUS VAN DER LEEUW, *Vom Heiligen in der Kunst* (1957; Eng. trans., *Sacred and Profane Beauty,* 1963), treats the relation beween art and religion.

The structure of religious experience: JOHN MACMURRAY, *The Structure of Religious Experience* (1936, reprinted 1971); J.B. PRATT, *The Religious Consciousness* (1920); PAUL TILLICH, *The Dynamics of Faith* (1957); and A.N. WHITEHEAD, *Religion in the Making* (1926), deal with psychological, theological, and metaphysical aspects; JOACHIM WACH, *The Sociology of Religion* (1944), is an indispensable study of the social expression of religious experience; WILLIAM A. CHRISTIAN, *Meaning artd Truth in Religion* (1964); F. FERRE, *Basic Modern Philosophy of Religion* (1967); NINIAN SMART, *Philosophers and Religious Truth* (1964); and J.E. SMITH, *Reason and God* (1961), deal with the issue of the cognitive import of religious experience; J.H. LEUBA, *The Psychology of Religious Mysticisnz* (1925), argues against its cognitive import; W.E. HOCKING, *Science and the Idea of God* (1944); W.T. STACE, *Religion and the Modern Mind* (1960); and H.N. WIEMAN, *The Wrestle of Religion with Truth* (1927), discuss the relation between religion and science; J. MACQUARRIE, *God-Talk* (1967); I.T. RAMSEY, *Christian Discourse* (1965) and *Models and Mystery* (1964), represent the linguistic approach to religious experience; MIRCEA ELIADE, *Le Mythe de l'éternel retour* (1949; Eng. trans., *Cosmos and History: The Myth of the Eternal Return,* 1954; rev. ed., 1965) and *The Sacred and the Profane: The Nature of Religion* (1959), interpret religious experience through myth, symbol, and ritual.

Situational contexts and forms of expression: EVELYN UNDERHILL, ***Worship*** (1936, reprinted 1957), invaluable for the meaning of worship and its forms; P. EDWALL *et al.* (eds.), ***Ways of Worship*** (1951), an account of the liturgies of the major Christian communities; EMILE DURKHEIM, ***Les Formes élémentaires de la* vie *religieuse*** (1912; Eng. trans., ***The Elementary Forms of the Religious Life,*** 1915, paperback 1961), presents the "group theory" of religion; C.C.J. WEBB, ***Group Theories of Religion and the Individual*** (1916), a critique of Durkheim; MIRCEA ELIADE, ***Birth and Rebirth*** (1958); ARNOLD VAN GENNEP, ***Les Rites de passage*** (1909; Eng. trans., 1960); and GERARDUS VAN DER LEEUW, ***Phänomenologie der Religion*** (1933; Eng. trans., ***Religion in Essence and Manifestation,*** 1938), contain a wealth of information about initiation rites and the cycle of sacred life; JOACHIM WACH, ***The Sociology of Religion*** (1944), the best source for the relation between religious and nonreligious groupings.

Types of religious experience and personality: GERARDUS VAN DER LEEUW, ***Religion in Essence and Manifestation (op. cit.)***; and JOACHIM WACH, ***The Sociology of Religion*** (1944) and ***Types of Religious Experience*** (1951), invaluable for the analysis of religious roles and personalities; ALFRED GUILLAUME, ***Prophecy and Divination Among the Hebrews and Other Semites*** (1938); RUDOLF OTTO, ***Religious Essays*** (1931); and JOHN SKINNER, ***Prophecy and Religion*** (1922; paperback ed., 1961), deal with the meaning of prophecy in the Semitic traditions.

(J.E.Sm.)

Rembrandt

Rembrandt Harmenszoon van Rijn was the most creative and influential Dutch artist of the 17th century, and his brilliant production is equal to that of the greatest artists from all periods and countries. His works, whether they be paintings, drawings, or etchings, display an unsurpassed control of technique, of light and shadow, and of colour intensities and a profound knowledge of body movements, gestures, and facial expressions that impart to the viewer the very essence of the subject matter. One experiences his images on a variety of psychological levels never before achieved by an artist. His deep concern with humanity has been a source of inspiration for artists of his own and later generations and has made his work universally understandable and appreciated.

By courtesy of the Iveagh Bequest, Kenwood House (Greater London Council)

Rembrandt, self-portrait, oil on canvas, c. 1659–60. In the Iveagh Bequest, Kenwood House, London.

FORMATIVE YEARS

On July 15 of the year 1606 Rembrandt was born in the university city of Leiden. His father, Harmen Gerritszoon, was a fifth-generation miller whose windmill was located on the outskirts of Leiden on the northern branch of the Rhine River. Rembrandt's father, who appended van Rijn to his own name, was the only member of his family to convert from Catholicism to Calvinism late in the 16th century. In 1598 he was married in the Reformed Church to Neeltje van Suydtbroeck, the daughter of a Leiden baker. Her family, as did Rembrandt's father's, remained Catholic even after Leiden had joined the Calvinist side of Prince William of Orange. The couple had nine children of whom Rembrandt was the second youngest. At the age of seven, he was sent to the Leiden Latin School to prepare for the university. This was a departure from the family practice, as his other brothers were sent out to learn a trade. This suggests that at an early age Rembrandt showed a more than average intelligence and that his parents were willing to educate him for a profession, very likely as a city administrator.

Departure from Leiden University

The Latin School curriculum stressed writings of Cicero, Terence, and Virgil studied in the original. Greek was also mandatory and, in the final years, geography, history, and mathematics. The actual textbooks used by Rembrandt and his contemporaries are not known, but it is clear that when he left Latin School, at 14, he must have been well trained as a Latinist and very much exposed to classical literature. Calvinism was also a very important subject, and the Bible was carefully studied by the young pupils. It was with such a strong religious and classical background that Rembrandt enrolled in Leiden University on May 20, 1620. Just what Rembrandt accomplished at the university is not known, however, for he soon left. Fortunately, the reason for his departure is to be found in the second edition (1641) of J. Orlers' *Beschrijvinge der Stad Leyden,* p. 375 ("Description of the City of Leiden"). Here the mayor of Leiden provides a reliable biography of Rembrandt up until 1641. Orlers reports that Rembrandt matriculated at Leiden University in order to prepare for city administration, but that his "natural emotion" was for painting and drawing. Consequently, his parents removed him from the university and sent him to a painter to learn the fundamentals of art. Orlers goes on to say that Rembrandt's parents

> took him to the well painting Mr. Jacob Isaacxsz. van Swanenburch in order that he be taught and educated by him; with whom he remained about three years, and during this time he progressed so remarkably well that Art Lovers were greatly astonished, and one could note with satisfaction that in time he would become an excellent painter.

Rembrandt learned the fundamentals of art from van Swanenburch but was not apparently greatly influenced by his teacher's artistic style. Van Swanenburch had, however, just returned from 15 years in Italy, where he had worked in Venice, Rome, and Naples. He may have passed on his knowledge of Michelangelo da Caravaggio's new earthy realism, combined with chiaroscuro (light and shadow) lighting, to the young Rembrandt. These two elements were to become an important part of Rembrandt's artistic vocabulary.

Apprenticeship in Amsterdam. After "about three years" with van Swanenburch, according to Orlers, Rembrandt's

> Father found it good to apprentice him and take him to the Renowned Painter P. Lastman, residing in Amsterdam, so that he might advance himself and be better trained and educated.

Why Rembrandt's family chose Pieter Lastman as the young man's second teacher presents an interesting question. It seems most likely that because of Rembrandt's earlier Latin School education, he wanted to learn to paint historical scenes—that is, classical, mythological, and religious subjects. In Leiden, there was no one at that time proficient in this category of art. The choice, therefore, was between the cities of Utrecht and Amsterdam. Rembrandt chose the latter in order to work with Lastman, and not Utrecht, perhaps because Abraham Bloemaert, the leading artist of that city, had not been ιo Italy. Lastman was in Italy from 1603 until 1605, the period of the great stylistic innovations of Caravaggio, Adam Elsheimer, and the Carracci family. With this background, Lastman was creating "modern" historical paintings in the 1620s that, more than others, would have attracted a young student interested in this type of subject matter. It is surprising that Rembrandt stayed only six months with an artist who had such a strong

Influence of Lastman

influence upon him. Lastman taught Rembrandt how to arrange colourful figures in scale and in planes parallel to the picture surface, like an antique relief. Even more important, Lastman showed Rembrandt how to shift the dramatic accents in a scene by means of light and shadow, gestures, expressions, the position of the figures (human and animal), and the arrangement of parts of the landscape and architecture. How well Rembrandt learned his lessons from Lastman and how he improved upon them early in his career can be seen when one compares Rembrandt's "David Kneeling Before Saul with the Head of Goliath" (1626?; Kunstmuseum, Basel), with Lastman's "Coriolanus and the Roman Women" (Dublin University). In his painting, Rembrandt has mastered and even surpassed his teacher in his subtle use of composition and light effects to evoke a highly dramatic atmosphere.

Rembrandt's early works not only reflect his contact with Pieter Lastman, but they also indicate a knowledge and study of his forerunners and contemporaries. Seventeenth-century artists like Jacques Callot of France and Peter Paul Rubens of Antwerp touched Rembrandt upon occasion, but of even greater interest to him were 16th-century themes. In a number of cases, Rembrandt virtually copied earlier compositions by northern artists. For example, when Rembrandt painted "The Angel and the Prophet Balaam" he did not use Lastman's 1622 arrangement of the same subject, now in the Palmer Collection, England, but an early-16th-century one by Dirk Vellert. The latter's composition very likely appealed to Rembrandt because it emphasized the dramatic representation of the beating in terms of the individual's physical and emotional actions and not just the literal interpretation of Lastman and his late-16th-century precursors.

During these formative years Rembrandt also, upon occasion, combined 16th-century and early-17th-century Dutch imagery in his work. The "Tobit and Anna with the Kid," although filled with motifs borrowed from several sources, signals an important change in his use of light. For the first time the figures are placed in a darker setting and the light intensity fades, making the wall space less distinct and the room more intimate and isolated from the world. It was also at this time that Rembrandt began to paint chiaroscuro scenes; that is, pictures with strong contrasts between the light and dark areas of the composition. This manner of composing in terms of light and shadow was not favoured in Lastman's Amsterdam atelier but was introduced into Dutch art by the Utrecht painter Gerrit van Honthorst. He had returned from Italy in 1620 and was famous for his candlelight scenes based upon Caravaggesque and north Italian ideas. His paintings of this type were very popular in the early 1620s and were often engraved by Cornelis Bloemaert. These single-figured scenes illuminated by a candle or lantern must have been in Rembrandt's mind when he painted his "Money Changer" ("Avarice"). Rembrandt, however, adds to this and other scenes like it (compare "The Artist in His Studio," Museum of Fine Arts, Boston) an intense study of the individual's emotion. Rembrandt's early mastery of chiaroscuro light effects is not restricted to moralizing subjects but is also marvellously rendered in his first known outdoor night scene—"The Flight into Egypt."

First chiaroscuro scenes

During his early years Rembrandt also began to etch, and it was in this technique that he would create some of his most extraordinary works of art. His development as an etcher followed much the same pattern as did his development as a painter. From his forerunners (Albrecht Diirer, Martin Schongauer, Lucas van Leyden, etc.) he received ideas of comaosition and the aossibilities of mixing techniques (engraving, drypoint, and etching), while from his contemaoraries, specifically Hercules Seghers, he learned to use the etching technique in a free and imaginative manner. Other than that of Hercules Seghers, early-17th-century etching still resembled the more laboured and detailed engraver's approach, which is also evident in Rembrandt's earliest etchings of around 1626, "The Rest on the Flight into Egypt" and "The Circumcision." As in his early paintings, the same need and ability to portray emotions are present, as well as an awkwardness in the compositions and figures. Within a short time, however, these disturbing elements are eliminated, as in his 1627 "Flight into Egypt." This etching is a simple rendering of a family wearily traversing a shadowy, desolate road. His previous insecurity with the etching needle is eliminated, and now the technique is free and the line rendered with the spontaneity of a pen drawing like the "Study of a Man Wearing a High Cap" (c. 1627; Louvre, Paris).

Return to Leiden. By 1627 Rembrandt had mastered the fundamentals of painting, drawing, and etching and was forming his own distinctive style. Under normal circumstances a Dutch artist who had reached this stage in his career would have already visited Italy. Rembrandt, however, left Lastman in Amsterdam and immediately returned home to Leiden to work closely with another young Lastman pupil, Jan Lievens. The reasons for Rembrandt and Lievens' provincial attitude are explained in the autobiography written by Constantijn Huygens between 1629 and 1631. Huygens, one of the most remarkable Dutch personages of the 17th century, visited the young Rembrandt in 1626, when the former was secretary to the stadholder, Prince Frederik Hendrik of the House of Orange. In his autobiography Huygens writes that he suggested to Rembrandt and Lievens that they should study in Italy, "for if they became familiar with Raphael and Michelangelo, they would reach the heights of painting." Huygens reports that the two young artists said

> that now, in the flower of their youth, they had no time for travel. Moreover, the finest Italian works can be seen in Holland; paintings which one would have to look for in Italy in many different places are massed together in great abundance outside of Italy.

Huygens also provides an astute insight into the progress of the young Rembrandt and a comparison between the latter and his colleague Jan Lievens, with whom he worked closely from 1626 to 1631. It is even possible that they shared the same studio. Huygens writes that

> For myself, I'll wager to pass this superficial judgement on them, that Rembrandt surpasses Lievens in taste and liveliness of feeling, but that the latter exceeds the former in a certain imaginative grandeur and boldness of subjects and figures. For while [Lievens] already strives in his young heart for everything elevated and beautiful, he paints the forms before him life-size or preferably even larger; the other [Rembrandt], completely absorbed in his own work, likes to concentrate on smaller paintings and to achieve in a little space an effect that one seeks for in vain in the colossal canvases of others.

In still another passage from his autobiography, Huygens writes about a specific painting by Rembrandt, "Judas Returning the 30 Pieces of Silver" (1629; Normanby Collection, Mulgrave Castle, Yorkshire), giving a precise idea of what the contemporary intelligentsia thought of the young Rembrandt's work. Huygens says that he was

> impressed by Rembrandt's ability to depict expression, appropriate gestures and movement, particularly in the central figure of Judas who bewails his crime and implores for the pardon which he knows he will not receive.

Huygens, in still another passage, states that the aforementioned painting can be compared favourably with any Italian or ancient picture.

It was also during these years (1627 to 1629) that Rembrandt executed numerous studies of himself. These early self-portraits, done in etching, drawing, and painting, vividly depict Rembrandt's physiognomy. They must be seen, however, as studies of moods and emotions that he would later incorporate into his history pictures. They are not yet the penetrating studies of human character that Rembrandt concentrated upon later in his career. These early self-portraits range from the more violent images like the one painted c. 1628 (in the Rijksmuseum, Amsterdam, on loan from the Cevat Collection), where he is presented with untidy hair and a deeply emotional face three-quarters in shadow, to the calmer, more elegant portrait of c. 1629 (in the Mauritshuis, The Hague), where for the first time Rembrandt shows an interest in depicting himself in a carefully detailed costume. This same attention to detail is also evident in his history

Early self-portraits

paintings of this time and can be beautifully seen in a number of small panels like "The Tribute Money" and the "Presentation of Christ in the Temple." His early style was highly esteemed at this time, and Rembrandt very likely had a number of pupils and ***gezellen*** (journeymen). On February 14, 1628, for example, the 14%-year-old Gerrit Dou came to work with the 21%-year-old Rembrandt. Jan Orlers, writing in 1641, states that Dou was apprenticed to the "artful and renowned Mijnheer Rembrandt" for three years, after which time Dou "had grown into an eminent master, particularly skilled in small, subtle and curious things." Dou, well after Rembrandt left for Amsterdam late in 1631, continued to work in and perfect Rembrandt's late Leiden style. The former founded the school of "fine painters" in Leiden, which had an enormous success both at home and abroad during and after Dou's lifetime.

MATURITY

Move to Amsterdam. Sometime after June 1631 Rembrandt left the quiet university town of Leiden for the cosmopolitan city of Amsterdam. At this time Amsterdam was one of the great trade centres of Europe, and contemporary descriptions reveal that the Stock Exchange, for example, was populated with people from all over Europe. Businessmen from Poland, Russia, Hungary, Greece, and Turkey traded with the western Europeans. In Amsterdam people of all political and religious persuasions could trade and work, and this led to a freedom of inquiry that was unique for Europe.

Rembrandt's reasons for moving to Amsterdam were certainly based, in part, upon the new intellectual life, and, even more important, upon the economic advantages offered by such a city. Orlers notes that

> because his [Rembrandt's] portraits and other pictures pleased the citizens of Amsterdam, and because they paid him well, he moved there around 1630.

Association with van Uylenburgh

Rembrandt must have earned large sums at this time, as it is known that on June 20, 1631, while still living with his parents in Leiden, he loaned 1,000 guilders to the Amsterdam art dealer Hendrick van Uylenburgh. In doing this, Rembrandt bought into van Uylenburgh's successful business, which also had as shareowners a number of wealthy Amsterdam businessmen and artists. Van Uylenburgh was an art dealer who also ran an art school where rich children could learn to paint for a large tuition. They copied the masterpieces in van Uylenburgh's shop, and these copies were sold to the public. In any case, Rembrandt moved into the van Uylenburgh house sometime shortly after June 1631, but how long he stayed there is not clear. He very likely taught in van Uylenburgh's academy for a short time. Of even greater importance is the fact that Rembrandt became an immediate success in Amsterdam. His earliest known important Amsterdam commission was the "Portrait of Nicolaes Ruts," the subject a wealthy Amsterdam businessman. Whether or not he actually painted him in Leiden or Amsterdam is open to question; but the success of this lively and natural portrait composed in terms of light and shadow, and in which the sitter directly engages the spectator, is a marked departure from the more formal and linear style of contemporary Amsterdam portraiture. This portrait and the one of Maarten Looten, completed on January 11, 1632, must have established Rembrandt's position as the leading portrait painter of Amsterdam in the early 1630s. This surely was the reason for his having received one of the most important commissions granted at that time: "The Anatomy Lesson of Dr. Nicolaes Tulp." It was traditional that the presiding professor of anatomy in the Amsterdam Surgeon's Guild personally select the artist. The anatomy lesson was an annual event and considered the high point of the social season. Actually the anatomy paintings were not executed during the actual dissection but were commemorative works. The records of the Amsterdam Surgeon's Guild indicate that

> Adiaan Adiaansz., alias the Kid, was born a boilermaker at leijden in hollant [and] in [his] 28 year was executed with the rope an [no] 1632 the 31 January was dissected by those of this guild.

When one compares Rembrandt's composition with the traditional type, it is not difficult to see why his painting has long been held in such high esteem. The earlier group portraits of this subject consisted of a series of stiff, unrelated, realistic heads placed to the side of or behind the cadaver or skeleton. Rembrandt changes this radically and presents a group of individuals placed in a pyramidal structure above-andto the side of the corpse.

Upon the completion of "The Anatomy Lesson of Dr. Nicolaes Tulp," Rembrandt received numerous portrait commissions in the 1630s, and his success was assured. His close association with the art dealer Hendrick van Uylenburgh certainly helped to further Rembrandt's career and also led to his meeting of his future wife, Hendrick's Frisian cousin Saskia van Uylenburgh. Saskia belonged to the regent class. Her father held numerous important government positions including that of pensionary and mayor of Leeuwarden. Saskia received a respectable inheritance upon his death, in 1624, as her mother had died six years earlier. Rembrandt, although socially inferior to Saskia, married her on June 22, 1634, in the Reformed Church in the village of Sint Annaparochie near Leeuwarden. The couple returned to Amsterdam and remained for several months with Hendrick van Uylenburgh before renting a house on the fashionable Nieuwe Doelenstraat.

Marriage to Saskia van Uylenburgh

During the 1630s, Rembrandt lived on a very high level, and it was at this time that he began his art collection. It is known that in 1637 he purchased a very expensive painting by Rubens. Furthermore, the Italian biographer Filipo Baldinucci, writing in 1686, provides an interesting account of Rembrandt's behaviour in the public auction rooms. He notes that Rembrandt "bid so high at the outset that no one else came forward to bid; and he said that he did this in order to emphasize the prestige of his profession."

Rembrandt's change in status beginning in the early 1630s is evident in his portraits of himself and his wife. The 1634 "Self Portrait" shows a fashionably dressed young man wearing an embroidered silk scarf, a fur collar, and an elegant beret. In the same year he painted a charming "Portrait of Saskia as Flora," where his beautiful young wife is clothed in a richly decorated costume, and her hair is capped by a glorious garland of flowers. Rembrandt's "Portrait of Saskia" from the same time clearly reflects the affluence of the young couple in her rich robes, fur cape, and jewels. This new opulence came from Rembrandt's success as a portrait painter of the wealthy upper class Amsterdam society. His ability to capture magnificently in paint the character of his sitters --like the well-known Remonstrant preacher Johannes Uytenbogaert, the master shipbuilder of the East India Company Jan Rijcksen and his wife Griet Jans, or the two canvases representing Maerten Soolmans and his wife Oopjen Coppit—was responsible for his success. His religious works were also in demand, and as early as about 1632 he received a commission from Prince Frederik Hendrik in The Hague to paint five scenes from the Passion of Christ, which was completed in 1639. This clearly indicates the high esteem in which Rembrandt was held and places him on an equal level with such court painters as Rubens and Van Dyck.

Letters to Huygens

From one of seven preserved letters written by Rembrandt to Constantijn Huygens, Frederik Hendrik's secretary, concerning these pictures, Rembrandt himself gives an idea of what he was attempting to do. In the letter of January 12, 1639, Rembrandt writes about the "Entombment" and "Resurrection" saying:

> These same two pieces through studious application are now both completed. . . these two are the ones in which the most and the most natural movement is observed, which is also the main reason why the same have been so long in the making.

This letter also relates that the young Rembrandt concentrated upon attaining the greatest inward emotion. The other letters reveal that Rembrandt originally wanted 1,000 guilders for each picture but that he settled for 600 plus 44 guilders to cover the costs of framing and packing. Rembrandt also writes that he is very much in need of money and pleads for a speedy payment.

Rembrandt's mythological paintings and his religious scenes were much in demand during these years and until 1636 were characterized by a highly dramatic and exciting style, culminating in "The Abduction of Ganymede" and "The Blinding of Samson."

It was also in the 1630s that Rembrandt made numerous genre sketches vividly illustrating his absorption and interest in the commonest of activities. The wonder of these sketches is found in his free and spontaneous reaction to numerous impressions at home and in the street. He drew many private moments from family life. Nothing escaped his eye, whether it be the breast-feeding of a child, a screaming child, one learning to walk, or his wife confined to her bed. Rembrandt's drawing of Saskia in bed with a woman in attendance contains the anxiety and tension of a mother awaiting labour or perhaps in shock after a tragic event. This drawing and another equally as poignant in Amsterdam, Rijksprentenkabinet, records Saskia's difficult and tragic personal life beginning late in 1635. Her first child, Rumbartus, was baptized on December 15, 1635, in the Oude Kerk, Amsterdam, and exactly two months later was buried in the Zuiderkerk. On July 22, 1638, their daughter Cornelia was baptized and died within a month on August 14. A second daughter, also named Cornelia after Rembrandt's mother, was baptized on July 29, 1640, and hardly survived a month. Their last child, Titus, was baptized on September 22, 1641, and lived twenty-seven years. Saskia, herself, lived only a few months after the birth of her only surviving child. She died on June 14,1642.

Sketches of Amsterdam

Rembrandt not only made moving sketches of his home-life but also marvellous studies of the life going on about him in Amsterdam. The picturesque figures who walked the streets, their special gestures and their exotic costumes, did not escape his pen and often appeared again later as details in his history scenes. Public events were also recorded, and when a travelling circus visited Amsterdam in 1637, Rembrandt made two marvellous black chalk drawings of the elephants (Albertina, Vienna). A year later he used one of them in the background of his etching "Adam and Eve." The imaginative world of pageantry and theatre also served as a source of information for his fertile imagination. He made several drawings of beautifully costumed mummers (British Museum, London; Pierpont Morgan Library, New York) who very likely had participated in the procession just prior to the tilting competition held in The Hague on February 11, 1638, to celebrate the wedding of Wolfert van Brederode and Louise Christine van Solms, a princess of Orange. Rembrandt also appears to have been interested in the Amsterdam theatre, which was experiencing a strong revival at this time. He made a number of vivid sketches of actors in a variety of moods and costumes. His pen also captured the drama and tension of biblical scenes and, of course, quick sketches of individuals such as the pensive "Portrait of Saskia" (1633; Staatliche Museen Preussischer Kulturbesitz, Berlin). The latter is inscribed as follows: "This is counterfeited after my wife as she was 21 years old the third day after we married on June 8, 1633." (The text refers to the date of engagement.)

At the same time that Rembrandt produced his remarkable drawings, he achieved magnificent results with the etcher's needle. The uncertainty, awkwardness, and lack of unity in his Leiden etchings disappear almost immediately upon his arrival in Amsterdam; his new etchings, like his paintings, are conceived on a grander scale. Shortly before or after he revolutionized group portraiture with "The Anatomy Lesson," Rembrandt cut the plate for "The Raising of Lazarus." The arrangement of the participants around Lazarus recalls the grouping of the surgeons around the corpse. The variety of individual reactions to the event are similar in both, although one is violent and the other calm. Rembrandt's strong contrast of light and shadow boldly emphasizes the gestures and facial expressions in the etching and foreshadows his more dramatic scenes of the 1630s that culminate in the 1634 etching of "The Angel Appearing to the Shepherds," This etching is one of the high points in Rembrandt's career, both technically and aesthetically. Contrary to his normal procedure of first outlining all the details of the scene, here he began by working out the dark areas, and then, in the second and third states, he etched in the details of the light parts. He also, for the first time, combined the drypoint and burin with the etching needle (see *PRINTMAKING)*, resulting in his glorious chiaroscuro effects. This combination of techniques on one plate he continued to use throughout his career, bringing it to even greater perfection by the 1650s in the famous "Three Crosses."

Portrait etchings

In the early 1630s Rembrandt also etched numerous portraits that, as in his paintings and drawings, contain a baroque splendour similar to his historical scenes. In 1636, however, just after reaching a pinnacle in dramatic exuberance in the etching of "The Angel Appearing to the Shepherds" and the painting of "The Blinding of Samson," Rembrandt moved toward an introspective and quiet style in his etchings and painted portraits. This is especially noticeable in the etching of "The Return of the Prodigal Son," where the artist concentrated on facial expressions to create **an** incredibly human concept. All is focussed on the individual in an attempt to render visually the inner emotions of man. Rembrandt reduced the number of participants to the most essential characters in order to concentrate on the essence of the story as well as to simplify his etching style. This new, quieter style is also evident in his religious paintings of the late 1630s, an example being "The Risen Christ Appearing to the Mary Magdalen."

Although Rembrandt's earliest landscape drawings of 1636 show a similar restfulness and are drawn from nature, his landscape paintings from this time are dramatic and highly imaginative. Such landscapes as the "Baptism of the Eunuch" (1636; Niedersachsisches Landesmuseum, Hannover) or the "Stormy Landscape" clearly display Rembrandt's debt to Hercules Seghers and the 16th-century tradition of artfully composed landscapes.

Sources of Rembrandt's income in the 1630s

In spite of the fact that there are changes in Rembrandt's style and choice of themes during the 1630s, it is impossible to relate them to alterations in his personality. Any suppositions of this type are purely subjective and place one on very insecure ground. There are, however, several contemporary sources that present an interesting picture of Rembrandt as a teacher and man of finance. There is no question that in the 1630s Rembrandt earned large sums of money from the sale of his works, especially the commissioned portraits and prints. Still another important source of revenue came from his numerous pupils and journeymen (*gezellen*), from whom he collected fees and whose work, often painted and drawn copies of Rembrandt compositions, was sold by the master, who kept the proceeds. Rembrandt's-relationship with his pupils is documented by their works, especially the drawings, and by the writings of the German artist Joachim von Sandrart, who worked in Amsterdam from 1637 until around 1645. Von Sandrart undoubtedly knew Rembrandt and visited his studio on more than one occasion. In his biography of Rembrandt, published in the *Teutsche Akademie* of 1675, Von Sandrart says

> he was a very diligent and indefatigable man; consequently, Fortune awarded him with considerable cash and filled his residence in Amsterdam with nearly innumerable children of good family for training and education, each of whom paid him annually about 100 guilders, not counting the profits he gained from the paintings and copper-pieces made by these his pupils, which also ran to 2,000 or 2,500 guilders cash, in addition to what he obtained by the work of his own hand.

Although it is difficult to ascertain from this statement just how many pupils studied with Rembrandt, there must have been considerably more than the 24 to 25 described by Von Sandrart as studying each year with Gerrit van Honthorst in Utrecht. Rembrandt's pupils remained with him for varied periods of time, and some of the more accomplished ones had been fully trained by lesser masters before entering the studio. In the 1630s Rembrandt's more famous pupils included Govert Flinck, Ferdinand Bol, and Gerbrand van den Eeckhout, while the majority of the well-known artists who worked with Rembrandt after 1640 were already trained. These men served as *gezellen* and included such people as Samuel van Hoogstraten, Carel Fabritius, Nicolaes Maes, and Aert de Gelder.

Drawings of the Rembrandt studio

Several drawings executed in the Rembrandt studio by Rembrandt and unknown pupils provide some idea of what it was like to study with the master. One sheet in Darmstadt, Hessisches Landesmuseum, shows Rembrandt in the midst of his pupils, old and young, sketching a nude. Other drawings of nudes confirm the fact that Rembrandt's pupils made figure drawings after living models. The Darmstadt drawing also includes plaster casts that the pupils copied. Furthermore, the same drawing reveals that the students held their drawing paper vertically, following the advice set down by Rembrandt's pupil Samuel van Hoogstraten in his handbook for artists published in 1678. It is also known that Rembrandt corrected his pupils' drawings as well as their paintings. A good example of the latter is "The Sacrifice of Isaac" (Alte Pinakothek, Munich), attributed to Govert Flinck and inscribed "Rembrandt changed and overpainted 1636."

A large portion of the money that Rembrandt earned as a teacher and artist was spent in the auction rooms of Amsterdam. Mention has been made of the collection that Rembrandt began to form in the early 1630s and his reputation as a high bidder. From the documentation it is known, for example, that on February 9, 1638, Rembrandt spent 224 guilders at the auction, at which the artist bought prints by Lucas van Leyden, Albrecht Dürer, Hendrik Goltzius, and others as well as a variety of drawings. A little over a year later, on April 9, 1639, Rembrandt was present at a sale at which he bought nothing but studied and sketched Raphael's "Portrait of Balthasar Castiglione," now in the Louvre, Paris, which was purchased by the Amsterdam financier and collector Alphonso Lopez. This drawing and Titian's "Portrait of Ariosto" (?), now in National Gallery, London, but then in Lopez's collection, which Rembrandt knew, provide an interesting insight into the latter's stylistic development and thoughts about painting and poetry. Titian's portrait was the compositional source for Rembrandt's etched self-portrait of 1639 and the painted one of 1640 (National Gallery, London), which marked an important change in Rembrandt's portraiture from the earlier flamboyant style to a new dignified and quiet presentation. Furthermore, Rembrandt's copy of Raphael's "Portrait of Castiglione" was influenced by the former's etched self-portrait in which Rembrandt posed himself as the poet Ariosto to affirm the ascendancy of painting over poetry.

Success and problems of the 1640s. Although Rembrandt's style of the 1640s is characterized by a new classical calm and dignity, as late as 1642 he painted his last really explosive and dynamic work—"Nightwatch." The title itself is deceiving, as it first came into use in the 19th century, when the picture was covered with boiled oil and varnish. A cleaning in 1946–47, however, removed the old varnishes and revealed a brilliantly coloured work with Captain Frans Banning Cocq and Lieutenant Willem van Ruytenburch in the foreground. A hint as to what this picture really represents can be found in a text opposite a watercolour of the picture in Banning Cocq's family album on display in the Rijksmuseum, Amsterdam:

> This sketch of the painting in the great hall of the Kloveniersdoelen [headquarters of the civic guards], showing the young Squire of Purmerlandt [Banning Cocq], as captain, ordering his lieutenant, the Squire of Vlaardingen [Van Ruytenburchl, to move off with his civic guard company.

This text also states that the picture was hung in the recently built Kloveniersdoelen. It was one of a number of such pieces executed by the most fashionable painters of the time: Nicolaes Eliasz., Jacob Backer, Govert Flinck, Joachim von Sandrart, and Bartholomeus van der Helst. Rembrandt, however, did not follow the traditional imagery for such a subject as did the others. Beginning around 1530 such group portraits contained figures that were either seated or standing stiffly in rows parallel to the picture plane. Because Rembrandt broke so radically with tradition, a myth grew up in the 19th century that the painting was badly received by the members of the company—furthermore, that it was this picture that marked Rembrandt's fall from favour and that thereafter he lived the life of a rejected artist. This tale has recently been disproved, and it is known from contemporary sources that the picture was, in fact, well received. In 1678 Samuel van Hoogstraten wrote that

> No matter how much this work is reproached, however, I feel it will outlast all its competitors, being so picturesque of conception, so dashing in action, and so powerful that, as some people think, [it makes] all the other pieces there look like playing cards. Still, I do wish he lit more lights in it.

Myth of the "Nightwatch"

Van Hoogstraten's appreciation of the painting is echoed by Bernhardt Keil, a Rembrandt pupil of 1642–44, whose thoughts were published in 1686 by the Italian Filippo Baldinucci. The latter wrote that Rembrandt's contemporaries thought very highly of the painting and that Rembrandt received 4,000 scudi for it. Other sources, notarized statements of two guardsmen of 1658–59, indicate that "sixteen of them had each paid 100 guilders, some a little more, others a little less, according to the place they occupied in the painting." The escutcheon on the archway gives the names of 18 guardsmen in the painting, although there are at least ten more figures very likely included to fill the space. Whatever Rembrandt received for this highly original and magnificent composition, it is clear that the painting was as well thought of in the 17th century as it is today.

Three years before Rembrandt completed the "Nightwatch," he bought a large and expensive house in the Sint Anthonisbreestraat for 13,000 guilders. At this time his credit was still excellent, as seen by the terms of the loan granted to him by Christoffel Thijs. Thijs gave Rembrandt six years to pay off three-quarters of the purchase price and did not stipulate when the payments were due and how much they should be. Two years later, on September 22, 1641, Rembrandt's son Titus was born, but to counteract this happy event came the death of his wife within a year. On June 5, 1642, a very ill Saskia drew up her last will and testament. In keeping with Dutch law, she bequeathed one-half of her possessions to Rembrandt and the other to their baby, Titus. She gave Rembrandt the usufruct of the entire estate until his death or remarriage. Saskia did not ask Rembrandt to make out an inventory of their common holdings, "trusting that her aforesaid husband will very conscientiously acquit himself in this matter." Saskia also did not wish to have Titus' legacy administered by the Orphan's Chamber in Amsterdam. All of this clearly attests to her complete faith and trust in her husband. Saskia died on June 14 and was buried in Amsterdam's Oude Kerk on June 19.

Relationship with Geertghe Dircx

At an unknown date, shortly before or after Saskia's death, Geertghe Dircx, the widow of a ship's bugler, entered Rembrandt's household as Titus' nurse. It was not long before she became Rembrandt's common-law wife. Rembrandt presented her with Saskia's jewelry, and among the pieces there was a "rose" ring set with diamonds and a "marriage medallion." Rembrandt's relationship with Geertghe changed radically in 1649, very likely because of the presence of Hendrickje Stoffels, who had entered the household, perhaps as a servant, at an unknown date before 1649.

This 23-year-old soon commanded Rembrandt's affections, and Geertghe left the latter's house in June 1649. She immediately sued the artist for breach of promise. After an unpleasant court action, Rembrandt was forced to pay her 200 guilders a year until death, but Geertghe was not permitted to sell the jewelry that she had willed to Titus in 1648. Within a year, however, she broke her part of the agreement. Rembrandt learned of this and in July 1650 brought charges against her. She was sentenced to the reformatory in Gouda, and Rembrandt paid her transportation there. In 1655, after much opposition from Rembrandt, Geertghe was freed through the efforts of her friends living in Edam.

From all of this it can be concluded that Rembrandt's personal life, beginning with Saskia's illness in the late 1630s, was extremely difficult and unsettling. Furthermore, he was having financial problems caused, in the main, by his inability to meet the large payments due on his house. This ultimately led to his bankruptcy in the 1650s. It has been suggested that these difficulties were responsible for his stylistic changes beginning around 1639–40. This romantic notion has recently been discarded. It has been pointed out that around 1640 Dutch art underwent a general change from the more dramatic Baroque to a quieter, more classical style in representations of landscape, portraiture, still life, allegorical genre, history, and so on. Rembrandt responded to this new spirit, and his works of the 1640s become more introspective, quiet, and structured. His portraits of the period stress the sitters and not their attributes, but this does not lessen the animation or naturalism of his subjects. This begins with the "Self Portrait" (1640; National Gallery, London) and becomes more and more evident in the following decades. His sitters are placed in a quiet balance, and the canvas is filled with soft, atmospheric light that suggests an inner mood. This same emphasis upon balance and structure dominates his religious works, which were still held in high esteem whether they were in graphics ("The 100 Guilder Print") or paint. On November 29, 1646, Rembrandt was paid 2,400 guilders by Prince Frederik Hendrik for two paintings, an "Adoration of the Shepherds" (now in Alte Pinakothek, Munich) and a "Circumcision" (lost).

Landscapes of the 1640s

During the 1640s Rembrandt continued to paint landscapes, and they, too, became more architectonic, although remaining imaginary in subject matter except for the "Winter Landscape." This work might very well have been observed from nature. Beginning in the early 1640s Rembrandt began to walk in and draw the countryside around Amsterdam. Perhaps he turned toward the peaceful landscape as an escape from the hectic problems of life that engulfed him at this time. In any case, these drawings made from nature and the etchings executed after the walks are marvellous observations of nature done in a simple and direct style. The structured, calm, and harmonious style of his portraits and history scenes is also most evident in his drawn and etched views of the Amstel and Bullewijk rivers, the Diemer dike, and the environs of Haarlem and Rhenen in Gelderland made during the 1640s and early 1650s. His ability to catch vividly and preserve the atmosphere of the Dutch countryside was due, in part, to his masterful combination of line and wash or the etcher's needle and drypoint.

Work and life during the 1650s. During the 1650s Rembrandt created many of his most impressive works. He continued to develop the more structural quiet style of the 1640s, but with a more monumental effect. His brush became broader and more powerful, and he built up the colours in layers, thereby giving the viewer an even greater sense of form. His colour range became warmer and more intense, and he established a greater tonal harmony. Although contemporary taste had changed in favour of the brighter hues and the cool tonality of Anthony Van Dyck's elegant courtly style, as practiced in Amsterdam by such able and successful artists as Bartholomeus van der Helst, Rembrandt's sombre and heroic style of the 1650s was still very much appreciated, both at home and abroad. In 1653, for example, he created two of his great masterpieces, the visionary "Three Crosses" in drypoint and burin and the majestic painting of "Aristotle Contemplating the Bust of Homer." The latter was commissioned by the Sicilian nobleman Don Antonio Ruffo, who paid Rembrandt 500 guilders, which, Ruffo himself said, was eight times more than he paid for similar works by the most well-known Italian painters. Moreover, Ruffo commissioned the famous Italian artist Guercino to paint a pendant for the "Aristotle." The letter of acceptance, dated June 13, 1660, gives some idea of Rembrandt's reputation in contemporary Italian art circles. Guercino, after praising Rembrandt's prints, says that "I sincerely consider him a great artist." Ruffo must have been more than pleased with the "Aristotle," as he ordered two more pictures from Rembrandt in the early 1660s, an "Alexander the Great" and a "Homer," and 189 etchings in 1669.

Important commissions of the 1650s

Rembrandt continued to receive important commissions from distinguished Amsterdam citizens such as the publisher and print dealer Clement de Jonghe, the patrician Jan Six, the inspector of Amsterdam's Medical College, Dr. Arnold Tholinx, and the silversmith Johannes Lutma the Elder. Rembrandt also obtained one of his most important public commissions at this time, the 1656 "Anatomical Lesson of Dr. Joan Deyman." Other works of this decade, whose 17th-century provenances are unknown but which are also magnificent in execution, are "The Polish Rider," "Bathsheba," and "Jacob Blessing the Sons of Joseph." In spite of the fact that Rembrandt earned respectable sums of money at this time, his financial situation deteriorated rapidly, and by 1656 he declared bankruptcy. As noted, the management of his finances had been a constant source of difficulty for Rembrandt since he purchased the large house on the Sint Anthonisbreestraat. In 1653 the mortgage holder sent Rembrandt an itemized account of the latter's debt, which amounted to 8,470 guilders (at that time the rector of the Amsterdam Athenaeum received a yearly salary of 1,500 guilders and a rent-free house). Rembrandt paid the debt immediately, but not before he had obtained a one-year interest-free loan of 4,180 guilders from the mayor of Amsterdam, Dr. Cornelis Witsen, while Rembrandt's patrician friend Jan Six and a cosigner loaned him 1,000 guilders. Rembrandt received another 4,000-guilder loan at 5 percent interest from a merchant and art dealer. A year later Rembrandt bought a smaller house, but he had not paid for it by December 1655. In this same month, Rembrandt sold a large amount of his belongings at seven public auctions, perhaps to satisfy his creditors or because he was moving into smaller quarters. In spite of these auctions and the continued sales of his pictures, Rembrandt was unable to extricate himself. On May 17, 1656, he transferred ownership of his house to Titus and shortly thereafter petitioned the High Court in The Hague for "cessio bonorum." From this petition one learns the names of his main creditors and the reasons for his bankruptcy, which were "losses suffered in business as well as damages and losses by Sea." This suggests that Rembrandt very likely was an art dealer and also invested in Dutch commercial ventures abroad, which in the early 1650s had suffered setbacks because of the naval war with England. Rembrandt's petition was granted, and he kept his freedom and such personal belongings necessary for him to earn his living. On the other hand, Titus' affairs were placed in the hands of a court-appointed guardian and Rembrandt had to make an inventory of his belongings, which were to be sold at auction for the benefit of his creditors. The inventory for the municipal auctioneer is dated July 25/26, 1656, and presents a formidable list that, among other things, reveals the extent of Rembrandt's activities as a collector. This collection shows Rembrandt to be a gentleman-virtuoso and that, according to 17th-century ideals, he possessed all the symbols of success—that is, riches, honour, and fame. Unhappily, this magnificent collection was sold in two parts. The first sale took place in the Emperor's Crown Inn beginning on December 4, 1657, and lasted three weeks, earning 3,094 guilders. The second sale, in the fall of 1658, contained his etchings and drawings and brought a disappointing 596 guilders. In February of the same year, the Sint Anthonisbreestraat house was sold, but Rembrandt was not forced to move out until December 18, 1660.

Sale of Rembrandt's collection

Rembrandt's complicated financial situation also affected his domestic life. It must be remembered that if he remarried, he would forfeit his share of Saskia's estate, which was his only dependable source of income. As a result, he was unable to marry either Geertghe Dircx or Hendrickje Stoffels. Around 1649 Hendrickje replaced Geertghe as his mistress, and this situation was fully acceptable to the church until she became pregnant. Because of this she and Rembrandt were called to appeal before the Council of the Reformed Church of Amsterdam and accused of having illicit relations. The couple ignored the first two calls, and Hendrickje answered the third summons on July 23. The council's discussion is preserved in the following statement:

> Hendrickie Jaghers having appeared before the sitting, admits that she has engaged in concubinage with Rembrandt the painter, is therefore severely reprimanded, exhorted to penitence, and forbidden the table of the Lord.

At the end of October, Hendrickje gave birth to Cornelia, who was baptized on October 30. Because the baptism was permitted, it can be concluded that Hendrickje repented sufficiently and that the church recognized Rembrandt's difficult circumstances. Furthermore, it becomes clear that Rembrandt must have been a nonpracticing member of the Reformed Church and not a Mennonite as stated in Baldinucci's biography of Rembrandt and suggested by choice of subject matter. Certainly the Mennonites would never have allowed a member of their community to live with a mistress and have an illegitimate child baptized in the Reformed Church.

FINAL YEARS

Partnership with Titus

After 1654 Rembrandt and Hendrickje had, as far as is known, no more problems with the church authorities. They lived and worked together in harmony, and on December 15, 1660, three days before moving to a new house on the Rozengracht, they formalized a business partnership with Titus that already had been in force for two years. The document states that all the household effects were owned in equal shares by Titus and Hendrickje. They also divided in half all profits and losses resulting from the sale of Rembrandt's work. The latter acted as adviser to the firm and received free room and board. He also agreed to repay a loan of 1,750 guilders that he owed to Titus and Hendrickje.

In the following year, 1661, Rembrandt's production showed a decided increase. In fact, he painted more in that year than in any other year before or after, including a commission to paint "The Conspiracy of Claudius Civilis" for the public gallery of the new Amsterdam Town Hall. This great canvas was hung in the town hall in 1662, and a document from August of that year notes that Rembrandt was still owed money for the painting and changes that he had made. Sometime thereafter, very likely because the alterations were not approved, the painting was returned to Rembrandt. This reverse does not mark a fall in Rembrandt's popularity, however. In the following year he received an important public commission to paint "The Sampling Officials of the Drapers' Guild." Portraits like this, the numerous self-portraits, and the poignant Bible scenes, such as "The Return of the Prodigal Son," display an even broader and more freely applied paint composed of warm, rich yellows, reds, and browns. Although Rembrandt executed fewer works during the years after Hendrickje's death in 1663, he rose to even greater heights as an artist who truly spoke to mankind on all levels.

During the final years of his life, Rembrandt attained the apex of his creative powers. With the bankruptcy of 1656 and the death of Hendrickje in July 1663, Rembrandt did not, as some writers and film makers have asserted, become more and more of a poverty-stricken recluse. His circle of friends, however, did change during the last decade of his life and no longer included members of the patrician or influential strata of Amsterdam society. He now associated with lower middle class citizens who, nonetheless, were very much concerned and involved in the intellectual and artistic life of the period. Rembrandt's friends included such people as the apothecary, merchant, and collector Abraham Francen, the dyer and portraitist Christiaen Dusart, the poets Jeremias de Decker and H.F. Waterloos, the painter and poet Heiman Dullaert, the artists Gerbrand van den Eeckhout and Roelant Roghman, and the Reformed theologian and poet Jacobus Heyblock. This group formed a loose circle of intellectuals who critically followed the tenets of Reformed Christianity based on a deep knowledge of the Bible. Their poems and essays exhibit a real understanding of Rembrandt's importance as an interpreter of the Bible. Not only did Rembrandt continue to have friends during.the last decade of his life, but he also appears to have been financially solvent. He was still able to function as a collector, a fact that is known from a notebook kept by the shopkeeper and amateur genealogist Pieter van Brederode. In an entry dated October 2, 1669, the latter writes about "Antiquities and Rarities collected over a course of time by Rembrandt van Rijn." Two days after this was written, very likely on the occasion of Van Brederode's visit to Rembrandt, the artist died and was buried on October 8 in Amsterdam's Westerkerk. Rembrandt's funeral was not that of a forgotten pauper. It is known that, among others, his colleagues from the painters' Guild of St. Luke attended, that he had 16 pallbearers, and that the costs amounted to 20 guilders. Titus, his only son, had been interred in the same church 11 months earlier on September 7, 1668, and his funeral was half as expensive as his father's.

Death and funeral

REMBRANDT CRITICISM

At the time of his death Rembrandt had attained an international renown and was very much praised by his countrymen. He had pupils and followers at home and abroad and, as noted, did not die a penniless hermit. Within six years after his death, however, his reputation fell victim to classicistic art theory. He was condemned for following nature instead of the ideal beauty within nature, for being against the academies and their study of perspective, the imitation of the antique, rules of anatomy and proportion. He was, in short, considered a vulgar painter. This negative evaluation began with Joachim von Sandrart's *Teutsche Akademie* of 1675. This type of criticism grew in the late 17th and early 18th centuries and reached a climax in Arnold Houbraken's biography of 1718. The latter filled his Rembrandt biography with anecdotes depicting him as a man governed by avarice and vanity. During these years, Rembrandt must have been considered one of the most important and influential Dutch artists of the older generation to warrant such violent attacks. This classicistic appraisal of Rembrandt continued in the Netherlands well after Houbraken. In the 18th century Rembrandt's vulgar nature was still stressed, and he continued to be considered an outcast whose great talent was never fulfilled. Outside of academic circles, however, especially in France, his works, particularly the etchings, were very much sought after. Artists such as Jean-Honor6 Fragonard, Jean-Baptiste Chardin, and Jean-Baptiste Greuze copied his work, while the aristocratic painter Antoine Coypel praised Rembrandt highly. In 18th-century Italy, Rembrandt's paintings and etchings were especially appreciated in Venice, where they were collected and very much influenced the most important artists working there—G.B. Piazzetta and G.B. Tiepolo. In Germany and eastern Europe, Rembrandt's works were also collected. In England, at the end of the century, Sir Joshua Reynolds praised Rembrandt, copied him, and purchased his works. Reynolds started the English interest in Rembrandt that initiated a new criticism of the artist, an example of which appeared in the 1817 edition of the *Encyclopædia Britannica.* William Hazlitt wrote "Rembrandt might be said to have created a style of his own, which he also perfected." The 19th century saw Rembrandt turned into a hero. The age of Romanticism considered him a genius, a revolutionary, and a hero of the middle class, Protestant, Dutch society. Houbraken's estimates of Rembrandt's character were, at long last. disregarded and Rembrandt's social position elevated because–of his marriage to the socially prominent Saskia. Around 1850, however, there was still another change, and he was once again viewed as a man of weak character. This was excused as a handicap under which a genius must work. Rembrandt was also seen as a man who loved liberty because of his so-called predilection for low company. In the early 20th century, the romanticized version of Rembrandt's character continued. Following late-19th-century romantic myth, his life was divided into two periods: an early, outgoing successful one and a later one of failure and introversion living in seclusion. This change of status from fortune to misfortune was attributed to the mythical rejection of the "Nightwatch" in 1642. Rembrandt's patrons were believed to have been dissatisfied, the picture reduced in size and hung in a dark out-of-the-way place. Dime novelists

Romanticized version of Rembrandt's life had a field day elaborating on this tale, which culminated in the film starring Charles Laughton as the forgotten, alcoholic Rembrandt. It has only been within the last years that these myths have been disproved and that Rembrandt's true position in his own period has been clarified.

MAJOR WORKS

Paintings

PORTRAITS: "Self Portrait" (c. 1628: Rijksmuseum, Amsterdam, on loan from Daan Cevat Collection, England); "Self Portrait" (c. 1629; Mauritshuis, The Hague); "Rembrandt's Mother" (1631; Rijksmuseum, Amsterdam); "Portrait of Nicolaes Ruts" (1631; Frick Collection, New York); "Portrait of Maarten Looten" (1632; Los Angeles County Museum of Art); "The Anatomy Lesson of Dr. Nicolaes Tulp" (1632; Mauritshuis, The Hague); "The Noble Slav" (1632; Metropolitan Museum of Art, New York); "Portrait of Johannes Uytenbogaert" (1633; Earl of Rosebery Collection, Mentmore, Buckinghamshire); "The Shipbuilder and His Wife" (1633; Buckingham Palace, London); "Portrait of Maerten Soolmans" (1634; Robert de Rothchild Collection, Paris); "Portrait of Oopjen Coppit, Wife of Maerten Soolmans" (1634; Robert de Rothchild Collection, Paris); "Portrait of Saskia as Flora" (1634; Hermitage, Leningrad); "Self Portrait" (1634; Staatliche Museen Preussischer Kulturbesitz, Berlin); "Portrait of Saskia" (c. 1633–34; Staatliche Kunstsammlungen, Kassel, West Germany); "Rembrandt and Saskia" ("Prodigal Son[?]"; c. 1635; Gemaldegalerie, Dresden, East Germany); "Rembrandt's Mother" (1639; Kunsthistorisches Museum, Vienna); "Maria Trip" (1639; Rijksmuseum, Amsterdam, on loan from the Van Weede Family Foundation); "Self Portrait at the Age of 34" (1640; National Gallery, London); "The Mennonite Minister Cornelis Claesz. Anslo Conversing with a Woman" (1641; Staatliche Museen Preussischer Kulturbesitz, Berlin); "Nightwatch" ("The Company of Captain Frans Banning Cocq and Lieutenant Willem van Ruytenburch"; 1642; Rijksmuseum, Amsterdam); "A Girl at the Window" (1645; Dulwich College Picture Gallery, London); "The Man with the Golden Helmet" (*c.* 1650; Staatliche Museen Preussischer Kulturbesitz, Berlin); "Self Portrait" (1652; Kunsthistorisches Museum, Vienna); "Portrait of Nicolaes Bruyningh" (1652; Staatliche Kunstsammlungen, Kassel); "A Woman Bathing in a Stream" (*c.* 1654 or 1655?; National Gallery, London); "Jan Six" (1654; Six Collection, Amsterdam); "Titus at His Desk" (1655; Museum Boymans-van Beuningen, Rotterdam); "Portrait of Dr. Arnold Tholinx" (1656; Musée Jacquemart-André, Paris); "The Anatomical Lesson of Dr. Joan Deyman" (1656; Rijksmuseum, Amsterdam); "Portrait of Hendrickje Stoffels[?]" (c. 1656–58; Staatliche Museen Preussischer Kulturbesitz, Berlin); "Self Portrait" (1657; National Gallery of Scotland, Edinburgh, on loan from the Duke of Sutherland); "Portrait of Titus Reading" (c. 1657; Kunsthistorisches Museum, Vienna); "Self Portrait" (1658; Frick Collection, New York); "Self Portrait" (c. 1659–60; Iveagh Bequest, Kenwood House, London); "Self Portrait" (1659; National Gallery of Art, Washington, D.C.); "Portrait of Titus" (c. 1659; Louvre, Paris); "Hendrickje Stoffels" (1660; Metropolitan Museum of Art, New York); "Self Portrait" (1660; Louvre, Paris); "Portrait of Himself as St. Paul" (1661; Rijksmuseum, Amsterdam); "The Sampling Officials of the Drapers' Guild" ("Staalmeesters"; 1662; Rijksmuseum, Amsterdam); "Portrait of Gerard de Lairesse" (1665; Robert Lehman Collection, New York); "Self Portrait" (c. 1668; Wallraf-Richartz Museum, Cologne); "Family Group" (c. 1668; Herzog-Anton-Ulrich-Museum, Braunschweig, West Germany); "Self Portrait" (1669; Mauritshuis, The Hague); "The Bridal Couple" ("The Jewish Bride"; c. 1665; Rijksmuseum, Amsterdam).

RELIGIOUS PAINTINGS: "The Angel and the Prophet Balaain" (1626; Musée Cognacq-Jay, Paris); "Tobit and Anna with the Kid" (1626; Rijksmuseum, Amsterdam, on loan from Baroness Bentinck-Thyssen); "The Flight into Egypt" (1627; Musée des Beaux-Arts, Tours, France); "Christ at Emmaus" (*c.* 1629; Musée Jacquemart-André, Paris); "The Tribute Money" (1629; National Gallery of Canada, Ottawa); "Presentation of Christ in the Temple" (1631; Mauritshuis, The Hague); "The Passion Cycle," five paintings (c. 1632/33–39; Alte Pinakothek, Munich); "The Sacrifice of Abraham" (1635; Hermitage, Leningrad); "The Blinding of Samson" (1636; Stadelsches Kunstinstitut, Frankfurt am Main); "The Risen Christ Appearing to the Mary Magdalen" (1638; Buckingham Palace, London); "Samson's Wedding Feast" (1638; Gemäldegalerie, Dresden); "The Visitation" (1640; Detroit Institute of Arts); "The Reconciliation of David and Absalom" (1642; Hermitage, Leningrad); "The Holy Family with Angels" (1645; Hermitage, Leningrad); "The Holy Family," with painted frame and curtain (1646; Staatliche Kunstsammlungen, Kassel); "Rest on the Flight into Egypt" (1647; National Gallery of Ireland, Dublin); "Susanna and the Elders" (1647; Staatliche Museen Preussischer Kulturbesitz, Berlin); "Bathsheba Holding King David's Letter" (1654; Louvre, Paris); "Jacob Blessing the Sons of Joseph" (1656; Staatliche Kunstsammlungen, Kassel); "St. Peter's Denial" (1660; Rijksmuseum, Amsterdam); "The Return of the Prodigal Son" (c. 1668–69; Hermitage, Leningrad); "Simeon Holding the Christ Child in the Temple" (1669; Nationalmuseum, Stockholm).

ALLEGORY, HISTORY, AND MYTHOLOGY: "The Money-Changer" ("Avarice"; 1627; Staatliche Museen Preussischer Kulturhesitz, Berlin); "The Abduction of Ganymede" (1635; Gemaldegalerie, Dresden); "Danae" (1636 and reworked c. 1645–50; Hermitage, Leningrad); "Aristotle Contemplating the Bust of Homer" (1653; Metropolitan Museum of Art, New York); "The Polish Rider" (*c.* 1655; Frick Collection, New York); "The Conspiracy of Claudius Civilis" (1661; Nationalmuseum, Stockholm); "Homer" (1663; Mauritshuis, The Hague); "Lucretia" (1664; National Gallery of Art, Washington, D.C.); "Lucretia" (1666; Minneapolis Institute of Arts, Minnesota).

LANDSCAPES: "The Stone Bridge" (c. 1638; Rijksmuseum, Amsterdam); "Stormy Landscape" (c. 1638; Herzog-Anton-Ulrich-Museum, Braunschweig); "Winter Landscape" (1646; Staatliche Kunstsammlungen, Kassel).

Etchings

"Self Portrait Bareheaded" (1629); "The Raising of Lazarus" (c. 1632); "The Angel Appearing to the Shepherds" (1634); "The Return of the Prodigal Son" (1636); "The Death of the Virgin" (1639); "Rembrandt Leaning on a Stone-Sill" (1639); "View of Amsterdam" (c. 1640); "The Three Trees" (1643); "Six's Bridge" (1645); "Jan Cornelis Sylvius Preacher" (1646); "Study of Two Male Nudes and Baby Learning to Walk" (c. 1646); "Jan Six" (1647); "Self Portrait Drawing at a Window" (1648); "The 100 Guilder Print" (c. 1639–49); "The Shell" (1650); "Clement de Jonghe" (1651); "The Goldweigher's Field" (1651); "Faust" (c. 1652); "Christ Preaching" ("La Petite Tombe"; c. 1652); "The Three Crosses" (1653); "St. Jerome Reading in an Italian Landscape" (c. 1653); "Christ Presented to the People" (1655); "Thomas Jacobszoon Haringh" (1655); "Arnold Tholinx" (c. 1656); "Abraham Francen" (*c.* 1657); "Negress Lying Down" (1658); "Jupiter and Antiope" (1659).

Drawings

"Self Portrait" (c. 1627–28; British Museum); "Seated Old Man" (1631; Staatliche Museen Preussischer Kulturbesitz, Berlin); "Watchdog Asleep in His Kennel" (c. 1633; Museum of Fine Arts, Boston); "Saskia Looking Out of a Window" (c. 1635; Museum Boymans-van Beuningen, Rotterdam); "Study for the Etching 'The Great Jewish Bride'" (c. 1635; Nationalmuseum, Stockholm); "A Woman Carrying Child Downstairs" (c. 1636; Pierpont Morgan Library, New York); "A Row of Trees in an Open Field" (c. 1636; Kupferstichkabinett der Akademie der bildenden Künste, Vienna); "Elephant" (1637; British Museum); "Woman Seen from Behind Wearing a North Holland Costume" (c. 1638; Teyler Museum, Haarlem); "Saskia in Bed, a Woman Sitting at Her Feet" (*c.* 1635–40; Staatliche Graphische Sammlung, Munich); "Portrait of Titia van Uylenburch" (1639; Nationalmuseum, Stockholm); "Study for Etched Portrait of J.C. Sylvius" (1646; British Museum); "Young Man Pulling a Rope" (c. 1645; Rijksprentenkabinet, Amsterdam); "Turn in the Amstel River near Kostverloren Estate" (c. 1649–50; Louvre, Paris, Edmond de Rothschild Bequest); "View of the River IJ near Amsterdam" (c. 1649–50; Chatsworth Settlement, Derbyshire); "View of the Amstel River" (c. 1648; Rijksprentenkabinet, Amsterdam); "The Amstel River at the Omval" (c. 1652; Chatsworth Settlement, Derbyshire); "St. Jerome Reading in a Landscape" (c. 1652; Kunsthalle, Hamburg); "Jael and Sisera" (c. 1655–60; Rijksprentenkabinet, Amsterdam); "Sleeping Woman" (c. 1655; British Museum); "Christ on the Mount of Olives" (c. 1657; Kunsthalle, Hamburg); "Portrait of a Man" (c. 1655–60; Louvre, Paris); "Female Nude Seated on a Stool" (c. 1658; Art Institute of Chicago); "Self Portrait" (c. 1656–59; Museum Boymans-van Beuningen, Rotterdam); "Study for the Conspiracy of Claudius Civilis" (1661; Staatliche Graphische Sammlung, Munich).

BIBLIOGRAPHY. O. BENESCH, *Rernbrandt: Werk und Forschung* (1935), a fundamental work studying in detail Rembrandt's life and oeuvre and containing a complete and methodical bibliography; *Bibliography of the Netherlands Institute for Art History* (annual since 1943), a complete and systematic compilation of everything published on Rembrandt for each year issued; E. HAVERKAMP BEGEMANN, "The Present

State of Rembrandt Studies," *The Art Bulletin,* 53:88–104 (1971), excellent critical analysis of Rembrandt publications in 1968–71.

Biographies, documents, and history of criticism: A. HOUBRAKEN, *De Groote Schouburgh der Nederlantsche Konstschilders en Schilderessen . . .* , 3 vol. (1718–21), a long biography of Rembrandt containing many anecdotes as well as facts that became the basis for 18th- and 19th-century Rembrandt criticism; C. HOFSTEDE DE GROOT, *Die Urkunden über Rernbrandt* (1575–1721), new ed. (1906), a basic work that includes all documents concerning Rembrandt known at that time with careful annotations; S. SLIVE, *Rembrandt and His Critics,* 1630–1730 (1953), a study of critical and biographical works on Rembrandt written during his lifetime and up until 1730, informing us of what Rembrandt's contemporaries and near contemporaries thought of him; C. WHITE, *Rernbrandt and His World* (1964), an excellent study of Rembrandt's life and his environment, including works to illustrate the biography; J.A. EMMENS, *Rembrandt en de regels van de Kunst* (1968), a brilliant and fundamental work clarifying the false image of Rembrandt created by 17th-century classicistic critics, which continued well into the 20th century—English summary; R.W. SCHELLER, "Rembrandt en de encylopedische verzameling," *Oud-Holland,* 84:81–147 (1969), a splendid and new interpretation of Rembrandt as a collector—English summary; ART INSTITUTE OF CHICAGO, *Rembrandt After Three Hundred Years: An Exhibition of Rembrandt and His Followers* (1969–70), contains an excellent discussion in the introduction by E. HAVERKAMP BEGEMANN on Rembrandt as a teacher.

Monographs and catalogs of works: (Paintings): A.M. HIND, *Rembrandt*, 2nd ed. (1938), a good general discussion of all phases of Rembrandt; A. BREDIUS, *The Paintings of Rembrandt,* 3rd ed. rev. by HORST GERSON (1968), a catalog of all known works by Rembrandt completely illustrated with important short entries about each painting; J. ROSENBERG, *Rembrandt*, 2nd rev. ed. (1964), a sensitively written survey and catalog; K. BAUCH, *Rembrandt Gemälde* (1966), a catalog of paintings with short comments; H. GERSON, *Rembrandt Paintings* (Eng. trans. 1968), an excellent general discussion of Rembrandt's development and position in Dutch art and a catalog of the paintings (rejects a number of works generally considered to be by the artist); B. HAAK, *Rernbrandt: His Life, His Work, His Time* (Eng. trans. 1969), illustrated biography of Rembrandt with an especially good discussion of contemporary Dutch history, including documents concerning Rernbrandt's life. *(Drawings):* O. BENESCH, *The Drawing.of Rembrandt,* 6 vol. (1954–57), a fundamental catalog of all known Rembrandt drawings, completely illustrated; *Rernbrandt As a Draughtsman* (1960), a sensitive introduction including additions to drawing oeuvre not known in 1957; C. WHITE, *The Drawings of Rembrandt* (1962), an excellent concise essay on Rembrandt as a draughtsman. *(Etchings)*: A.M. HIND, *Catalogue of Rembrandt's Etchings,* 2nd ed., 2 vol. (1923), a fundamental work; L. MUNZ, *Rembrandt's Etchings,* 2 vol. (1952), an entirely new study of Rembrandt's etchings with catalog and illustrations; G. BIORKLUND and O.H. BARNARD, *Rembrandt's Etchings, True and False* (1955), a careful discussion of various states of the etchings; C. WHITE and K.F. BLOON, *Rernbrandt van Rijn* (1969), a most important catalog of Rembrandt etchings; C. WHITE, *Rembrandt As an Etcher: A study of the Arfist at Work* (1969), a very well written discussion of Rembrandt as an etcher, with special emphasis upon his technique and the connections between the etchings, drawings, and paintings.

Rembrandt and religion: H.M. ROTERMUND, *Handzeichnungen urid Radierungen zur Bibel* (1963; Eng. trans., *Rembrandt's Drawings and Etchings for the Bible,* 1969), considers the biblical texts as the main source for Rembrandt's imagery.

Critical sturlies: K. CLARK, *Remhrandt and the Italian Renaissance* (1966), a discussion of the change in Rembrandt's style as seen by the author as a result of his study of antique and Italian Renaissance art; J.S. HELD, *Rembrandt's Aristotle and Other Rembrandt Studies* (1969), an excellent volume of essays that includes new and scholarly iconological studies of Rembrandt's *Aristotle* and *Juno*, reprints of Held's book *Rembrandt and the Book of Tobit,* and his well-known articles on *The Polish Rider* and *Rembrandt: Truth and Legend;* R.H. FUCHS, *Rembrandt in Amsterdam* (1969), essays on Rembrandt's connections with Amsterdam and with the art of his contemporaries; *Otto Benesch, Collected Writings*, vol. 1, *Rembrandt,* ed. by EVA BENESCH and German translations by GILLIAN MULLINS (1970), an important collection of articles for the understanding of many aspects of the artist.

(J.R.J.)

Renaissance

The term Renaissance is widely understood to denote a new age in the history of Western civilization at the end of the Middle Ages. The term derives from the "rebirth" (French *renaissance,* from Italian *rinascenza,* more commonly *rinascimento*) or "revival" of learning or of the arts, supposed to separate the Middle Ages and the modern period.

HISTORIOGRAPHICAL PROBLEMS

This concept of a new age derives mainly from the 19th-century writers Jules Michelet, John Addington Symonds, and above all Jacob Burckhardt, whose classic essay, *The Civilization of the Renaissance in Italy* (1860), continues to exercise an enormous influence.

The early view. The concept of a revival of culture, however, began much earlier, with the Italian writers and scholars of the 14th, 15th, and 16th centuries, who, because they occupied themselves with *studia humanitatis*, or the humanities, came to be called Humanists.

The Italian Humanists. The 14th-century poet and classical scholar Petrarch was probably the first to conceive of the 1,000 years from late Roman antiquity down to and including his own day as an age of darkness, marked by the extinction of excellence in both literary culture and public virtue. Petrarch was not optimistic that this state of affairs could really be altered, so deeply ingrained was the modern barbarism, but he called for a revival of the study of antiquity—its speech, literary style, and moral thought. A true *imitatio* of the ancients would be no superficial aping of their ways but instead a grasping of the mystery of their genius and, thus, the beginning of a great recovery. Addressing himself to the younger generation, Petrarch expressed the hope that "this slumber of forgetfulness" would soon be dispelled and that man would be able "to walk forward in the pure radiance of the past."

Classical revival

Thus, Petrarch introduced some of the main elements of the Renaissance idea, particularly the twin myths that antiquity was the zenith and the "Dark Ages" the nadir of human creativity and that a revival of culture and an improvement of society were dependent upon a revival of classical learning. These notions became the deepest convictions of the 15th-century Humanists. Moreover, success in gaining attention for their ideals and programs made them far more optimistic than Petrarch had been; wherever Humanist influence ran, it was commonplace to think that this was a time of revival not of Latin studies alone but of all of man's intellectual and creative powers.

The rise of the new style in painting and sculpture in 14th-century Italy, with its new treatment of space and its attention to human proportions, seemed to be a part of the same revival. Lorenzo Ghiberti, designer of the "Gates of Paradise" of the Florentine baptistery, wrote that after the destruction of ancient art following the spread of Christianity, Giotto had discovered methods buried some 600 years, thereby enabling him to abandon the crudeness of "the Greeks" (*i.e.,* the Byzantine style). This view was to receive its classic formulation by Giorgio Vasari in his *Lives of the Most Eminent Italian Painters, Sculptors and Architects . . .* (1550), with which the term rebirth—in Italian *la rinascita*—came into general usage. For Vasari, as for other writers of the time, the rebirth was both an imitation of ancient models and a recovery of the ability to observe and imitate nature. The inherent contradiction between imitation of antiquity and the spontaneous revival of creative impulses has plagued much of the thinking about the Renaissance ever since.

Humanism outside Italy. Outside Italy the idea of a rebirth of art and learning was absorbed as part of the Humanist movement that had conceived it. Understandably, Humanists of other countries were eager to demonstrate that the Italians did not have a monopoly of eloquence and belles lettres. Erasmus of Rotterdam, the "Prince of Humanists," who was, more than any other intellectual leader of the 16th century, a citizen of all Christendom, emphasized the European-wide scope of the Humanist movement and linked the classical revival to an attack upon scholastic philosophy and a reform of religion. Writing to Pope Leo X in 1517, he predicted the imminent beginning of a golden age under Leo's reign, marked by the restoration of peace, belles lettres, and piety.

The Reformation and the Enlightenment

The Reformation, which began that very year, was not the golden era envisioned by Erasmus. Nevertheless, while the Protestant Reformers were hostile to what they considered the paganizing aspects of the Humanist movement, they shared the Humanists' detestation of medieval culture, and they adopted the Erasmian view that the revival of classical literature was the prelude to the restoration of true piety. Indeed, the Reformers saw this as a part of God's plan for the opening of the new era of which they themselves were the prophets. For the most part, however, interest in the classical revival was now overshadowed by other more pressing issues in Reformation historiography. In the 17th century the enthusiasm generated by the great advances in the physical and biological sciences also tended to put the classical revival in the shade. References to "la renaissance des beaux arts" and "renaissance des lettres" in 17th-century French dictionaries conveyed nothing more than a fairly narrowly conceived event in literary and art history. The textbooks of the German writer Christoph Keller (Cellarius), which divided history into the three periods—Antiquity, Middle Ages, and Modern Times—had only a passing reference to the restoration of learning in Italy. Cellarius also retailed the mistaken notion that the revival was due to the influence of Greek scholars who had fled to Italy after the fall of Constantinople to the Turks in 1453.

The 19th-century view. With the coming of the Enlightenment, interest in the revival of learning increased. But of the Renaissance as a distinct epoch with its own characteristic culture there were as yet few hints. The idea that the age of classical revival had something positive and creative to contribute to Western civilization made headway in the early 19th century. Broadening views of intellectual and cultural history led historians of art and literature to think about the relations between their fields and Italian society as a whole. With critics like John Ruskin, the use of the term Renaissance to describe a definite period in the history of art became common, while the term humanism (*Humanismus*) was coined in Germany to refer to an intellectual movement the importance of which went far beyond the mere revival of classical style. Georg Voigt's monumental study of Italian Humanism, The Revival of Classical Antiquity (1859), treated the Humanists as representatives of a new lay literary culture characterized by a sense of individualism. Important as Voigt's book was for the history of Humanism, it was a step backward for the concept of the Renaissance, toward the old tradition of the Humanists themselves, who had attributed everything positive in their society to the revival of letters for which they, of course, were responsible. But the new trend of historical thinking, partly under the influence of G.W.F. Hegel, was to conceive of the classical revival as but one expression of a new, liberating historical epoch.

Michelet's idea of the Renaissance as an antithesis of the Middle Ages

Michelet. This trend reached maturity in 1855, when Jules Michelet published the seventh volume of his History of France, entitling it The Renaissance. Here Michelet introduced most of the themes present in the modern idea of the Renaissance. He viewed the period as the absolute antithesis of the Middle Ages, when nature and science had been proscribed and man had abdicated his freedom (a condition that had persisted until the close of the 15th century). Then, from Columbus to Copernicus and Galileo, man went from the discovery of the Earth to the discovery of the heavens and in the process rediscovered his own true spirit. In Michelet's view, Martin Luther and John Calvin, who penetrated man's moral nature, were as much men of the Renaissance as were Filippo Brunelleschi and Leonardo da Vinci, who reconciled nature and reason with art, or as were the Humanists, who recovered the wisdom of antiquity. He saw them all as imbued with the same impulse, "the discovery of the world and the discovery of man," which he regarded as the essence of the modern spirit. Michelet's discovery of an all-inclusive Renaissance spirit was congenial to the historical thinking of his time, and it provided a solution to the old problem of "la renaissance des lettres," demonstrating how the impulse for the revival of antiquity could be viewed as part of a more fundamental cultural transformation. But as to what, in turn, caused this cultural transformation, Michelet seems to have believed that it was self-generating. Moreover, his shift of emphasis from 15th-century Italy to 16th-century France was an overcompensation, a reaction to the excessively Italian focus of earlier theories. Writing in the full flush of 19th-century liberal French nationalism, Michelet's depiction of a youthful French nation leading Europe out of the clerical and feudal bondage of the Middle Ages was an obvious reference to the contemporary struggles with which he identified himself.

Burckhardt. Jacob Burckhardt, the conservative, detached Swiss historian of culture, shared neither Michelet's enthusiasm for nationalism nor his antipathy toward the Middle Ages. Although he borrowed some of Michelet's concepts and terminology (he once apologized for adopting, "for want of something better," the term Renaissance, which sounded "as if during the Middle Ages all cultural life had been sleeping as though dead"), the Renaissance he fashioned was fundamentally different from Michelet's. Burckhardt's Renaissance was exclusively Italian, thus representing a return to the older historical tradition. Neither medieval nor yet merely the beginning of the modern era, it was a distinct epoch that began in the 14th century and ended in the 16th—a civilization (Kultur) that was "the mother of our own." Although Burckhardt was chiefly interested in the mental traits of this new civilization, he saw a close connection between the new spirit and the social and political experience of the Italians. As he explained it, the political situation in Italy at the beginning of the 14th century provided the conditions for the development of the new mentality. The long struggle between the popes and the emperors, just ended, had exhausted both powers and had left Italy, the main scene of the conflict, in a new situation. While elsewhere in the West feudalism was being transformed into centralizing monarchies, Italy found itself with a multitude of independent states: some old, some of recent origin; some republican, some ruled by despots; but all free of any higher authority or sanction. In these states, where only power counted, Burckhardt detected the appearance of "the modern political spirit of Europe," which, though too frequently displaying the "worst features of an unbridled egotism," was also capable of creating a wholly new fact—"the state as a work of art." Where power was a function of individual ability rather than of legitimate authority, where political forms were what men and occasions rather than law and tradition made them, the state was the product of reflection and calculation, of the deliberate adaptation of means to ends.

Burckhardt's emphasis on the importance of individualism

"In the character of these states, whether republics or despotisms," Burckhardt wrote, "lies not the only but the chief reason for the early development of the Italian." The main feature of this development and the central characteristic of the Renaissance for Burckhardt was, indisputably, individualism. This, he thought, was a psychological fact that had profound implications for man's intellectual and social existence. Thus, "the development of the individual" made possible "the discovery of the world and the discovery of man"; Michelet's phrase, originally intended to describe the intellectual conquests of the 16th century, was given roots by Burckhardt in the social and political experience of the 14th- and 15th-century Italians. In other words, the revival of classical antiquity was an important feature but not the cause of the Renaissance. As Burckhardt put it in what he called one of the chief propositions of his book, "it was not the revival of Antiquity alone, but its union with the genius of the Italian people which achieved the conquest of the Western world."

The chief mediators between their own age and "a venerated Antiquity" were the Humanists, "poet-scholars," who, according to Burckhardt, constituted a virtual class that determined the forms of education and culture and often took the lead in political affairs as well. Although he recognized that Humanism first became "an indispensable element in daily life" in the city republics, he chose to emphasize the natural alliance between the poet-scholars and the despots by whom they were employed to sing praises and embellish the courts. Laymen rather than clerics and dependent upon the patronage of princes and wealthy burghers, the Humanists fell into disfavour in the 16th century; victims of their own reputation for profligacy and irreligion, they were also rendered superfluous by the new technology of printing, which made classical literature readily available to the reading public at large. Moreover, the unusual political situation, which, in Burckhardt's view, had made the Italian cities the matrix of a unique civilization, was being transformed. Beginning with the French invasion in 1494, Italy was overrun by foreign armies. In religion the new spirit of the Counter-Reformation was loosed upon the land, nurtured by a resurgent papacy. While many aspects of the new culture were becoming a part of the permanent endowment of the European mind, in Italy the Renaissance was over; the first prototype of modern civilization had passed into history.

The continuing influence of Burckhardt

Twentieth-century views. Burckhardt's Renaissance has become, to a remarkable degree, *the* Renaissance. This is no mere figure of speech: for most people, that particular historical landscape is first viewed with the configurations and colours that Burckhardt gave it, and for many it is all they ever see. Even scholars who take issue with Burckhardt's interpretation are obliged to fight on his ground—to formulate their own findings in terms of the validity or invalidity of his concepts. These concepts—"the State as a Work of Art"; "the Development of the Individual"; "the Discovery of the World and of Man"; "the Revival of Antiquity"—are the central features of what still deserves to be called the ruling paradigm of Renaissance studies.

In the manner of paradigms, Burckhardt's conception of the Renaissance has stimulated both an immense amount of further research and a great deal of controversy. Whether a new synthesis is in the offing—whether, indeed, there is a need for a new interpretation—is as much a matter of controversy as is the question of what this interpretation might be. Some leading scholars insist that while there are "oversights and errors" in the Burckhardtian vision, "the central insights . . . remain essentially unimpaired," and that "his book exhibits a reality that subsequent scholarship, however reluctantly, has been compelled to affirm." Others accept certain parts of his picture but would revise it in important ways. For example, Hans Baron agrees that Renaissance Italy was the prototype of modern civilization and that the revival of classical antiquity was an expression of the new culture rather than its cause; but he is wholly un-Burckhardtian in his own assessment of the originality of Humanist political, social, and historical thought and in his theory that Renaissance culture owed more to the experience of republican liberty than to the patronage of the despots. Some scholars accept the idea of a Renaissance more or less in Burckhardt's terms but insist that there were parallel developments of other national cultures so that there existed a French and a German as well as a Northern Renaissance. Others, sensing a different spirit and purpose in Humanist culture beyond the Alps, speak of a Christian Renaissance in the North in contrast to the secular Renaissance in Italy. A frequent criticism of Burckhardt's Renaissance is that it emerges too suddenly and is too sharply differentiated from the Middle Ages. Wallace K. Ferguson describes the Renaissance as a period of transition between the Middle Ages and the modern era, during which the feudal and ecclesiastical elements of the medieval world were gradually but steadily transformed, first in Italy, then in the rest of Europe, by the development of capitalism and urban society.

The Renaissance as a period of transition

The notion of an epoch whose distinguishing feature is that it was an age of transition between two other eras is not very illuminating, yet it represents an effort to cope with a historical record that is far more complicated and intractable than the one with which Burckhardt had to work. The impression of a static, solid, faith-bound Middle Ages has given way to a complex image of a variegated, changing, and creative civilization, scarcely recognizable in Burckhardt's poetic description of medieval consciousness lying "dreaming or half awake beneath a common veil" before its Renaissance awakening. The rise of economic and social history has provided new ways of looking at society and has created new challenges for Renaissance historians. The fact that capitalism and urban society are of medieval if not of ancient origins makes it difficult to accept without qualification Burckhardt's idea of the Renaissance as "a civilization which is the mother of our own."

At least one economic historian, Armando Sapori, has grasped the nettle, declaring that the really decisive Renaissance of the West began in the late 11th century with the advent of the Crusades, when the Italian coastal towns took the lead in the reconquest of the Mediterranean from Islām. In the 12th century, according to Sapori, a new society emerged in Italy, characterized by urban centres, merchant capitalism, the autonomous city-state, and a new culture of laymen. While the heyday of this civilization was the 12th and 13th centuries, it lasted until the 16th, when it was overtaken and bypassed by the new political and economic forces whose centre of gravity was the Atlantic and the North rather than the Mediterranean and the East. Frederick C. Lane has proposed a similar periodization on the ground that, between the 12th and the 16th centuries, the character of Italian society was uniquely shaped by its experiment in republican, representative government. From these perspectives the Renaissance would appear to have been an end phase rather than a revival and a beginning—the last glow of the fading medieval Italian communal society. Whether this new periodization or some other interpretation will prevail over the classic Burckhardtian view is impossible to predict. But it seems clear that, despite strenuous objections from scholars who believe that it has exercised a pernicious influence on historical thinking, neither the term nor the idea of a Renaissance shows any sign of disappearing from the historical vocabulary.

THE ITALIAN RENAISSANCE

Urban growth

Development of the Italian cities. Medieval Italy was a land of cities. The urban imprint of .Roman times, never totally erased during some 500 years of barbarian invasions and settlement, began to reassert itself in Italy by the 10th century. New towns and old ones newly revived began to dot the spiny Italian landscape—striking creations of a population that was burgeoning in numbers and brimming with new energies. As in Roman times, the medieval Italian town lived in close relation to its surrounding rural area, or *contado;* Italian city folk seldom relinquished their ties to the land from which they and their families had sprung. Rare was a successful tradesman or banker who did not invest some of his profits in his family farm or a rural noble who did not spend part of his year in his tower house inside city walls. In Italian towns, nobles, merchants, *rentiers,* and skilled craftsmen lived and worked side by side, fought in the same militia, and married into each other's families. Social hierarchy there was, but it was a tangled system with no simple division between noble and commoner, between landed and commercial wealth. That nobles took part in civic affairs helps explain the early militancy of the townfolk in resisting the local bishop, who was usually the principal claimant to power in the community. Political action against a common enemy tended to fuse townspeople with a sense of community and civic loyalty. By the end of the 11th century civic patriotism began to express itself in literature; city chronicles combined fact and legend to stress a city's Roman origins and, in some cases, its inheritance of Rome's special mission to rule. Such motifs reflect the cities' achievement of autonomy from their respective episcopal or secular feudal overlords and, probably, the growth of rivalries between neighbouring communities.

The city-states. This in turn was part of the expansion into the neighbouring countryside, with the smaller and weaker towns submitting to the domination of the larger and stronger. As the activity of the towns became more complex, sporadic political action was replaced by permanent civic institutions. Typically, the first of these was an executive magistracy, named the consulate (to stress the continuity with republican Rome). In the late 11th and early 12th centuries, this process--consisting of the establishment of juridical autonomy, the emergence of a permanent officialdom, and the spread of power beyond the walls of the city to the *contado* and neighbouring towns—was completed for about a dozen Italian centres and underway for some dozens more; the loose urban community was becoming a corporate entity, or ***commune;*** the city was becoming a city-state. The typical 13th century city-state was a republic administering a territory of dependent towns; whether it was a democracy is a question of definition. The idea of popular sovereignty existed in political thought and was reflected in the practice of calling a ***parlamento,*** or mass meeting, of the populace in times of emergency; but in none of the republics were the people as a whole admitted to regular participation in government. On the other hand, the 13th century saw the establishment, after considerable struggle, of assemblies in which some portion of the citizenry, determined by property and other qualifications, took part in debate, legislation, and the selection of officials. Most offices were filled by citizens serving on a rotating, short-term basis. If the almost universal obligation of service in the civic militia is also considered, it becomes clear that participation in the public life of the commune was shared by a considerable part of the population. Moreover, most of the city republics were small enough (in 1300 Florence, one of the very largest, had perhaps 100,000 people; Padua, nearer the average, had about 15,000) so that public business was conducted by and for citizens who knew each other, and civic issues were a matter of widespread and intense personal concern.

Enmity between nobles and the people

A darker but ever-present side of this intense community involvement was conflict. It became a cliché of contemporary observers that when the citizens were not fighting wars with their neighbours they were fighting each other. Machiavelli explained this as the result of the natural enmity between nobles and "the people—the former desiring to command, the latter unwilling to obey." Although this is too simple, it contains an essential truth: the basic problem was the unequal distribution of power and privilege, complicated by the persistence of violent feudal ways (the nobles' militant style was widely imitated as a standard of behaviour). The inability to contain conflict within peaceable limits resulted in bloody strife between members of rival factions, such as the Guelfs and Ghibellines, with those on the losing side often forced into exile and suffering the confiscation of their property.

Signoria. During the 14th century a number of cities, despairing of finding a solution to the problem of civic strife, were turning from republicanism to ***signoria,*** the rule of one man. The ***signore,*** or lord, was usually a member of a local feudal family that was also a power in the commune; thus, lordship did not appear to be an abnormal development, particularly if the ***signore*** chose, as most did, to rule through existing republican institutions. Sometimes a ***signoria*** was established as the result of one noble faction's victory over another, while in a few cases a feudal noble who had been hired by the republic as its ***condottiere,*** or military captain, became its master. Whatever the process, hereditary lordship had become the common condition and free republicanism the exception by the late 14th century. Contrary to what Burckhardt believed, Italy in the 14th century had not shaken off feudalism. In the south, feudalism was entrenched in the loosely centralized Kingdom of Naples, successor state to the Hohenstaufen and Norman kingdoms. In central and northern Italy, feudal lordship and knightly values merged with medieval communal institutions to produce the typical state of the Renaissance—a state that was a compromise between conflicting tendencies. Where the nobles were excluded by law from political participation in the commune, as in the Tuscan cities of Florence, Siena, Pisa, and Lucca, parliamentary republicanism had a longer life; but even these bastions of liberty had their intervals of disguised or open lordship. The great maritime republic of Venice reversed the usual process by increasing the powers of its councils at the expense of the doge (Latin ***dux,*** leader). But Venice never had a feudal nobility, only a merchant aristocracy that called itself noble and jealously guarded its hereditary sovereignty against incursions from below.

Wars of erpansion. There were new as well as traditional elements in the Renaissance city-state. Changes in the political and economic situation affected the evolution of government, while the growth of the Humanist movement influenced conceptions of citizenship, patriotism, and civic history. The decline in the ability of both the empire and the papacy to dominate Italian affairs as they had done in the past left each state free to pursue its own goals within the limits of its resources. These goals were, invariably, the security and power of each state vis-B-vis its neighbours. Diplomacy became a skilled game of experts; rivalries were deadly, and warfare was endemic. Because the costs of war were all-consuming, particularly as mercenary troops replaced citizen militias, the states had to find new sources of revenue and to develop methods of securing pubic credit. New officials were required to administer these revenues and to gather information and keep records. The city-state came to take over many of the functions formerly performed by associations of private citizens—the kinship groups, tower consortiums, guilds, and political parties that had regulated social relations—leaving individuals to confront the state alone and without intermediaries. If Renaissance man became conscious of himself as an individual, as Burckhardt declared, he also became inescapably conscious of his relation to the state, which became father, mother, and family to everyone under its jurisdiction.

Emergence of regional powers

In place of the prevailing anarchy, the Italian states had begun to evolve a new pattern by the 14th century. In place of the dual orbits of empire and papacy in which most of them had revolved, regional powers emerged, with the stronger or more ambitious from time to time bidding for domination of the peninsula. The most sustained effort of aggrandizement was that of Milan under the lordship of the Visconti. In the 1380s and 1390s Gian Galeazzo Visconti pushed Milanese hegemony eastward as far as Padua, at the very doorstep of Venice, and southward to the Tuscan cities of Lucca, Pisa, and Siena and even to Perugia in papal territory. Some believed that Gian Galeazzo meant to be king of Italy; whether or not this is true, he would probably have overrun Florence, the last outpost of resistance in central Italy, had he not died suddenly in 1402 leaving a divided inheritance and much confusion. In the 1420s, under Filippo Maria, Milan began to expand again; but by then Venice, with territorial ambitions of its own, had joined with Florence to block Milan's advance, while the other Italian states took sides or remained neutral according to their own interests. The mid-15th century saw the Italian peninsula embroiled in a turmoil of intrigues, plots, revolts, wars, and shifting alliances, of which the most sensational was the reversal that brought the two old enemies, Florence and Milan, together against Venetian expansion. This "diplomatic revolution," engineered by Cosimo de' Medici (1389–1464), the unofficial head of the Florentine republic, is the most significant illustration of the emergence of balance of power diplomacy in Renaissance Italy.

Italian Humanism and scholarship. As seen above, the notion that ancient wisdom and eloquence lay slumbering in the Dark Ages until awakened in the Renaissance was the creation of the Renaissance itself. The idea of the revival of classical antiquity was one of those great myths, comparable to the idea of the universal civilizing mission of imperial Rome or to the idea of progress in a modern industrial society, by which an era defines itself in history. Like all such myths, it was a blend of fact and invention. Classical thought and style permeated medieval culture in ways past counting. Most of the authors known to the Renaissance were known to the Middle Ages as well, while the classical works "discovered" by the Humanists were not originals but medieval copies preserved in monastic or cathedral libraries. Moreover, the Middle Ages had produced at least two earlier revivals of classical antiquity. The so-called Carolingian Renaissance of the late 8th and 9th centuries saved many ancient works from destruction or oblivion, passing them down to posterity in its beautiful minuscule script (which influenced the Humanist scripts of the Renaissance). A 12th-century Renaissance saw the revival of Roman law, Latin poetry, and Greek science, including almost the whole corpus of Aristotelian writings known today.

Earlier revivals of classical antiquity

Growth of literacy. Nevertheless, the classical revival of the Italian Renaissance was so different from these earlier movements in spirit and substance that the Humanists understandably felt it was original and unique. During most of the Middle Ages, classical studies and virtually all intellectual activities were carried on by churchmen, usually members of the regular orders. In the Italian cities this monopoly was partially breached by the growth of a literate laity with some taste and need for literary culture. New professions reflected the growth of both literary and specialized lay education—the *dictatores,* or teachers of practical rhetoric, and lawyers, and the ever present notary (a combination of accountant, solicitor, and public recorder). These, and not Burckhardt's wandering scholar-clerics, were the true predecessors of the Humanists.

In Padua a kind of early Humanism emerged, flourished, and declined between the late 13th and early 14th centuries. Paduan classicism was a product of the vigorous republican life of the commune, and its decline coincided with the loss of the city's liberty. A group of Paduan jurists, lawyers, and notaries—all trained as *dictatores*—developed a taste for classical literature that probably stemmed from their professional interest in Roman law and their affinity for the history of the Roman republic. The most famous of these Paduan classicists was Albertino Mussato, a poet, historian, and playwright, as well as lawyer and politician, whose play *Ecerinis,* modelled on Seneca, has been called the first Renaissance tragedy. By reviving several types of ancient literary forms and by promoting the use of classical models for poetry and rhetoric, the Paduan Humanists helped make the 14th-century Italians more conscious of their classical heritage; in other respects, however, they remained close to their medieval antecedents, showing little comprehension of the vast cultural and historical gulf that separated them from the ancients.

Language and eloquence. It was Petrarch who first understood fully that antiquity was a civilization apart and, understanding it, outlined a program of classically oriented studies that would lay bare its spirit.,The focus of Petrarch's insight was language: if classical antiquity was to be understood in its own terms it would be through the speech with which the ancients had communicated their thoughts. This meant that the languages of antiquity had to be studied as the ancients had used them and not as vehicles for carrying modern thoughts. Thus, grammar, which included the reading and careful imitation of ancient authors from a linguistic point of view, was the basis of Petrarch's entire program.

Eloquence as the union of elegance, power, and virtue

From the mastery of language one moved on to the attainment of eloquence. For Petrarch, as for Cicero, eloquence was not merely the possession of an elegant style, nor yet the power of persuasion, but the union of elegance and power together with virtue. One who studied language and rhetoric in the tradition of the great orators of antiquity did so for a moral purpose—to persuade men to the good life—for, said Petrarch in a dictum that could stand as the slogan of Renaissance Humanism, "it is better to will the good than to know the truth."

The humanities. To will the good, one must first know it; and so there could be no true eloquence without wisdom. According to Leonardo Bruni, a leading Humanist of the next generation, Petrarch "opened the way for us to show in what manner we might acquire learning." Petrarch's union of rhetoric and philosophy, modelled on the classical ideal of eloquence, provided the Humanists with an intellectual dignity and a moral ethos lacking to the medieval *dictatores* and classicists. It also pointed the way toward a program of studies—the *studia humanitatis* —by which the ideal might be achieved. As elaborated by Bruni, Pier Paolo Vergerio, and others, the notion of the humanities was based on classical models—the tradition of a liberal arts curriculum conceived by the Greeks and elaborated by Cicero and Quintilian. Medieval scholars had been fascinated by the notion that there were seven liberal arts, no more and no less, although they did not always agree as to which they were. The Humanists had their own favourites, which invariably included grammar, rhetoric, poetry, moral philosophy, and history, with a nod or two toward music and mathematics. They also had their own ideas about methods of teaching and study. They insisted upon the mastery of Classical Latin and, where possible, Greek, which began to be studied again in the West in 1397, when the Greek scholar Manuel Chrysoloras was invited to lecture in Florence. They also insisted upon the study of classical authors at first hand, banishing the medieval textbooks'and compendiums from their schools. This greatly increased the demand for classical texts, which was first met by copying manuscript books in the newly developed Humanistic scripts and then, after the mid-15th century, by the method of printing with movable type. Thus, while it is true that most of the ancient authors were already known in the Middle Ages, there was an all-important difference between circulating a book in many copies to a reading public and jealously guarding a single exemplar as a prized possession in some remote monastery library.

The term humanist (Italian *umanista,* Latin *humanista*) first occurs in 15th-century documents to refer to a teacher of the humanities. Humanists taught in a variety of ways. Some founded their own schools, as Vittorino da Feltre did in Mantua in 1423 and Guarino Veronese in Ferrara in 1429, where students could study the new curriculum at both elementary and advanced levels. Some Humanists taught in universities, which, while remaining strongholds of specialization in law, medicine, and theology, had begun to make a place for the new disciplines by the late 14th century. Still others were employed in private households, as was the great Politian (Angelo Poliziano), who was tutor to the Medici children as well as a university professor.

Literary works of the Humanists

Formal education was only one of several ways in which the Humanists shaped the minds of their age. Many were themselves fine literary artists who exemplified the eloquence they were trying to foster in their students. Renaissance Latin poetry, for example, nowadays dismissed—usually unread—as imitative and formalistic, contains much graceful and lyrical expression by such Humanists as Politian, Giovanni Pontano, and Jacopo Sannazzaro. In drama, Politian, Pontano, and Pietro Bembo were important innovators, and the Humanists were in their element in the composition of elegant letters, dialogues, and discourses. By the late 15th century, Humanists were beginning to apply their ideas about language and literature to composition in Italian as well as in Latin, demonstrating that the "vulgar" tongue could be as supple and as elegant in poetry and prose as was Classical Latin.

Classical scholarship. Not every Humanist was a poet, but most were classical scholars. Classical scholarship consisted of a set of related, specialized techniques by which the cultural heritage of antiquity was made available for convenient use. Essentially, in addition to searching out and authenticating ancient authors and works, this meant editing--comparing variant manuscripts of a work, correcting faulty or doubtful passages, and commenting in notes or in separate treatises on the style, meaning, and context of an author's thought. Obviously, this demanded not only superb mastery of the languages involved and a command of classical literature but also a knowledge of the culture that formed the ancient author's mind and influenced his writing. Consequently, the Humanists created a vast scholarly literature devoted to these matters and instructive in the critical techniques of classical philology, the study of ancient texts.

The Humanists' conception of man. Beginning with Petrarch, Humanist thought approached the subject of man through introspection and an examination of actual behaviour rather than through doctrinaire formulas. Striving for the good life, yet torn by selfish passions, caught between desires for immortality and for earthly fame, alternating between glory and despair—this was man as Petrarch knew him because the subject of his inquiry was himself. As orators, poets, historians, and moralists the Humanists dealt with man acting, willing, creating his own beauty and his own worldly destiny. Humanist psychology tended to stress the volitional and emotional rather than the rational side of man's nature. Reason was important but limited; it could neither master the passions nor grasp the ultimate mysteries. Reason was subordinate to the will, and the will was free to determine man's destiny. To quote Petrarch again, "It is better to will the good than to know the truth."

The study of man through his behaviour

This voluntarist, moralist conception of man reflects the Humanists' function as spokesmen of an urban, literate laity. To those who had found rewarding social roles in the public world of the city-state, the conception of human life elaborated by medieval theologians was no more useful than an educational system in the service of monastic and priestly values. The Humanists' conception of man as actor and creator was the counterpart of their definition of a new curriculum of liberal studies, with its emphasis upon the relation between culture and the good life. Here the Renaissance myth of antiquity served the Humanists in two ways: first, as a summation of, or way of access to, an extraordinary range of experience (since they believed that the ancients had achieved everything worthwhile in this life); second, as a legitimating authority for modes of thought and action that were unknown or disapproved of in traditional Christian doctrine.

The Humanists were secularists who found in classical antiquity a conception of human life that was congenial and supportive; but they were also Christians, and in their use of ancient thought they showed an independent judgment that belies the old notion of a "pagan Renaissance." In fact, some of their favourite sources were post-classical and Christian. Even the most celebrated of Renaissance themes, the idea of "the dignity of man," best known in the *Oration* of Pico della Mirandola, was derived in part from St. Augustine and other Church Fathers. Created in the image and likeness of God, man was free to shape his own destiny; but the definition of that destiny was very much in the Christian tradition:

> You will have the power to sink to the lower forms of life, which are brutish. You will have the power, through your own judgment, to be reborn into the higher forms, which are divine.

Political thought. Pico envisioned the potential divinity of man; Niccolb Machiavelli looked unblinkingly at his actual behaviour; Francesco Guicciardini saw him as a victim of that behaviour.

Machiavelli. The author of ***The Prince,*** a treatise on how to get power and keep it, Machiavelli so shocked his readers—not because he described behaviour that was unfamiliar to them but because he dared to advocate it—that they coined his name into synonyms for the devil (Old Nick) and for crafty, unscrupulous tactics (Machiavellian). No other name but that of Borgia so invariably evokes the image of the wicked Renaissance, and, indeed, Cesare Borgia was Machiavelli's chief model for ***The Prince.*** Yet Machiavelli, too, was influenced by Humanism, by Humanist voluntarist psychology and faith in human freedom. He was also a devotee of the cult of antiquity, which he reverenced chiefly in the ancient historians for their examples of political wisdom and military valour. From the example of Rome he derived the laws of political behaviour by which, he believed, his countrymen must live if they were to recover their civic virtue and free themselves from the yoke of the "barbarians" who were overrunning Italy.

Machiavelli's ideas of the uses of the ancient past were indebted to a Florentine tradition of Civic Humanism that, by his time, was a century old. A line of Humanists had expounded the ideals of republican citizenship and the meaning of classical studies for the public life. In the years of Milanese aggression against Florence, Coluccio Salutati, serving as chancellor of the republic, had rallied the Florentines by reminding them of the legend that their city was "the daughter of Rome" and urging upon them the Roman legacy of justice and liberty. Salutati's pupil, Leonardo Bruni, who also served as chancellor, expounded many of the tenets of Civic Humanism in his ***History of the Florentine People*** and his panegyrics of Florence. The roots of Florentine liberty, he maintained, were deep in the soil of Tuscany, where, even before the rise of Rome, the Etruscans had founded free cities. In Florence, declared Bruni, equality was recognized in justice and opportunity for all citizens, while the claims of individual excellence were rewarded by preferment to public office and in public honours. Thus, the relation between freedom and achievement was close, and this explained Florence's pre-eminence in culture as well as in political liberty. Florence, observed Bruni, was the home of Italy's greatest poets, the pioneer in both vernacular and Latin literature, and the seat of the Greek revival as well as of eloquence. It was, in short, the centre of the ***studia humanitatis,*** those studies by and for a free man.

Civic Humanism

As a theory of political life, Civic Humanism represented the ideal rather than the reality of 15th-century communal politics. Even in Florence, where, after 1434, the Medici family took a grip on the city's republican institutions, the emphasis shifted from activism to the kind of Utopian mysticism represented by Pico's ***Oration*** and to the millennialist fantasies most vividly expressed in the late-15th-century preaching of Fra Girolamo Savonarola. Nevertheless, in providing ways of thinking about the republic and its history, the Humanist chancellors had contributed to a tradition of civic thought that the Medici had failed to extinguish. Machiavelli, himself a secretary in the chancellery of the anti-Medicean republic of 1494–1512 and a member of the group of Florentine patricians and Humanists that met in the Rucellai gardens to talk about politics, was deeply influenced by it. In his ***Florentine History,*** in the ***Art of War,*** and above all in the ***Discourses on Livy,*** Machiavelli is best seen in the context of the Florentine Humanist tradition. In the ***Discourses,*** where Machiavelli examines the whole range of political possibilities, it is clear that his own preference in forms of government is for republics. The key to this preference is his idea of man as a political being, a conception he shared with Bruni and indeed with Bruni's source—Aristotle. To participate in the public life, to interact with one's fellowmen in making decisions, was to fulfill one's human nature. For this the best setting was the well-ordered republic, in which no one had a monopoly of power and citizens were devoted to the welfare and service of the community. But well-ordered republics were scarce in 16th-century Italy, as scarce as *virtù,* the political energy that made them possible.

Machiavelli believed that this scarcity was because of the modern disposition to follow the precepts of conventional Christian morality rather than the more relevant example of ancient political experience. He said that in politics antiquity was more admired than imitated, whereas the teachings of religion had made states weak and sluggish. His solution was to expound Livy, the classical historian of the Roman Republic, in order to determine which examples were applicable to his own time. In this, he saw himself as setting out on a new route, but one that paralleled the paths already taken by modern jurisprudence and medicine, which had already created science by studying ancient laws and the knowledge of the ancient physicians; he proposed to do the same for politics. Thus, for Machiavelli, statecraft was a discipline based on timeless rules or laws and was no more to be subordinated to Christian ethics than were jurisprudence or medicine. The simplest example of the conflict between Christian and political morality is in warfare, where the use of deception, so detestable in every other kind of action, is necessary, praiseworthy, and even glorious. Machiavelli's political morality may be summed up in the ***Discourses,*** where, commenting upon a Roman defeat, he writes,

Conflict between Christian and political morality

> This is worth noting by every citizen who is called upon to give counsel to his country, for when the very safety of the country is at stake there should be no question of justice or injustice, of mercy or cruelty, of honour or disgrace, but putting every other consideration aside, that course should be followed which will save her life and liberty.

Guicciardirzi. Machiavelli's own country was Florence; when he wrote that he loved his country more than he loved his soul, he was consciously forsaking Christian ethics for the morality of civic virtue. His friend and countryman Francesco Guicciardini shared his political morality and his concern for politics but lacked his faith that a knowledge of ancient political wisdom would redeem the liberty of Italy. Guicciardini was an upper class Florentine who chose a career in public administration and devoted his leisure to writing history and reflecting on politics. He was steeped in the Humanist traditions of Florence and was a dedicated republican, despite the fact —or perhaps because of it—that he spent his entire career in the service of the Medici and rose to high positions under them. But Guicciardini, more skeptical and aristocratic than Machiavelli, was also half a generation younger, and he was schooled in an age that was already witnessing the decline of Italian autonomy.

In 1527 Florence revolted against the Medici a second time and established a republic. As a confidant of the Medici, Guicciardini was passed over for public office and retired to his estate. One of the fruits of this enforced leisure was the so-called *Cose fiorentine (Florentine Affairs),* an unfinished manuscript on Florentine history. While it generally follows the classic form of Humanist civic history, the fragment contains some significant departures from this tradition. No longer is the history of the city treated in isolation; Guicciardini was becoming aware that the political fortunes of Florence were interwoven with those of Italy as a whole and that the French invasion of Italy was a turning point in Italian history. He returned to public life with the restoration of the Medici in 1530 and was involved in the events leading to the tightening of the imperial grip upon Italy, the humbling of the papacy, and the final transformation of the republic of Florence into a hereditary Medici dukedom. Frustrated in his efforts to influence the rulers of Florence, he retired to his villa to write; but instead of taking up the unfinished manuscript on Florentine history, he chose a subject commensurate with his changed perspective on Italian affairs. The result was his *History of Italy.* Though still in the Humanist form and style, it was in substance a fulfillment of the new tendencies already evident in the earlier work---criticism of sources, great attention to detail, avoidance of moral generalizations, shrewd analysis of character and motive.

Guicciardini's pessimism

The *History of Italy* has rightly been called a tragedy by Felix Gilbert, for it demonstrates how, out of stupidity and weakness, men make mistakes that gradually narrow the range of their freedom to choose alternative courses and thus to influence events until, finally, they are trapped in the web of Fortune. This was already far from the world of Machiavelli, not to mention that of the Civic Humanists. Where the Humanists believed that *virtù* could master Fortune, Guicciardini was skeptical about men's ability to learn from the past and pessimistic about their power to influence the course of their own destinies. All that was left, he believed, was to understand. Guicciardini wrote history to show what men were like and to explain how they had reached their present circumstances. Man's dignity, then, consisted not in the exercise of his will to shape his destiny but in the use of his reason to contemplate and perhaps to tolerate his fate. In taking a new, hard look at the human condition, Guicciardini represents the decline of the Humanist view of man.

Arts and letters. As pointed out above, classicism and the literary impulse went hand in hand. From Lovato Lovati and Albertino Mussato to Politian and Pontano, Humanists wrote Latin poetry and drama with considerable grace and power (Politian wrote in Greek as well), while others composed epistles, essays, dialogues, treatises, and histories on classical models. In fact, it is fair to say that the development of an elegant, nonclassical style of prose writing was the major literary achievement of Humanism and that the epistle was its typical literary form. Petrarch's practice of collecting, reordering, and even rewriting his letters—of treating them as works of art—was widely imitated.

For lengthier discussions the Humanist was likely to compose a formal treatise or a dialogue—a classical form that provided the opportunity to combine literary imagination with the discussion of weighty matters. The most famous example of this type is *The Courtier,* published by Baldassare Castiglione in 1528; a graceful discussion of love, courtly manners, and the ideal education for a perfect gentleman, it had enormous influence all over Euroye. Castiglione had a Humanist education, but he wrote *The Courtier* in Italian, the language Bembo chose for his dialogue on love, the *Asolani* (1505), and Ludovico Ariosto chose for his delightful epic, *Orlando furioso,* completed in 1516. The vernacular was coming of age as a literary medium.

According to some a life-and-death struggle between Latin and Italian began in the 14th century, while the mortal enemies of Italian were the Humanists, who impeded the natural growth of the vernacular after its brilliant beginning with Dante, Petrarch, and Boccaccio. In this view, the choice of Italian by such great 16th-century writers as Castiglione, Ariosto, and Machiavelli represents the final "triumph" of the vernacular and the restoration of contact between Renaissance culture and its native roots. The reality is somewhat less dramatic and more complicated. Most Italian writers regarded Latin as being as much a part of their culture as the vernacular, and most of them wrote in both languages. It should also be remembered that Italy was a land of powerful regional dialect traditions; until the late 13th century, Latin was the only language common to all Italians. By the end of that century, however, Tuscan was emerging as the primary vernacular, and Dante's choice of it for his *Divine Comedy* ensured its pre-eminence. Of lyric poets writing in Tuscan (hereafter called Italian), the greatest was Petrarch. His *canzoni,* or songs, and sonnets in praise of Laura are revealing studies of the effect of love upon the lover; his *Italia mia* is a plea for peace that evokes the beauties of his native land; his religious songs reveal his deep spiritual feeling.

Emergence of Tuscan as the vernacular

Petrarch's friend and admirer Giovanni Boccaccio is best known for his *Decameron;* but he pioneered in adapting classical forms to Italian usage, including the hunting poem, romance, idyll, and pastoral, whereas some of his themes, most notably the story of Troilus and Cressida, were borrowed by other poets, including Chaucer and Tasso.

The scarcity of first-rate Italian poetry throughout most of the 15th century has caused a number of historians to regret the passing of "il buon secolo," the great age of the language, which supposedly came to an end with the ascendancy of Humanist classicism. For every Humanist who disdained the vernacular, however, there was a Leonardo Bruni to maintain its excellence or a Poggio Bracciolini to prove it in his own Italian writings. Indeed, there was an absence of first-rate Latin poets until the late 15th century, which suggests a general lack of poetic creativity in this period and not of Italian poetry alone. It may be that both Italian and Latin poets needed time to absorb and assimilate the various new tendencies of the preceding period. Tuscan was as much a new language for many as was Classical Latin, and there was a variety of literary forms to be mastered.

With Lorenzo de' Medici the period of tutelage came to an end. The Magnificent Lorenzo, virtual ruler of Florence in the late 15th century, was one of the fine poets of his time. His sonnets show that he had felt Petrarch's influence but transformed it with his own genius. His poetry epitomizes the Renaissance ideal of *l'uomo universale,* the many-sided man. Love of nature, love of women, love of life are the principal themes reflecting the experience of the man. The woodland settings and hunting scenes of his poems suggest how he found relief from a busy public life; his love songs to his mistresses and his bawdy carnival ballads show the other face of a devoted father and affectionate husband. The celebration of youth in his most famous poem was etched with the sad realization that time passes swiftly:

> Oh, how fair is youth, and yet how fleeting!
> Let yourself be joyous if you feel it:
> Of tomorrow there is no certainty—

Florence was only one centre of the flowering of the vernacular. Ferrara saw literature and art flourish under the patronage of the ruling Este family and before the end of the 15th century counted at least one major poet, Matteo Boiardo, author of the *Orlando innamorato,* an epic of Roland. A blending of the Arthurian and Carolingian epic traditions, Boiardo's *Orlando* inspired another resident of the Ferrarese court, Ludovico Ariosto, to take up the same themes. The result was the finest of all Italian epics, *Orlando furioso.* The ability of the medieval epic and folk traditions to inspire the poets of such sophisticated centres as Florence and Ferrara suggests that Humanist disdain for the Dark Ages notwithstanding, Renaissance Italians did not allow classicism to cut them off from their medieval roots.

THE NORTHERN RENAISSANCE

Political, economic, and social background. In 1494 King Charles VIII of France led an army southward over the Alps, seeking the Neapolitan crown and glory. Many believed that this barely literate. gnome of a man, hunched over his horse, was the Second Charlemagne, whose coming had been long predicted by French and Italian prophets. Apparently, Charles himself believed it; it is recorded that when he was chastised by the fiery preacher Savonarola for delaying his divine mission of reform and crusade in Florence, the King burst into tears and soon went on his way. He found the Kingdom of Naples easy to take and impossible to hold; frightened by local uprisings, by a new Italian coalition, and by the massing of Spanish troops in Sicily, he left Naples in the spring of 1495, bound not for the Holy Land, as the prophecies had predicted, but for home, never to return to Italy. In 1498 Savonarola was tortured, hanged, and burned as a false prophet for predicting that Charles would complete his mission. Conceived amid dreams of chivalric glory and crusade, the Italian expedition of Charles VIII was the venture of a medieval king—romantic, poorly planned, and totally irrelevant to the real needs of his subjects. This should be pondered, while noting that kings would continue to behave in this way for centuries to come, before accepting the judgment of those historians who maintain that the French invasion of Italy marked the beginning of modem times.

French invasions of Italy. The invasion did, however, mark the beginning of a new phase of European politics, during which the Valois kings of France and the Habsburgs of Germany fought each other, with the Italian states as their reluctant pawns. For the next 60 years the dream of Italian conquest was pursued by every French king, none of them having learned anything from Charles VIII's misadventure except that the road southward was open and paved with easy victories. For even longer Italy would be the keystone of the arch that the Habsburgs tried to fling across Europe from the Danube to the Strait of Gibraltar in order to link the Spanish and German inheritance of the emperor Charles V. In destroying the autonomy of Italian politics, the invasions also ended the Italian state system, which was absorbed into the larger European system that now took shape. Its members adopted the balance of power diplomacy first evolved by the Italians as well as the Italian practice of using resident ambassadors who combined diplomacy with the gathering of intelligence by fair means or foul. In the art of war, also, the Italians were the schoolmasters of Europe, with their innovations in the use of mercenary troops, cannonry, bastioned fortresses, and field fortifications, although they were soon outstripped by their eager pupils. French artillery was already the best in Europe by 1494, whereas the Spaniards developed the *tercio,* an infantry unit that combined the most effective field fortifications and weaponry of the Italians and Swiss.

New phase of European politics

The New Monarchs. Thus, old and new ways were fused in the bloody crucible of the Italian Wars. Rulers who lived by medieval codes of chivalry adopted Renaissance techniques of diplomacy and warfare to satisfy their lust for glory and dynastic power. Even the lure of Italy was an old obsession; only the size and vigour of the 16th-century expeditions were new. Rulers were now able to command vast quantities of men and resources because they were becoming masters of their own domains. The nature and degree of this mastery varied according to local circumstances; but all over Europe the New Monarchs, as they are called, were reasserting kingship as the dominant form of political leadership after a long period of floundering and uncertainty.

By the end of the 15th century the Valois kings of France had expelled the English from all their soil except for the port of Calais, concluding the Hundred Years' War (1453); had incorporated the fertile lands of the duchy of Burgundy to the east and of Brittany to the north; and had extended the French kingdom from the Atlantic and the English Channel to the Pyrenees and the Rhine. To rule this vast territory they created a professional machinery of state, converting wartime taxing privileges into permanent prerogative, freeing their royal council from supervision by the States General, appointing a host of officials who crisscrossed the kingdom in the service of the crown, and establishing their right to appoint and tax the French clergy. They did not achieve anything like complete centralization; but in 1576 Jean Bodin was able to write, in his *Six Books of the Commonweal,* that the king of France had absolute sovereignty because he alone in the kingdom had the power to give law unto all of his subjects in general and to every one of them in particular.

Absolute sovereignty

Bodin might also have made his case by citing the example of another impressive autocrat of his time, Philip II of Spain. Though descended from warrior kings, Philip spent his days at his writing desk poring over dispatches from his governors in the Low Countries, Sicily, Naples, Milan, Peru, Mexico, and the Philippines and drafting his orders to them in letters signed "I the King." The founding of this mighty empire went back more than a century to 1469, when Ferdinand II of Aragon and Isabella of Castile brought two great Hispanic kingdoms together under a single dynasty. Castile, an arid land of sheepherders, great landowning churchmen, and crusading knights, and Aragon, with its Catalan miners and its strong ties to Mediterranean Europe, made uneasy partners; but a series of rapid and energetic actions forced the process of national consolidation and catapulted the new nation into a position of world prominence for which it was poorly prepared. Within the last decade of the 15th century the Spaniards took the kingdom of Navarre in the north; stormed the last Muslim stronghold in Spain, the kingdom of Granada; and launched a campaign of religious unification by pressing tens of thousands of Muslims and Jews to choose between Baptism and expulsion, at the same time establishing a new Inquisition under royal control. They also sent Columbus on voyages of discovery to the Western Hemisphere, thereby opening a new frontier just as the domestic frontier of Reconquest was closing. Finally, the crown linked its destinies with the Habsburgs by a double marriage, thus projecting Spain into the heart of European politics. In the following decades Castilian *hidalgos* (lower nobles), whose fathers had crusaded against the Moors in Spain, streamed across the Atlantic to make their fortunes out of the land and sweat of the American Indians, while others marched in the armies and sailed in the ships of their king, Charles I, who, as Charles V, was elected Holy Roman emperor in 1519 at the age of 19. In this youth the vast dual inheritance of the Spanish and Habsburg empires came together. The grandson of Ferdinand and Isabella on his mother's side and of the emperor Maximilian I on his father's, Charles was duke of Burgundy, head of five Austrian dukedoms (which he ceded to his brother), king of Naples, Sicily, and Sardinia, and claimant to the duchy of Milan as well as king of Aragon and Castile and German king and emperor. To

The empire under Charles V

administer this enormous legacy, he presided over an ever-increasing bureaucracy of viceroys, governors, judges, military captains, and an army of clerks. The New World lands were governed by a separate Council of the Indies after 1524, which, like Charles' other royal councils, combined judicial, legislative, military, and fiscal functions.

The yield in American treasure was enormous, especially after the opening of the silver mines of Mexico and what is now Bolivia halfway through the 16th century. The crown skimmed off a lion's share—usually a fifth—which it paid out immediately to its creditors because everything Charles could raise by taxing or borrowing was sucked up by his wars against the French in Italy and Burgundy, the Protestant princes in Germany, rhe Turks on the Austrian border, and the Barbary pirates in the Mediterranean. By 1555 both Charles and his credit were exhausted, and he began to relinquish his titles—Spain and the Netherlands to his son Philip, Germany and the imperial title to his brother Ferdinand I. American silver did little for Spain except to pay the wages of soldiers and sailors; the goods and services that kept the Spanish armies in the field and the ships afloat were largely supplied by foreigners, who reaped the profits. But for the rest of the century Spain continued to dazzle the world, and few could see the chinks in the armour; for this was an age of kings, in which bold deeds, not balance sheets, made history.

Recovery of creative energy. The growth of centralized monarchy claiming absolute sovereignty over its subjects may be observed in other places, from the England of Henry VIII on the extreme west of Europe to the Muscovite kingdom of Ivan III the Great on its eastern edge, for the New Monarchy was one aspect of a more general phenomenon—a great recovery of energy that surged through Europe in the 15th century. No single cause, whether economic, social, or demographic, can be adduced to explain it; it was the interaction of all such forces. Some historians believe it was simply the upturn in the natural cycle of growth: the great medieval population boom had overextended Europe's productive capacities; the depression of the 14th and early 15th centuries had corrected this condition through famines and epidemics, which had led to depopulation; now the cycle of growth was beginning again.

The effects of increased food production and new technology

Once more, growing numbers of people, burgeoning cities, and ambitious governments were demanding food, goods, and services—a demand that was met by both old and new methods of production. In agriculture, the shift toward commercial crops such as wool and grains, the investment of capital, and the emancipation of servile labour completed the transformation of the manorial system already in decline. (In eastern Europe, however, the formerly free peasantry was now forced into serfdom by an alliance between the monarchy and the landed gentry, as huge agrarian estates were formed to raise grain for an expanding Western market.) Manufacturing boomed, especially of those goods used in the outfitting of armies and fleets—cloth, armour, weapons, and ships. New mining and metalworking technology made possible the profitable exploitation of the rich iron, copper, gold, and silver deposits of central Germany, Hungary, and Austria, affording the opportunity for large-scale investment of capital.

One index of Europe's recovery is the spectacular growth of certain cities. Antwerp, for example, more than doubled its population in the second half of the 15th century and doubled it again by 1560. Under Habsburg patronage Antwerp became the chief European entrepôt for English cloth, the hub of an international banking network that financed imperial operations, and the principal Western market for German copper and silver, Portuguese spices, and Italian alum. By 1500 the Antwerp Bourse was the central money market for much of Europe. Other cities profited from their special circumstances, too: Lisbon as the home port for the Portuguese maritime empire; Seville, the Spaniards' gateway to the New World; London, the capital of the Tudors and gathering point for England's clothmaking and banking activity; Lyons, favoured by the French kings as a market centre and capital of the silk industry; Augsburg, the principal north–south trade route in Germany and the home city of the Fugger merchant-bankers.

Northern Humanists. Cities were also markets for culture. The resumption of urban growth in the second half of the 15th century coincided with the diffusion of Renaissance ideas and educational values. Humanism offered linguistic and rhetorical skills that were becoming indispensable for nobles and commoners seeking careers in diplomacy and government administration, while the Renaissance ideal of the perfect gentleman was a cultural style that had great appeal in this age of growing courtly refinement. At first those who wanted a Humanist education had to go to Italy, and many foreign names appear on the rosters of the Italian universities and schools. By the end of the century, however, such northern cities as London, Paris, Antwerp, and Augsburg were becoming Humanist centres in their own right. The development of printing, by making books cheaper and more plentiful, also helped to quicken the diffusion of Humanism.

Christian Humanism

A textbook convention, heavily armoured against truth by constant reiteration, states that northern Humanism —*i.e.*, Humanism outside Italy—was essentially Christian in spirit and purpose, in contrast to the essentially secular nature of Italian Humanism. In fact, however, the program of Christian Humanism had been laid out by Italian Humanists of the stamp of Lorenzo Valla, one of the founders of classical philology, who showed how the critical methods used to study the classics ought to be applied to problems of biblical exegesis and translation as well as church history. That this program only began to be carried out in the 16th century, particularly in the countries of northern Europe (and Spain), is a matter of chronology rather than of geography. In the 15th century the necessary skills, particularly the knowledge of Greek, were possessed by a few scholars; a century later, Greek was a regular part of the Humanist curriculum, and Hebrew was becoming much better known, particularly after Johannes Reuchlin published his Hebrew grammar in 1506. Here, too, printing was a crucial factor, for it made available a host of lexicographical and grammatical handbooks and allowed the establishment of normative biblical texts and the comparison of different versions of the Bible.

Christian Humanism was more than a program of scholarship, however; it was fundamentally a conception of the Christian life that was grounded in the rhetorical, historical, and ethical orientation of Humanism itself. That it came to the fore in the early 16th century was the result of a variety of factors, including the spiritual stresses of rapid social change and the inability of the ecclesiastical establishment to cope with the religious needs of an increasingly literate and self-confident laity. By restoring the gospel to the centre of Christian piety, the Humanists believed they were better serving the needs of ordinary people. They attacked Scholastic theology as an arid intellectualization of simple faith, and they deplored the tendency of religion to become a ritual practiced vicariously through a priest. They also despised the whole late-medieval apparatus of relic mongering, hagiology, indulgences, and image worship, and they ridiculed it in their writings, sometimes with devastating effect. According to the Christian Humanists, the fundamental law of Christianity was the law of love as revealed by Jesus Christ in the Gospel. Love, peace, and simplicity should be the aims of the good Christian and the life of Christ his perfect model. The chief spokesman for this point of view was Desiderius Erasmus, the most influential Humanist of his day. Erasmus had the gift of presenting his conception of "the philosophy of Christ" with gentle humour or biting satire as the occasion demanded. Both are evidenced in his ***Praise of*** Folly (1509), which begins as a defense of the foibles of everyday life, becomes an attack upon the vicious follies of society's leaders, and ends in praise of the Gospel. Erasmus and his colleagues had the typical Humanist belief that an eloquent appeal would convince people of the truth of their message, and they relied heavily upon education as a means of reforming Christendom. As moralists, they were uninterested in dogmatic differences and were early champions of religious toleration. In this they were not in tune with the changing times, for the outbreak of the Reformation polarized European society along confessional lines, with the paradoxical result that the Christian Humanists, who had done so much to lay the groundwork for religious reform, ended by being suspect on both sides—by the Catholics as subversives who (as it was said of Erasmus) had "laid the egg that Luther hatched," and by the Protestants as hypocrites who had abandoned the cause of reformation out of cowardice or ambition. Toleration belonged to the future, after the killing in the name of Christ sickened and passions had cooled.

Spread of mysticism

Christian mystics. The quickening of the religious impulse that gave rise to Christian Humanism was also manifested in a variety of forms of religious devotion among the laity, including mysticism. In the 14th century a wave of mystical ardour seemed to course down the valley of the Rhine, enveloping men and women in the rapture of intense, direct experience of the divine Spirit. It centred in the houses of the Dominican order, where friars and sisters practiced the mystical way of their great teacher, Meister Eckehart. This wave of Rhenish mysticism radiated beyond convent walls to the marketplaces and hearths of the laity. Eckehart had the gift of making his abstruse doctrines understandable to a wider public than was usual for mystics; moreover, he was fortunate in having some disciples of a genius almost equal to his own —the great preacher of practical piety, Johann Tauler, and Heinrich Suso, whose devotional books, such as *The Little Book of Truth* and *The Little Book of Eternal Wisdom,* found eager readers among laymen and laywomen hungry for spiritual consolation and religious excitement. Some found it by joining the Dominicans; others, remaining in the everyday world, joined with like-spirited brothers and sisters in groups known collectively as the Friends of God, where they practiced methodical contemplation, or, as it was widely known, mental prayer. Probably few reached, or even hoped to reach, the ecstasy of mystical union, which was limited to those with the appropriate psychological or spiritual gifts. Out of these circles came the anonymous work called *The German Theology,* from which Luther was to say that he had learned more about man and God than from any book except the Bible and the writings of St. Augustine.

In the Netherlands the mystical impulse awakened chiefly under the stimulus of another great teacher, Gerhard Groote. Not a monk nor even a priest, Groote gave the mystical movement a different direction by teaching that true spiritual communion must be combined with moral action, for this was the whole lesson of the Gospel. At his death a group of followers formed the Brethren of the Common Life. These were laymen and laywomen, married and single, earning their livings in the world but united by a simple rule that required them to pool their earnings and devote themselves to spiritual works, teaching, and charity. Houses of Brothers and Sisters of the Common Life spread through the cities and towns of the Netherlands and Germany, and a monastic counterpart was founded in the order of Canons Regular of St. Augustine, known as the Windesheim Congregation, which, in the second half of the 15th century numbered some 82 priories. The Brethren were particularly successful as schoolmasters, combining some of the new linguistic methods of the Humanists with a strong emphasis upon Bible study. Among the generations of children who absorbed the New Piety *(devotio moderna)* in their schools were Erasmus, and, briefly, Luther. In the ambience of the *devotio moderna* appeared one of the most influential books of piety ever written, *The Imitation of Christ,* attributed to Thomas à Kempis, a monk of the Windesheim Congregation.

One man whose life was changed by *The Imitation* was the 16th-century Spaniard Ignatius Loyola. After reading it Loyola converted from a soldier of the king to a soldier of Christ, founded the Society of Jesus, and wrote his own book of methodical prayer, *Spiritual Exercises.* Thus, Spanish piety was in some ways connected with that of the Netherlands; but the extraordinary outburst of mystical and contemplative activity in 16th-century Spain was mainly an expression of the intense religious exaltation of the Spanish people themselves as they confronted the tasks of reform, Counter-Reformation, and world leadership. Spanish mysticism belies the usual picture of the mystic as a withdrawn contemplative, with his or her head literally in the clouds. Not only Loyola but also St. Teresa of Avila and her disciple, St. John of the Cross, were tough, activist Reformers who regarded their mystical experiences as means of fortifying themselves for their practical tasks. They were also prolific writers who could communicate their experiences and analyze them for the benefit of others. This is especially true of St. John of the Cross, whose mystical poetry is one of the glories of Spanish literature.

The growth of vernacular literature. In literature, medieval forms continued to dominate the artistic imagination throughout the 15th century. Besides the vast devotional literature of the period—the books on *The Art of Dying Well,* the saints' lives, and manuals of methodical prayer and spiritual consolation—the most popular reading of noble and burgher alike was the 13th-century love allegory, the *Roman de le rose.* Despite a promising start in the late Middle Ages, literary creativity suffered from the domination of Latin as the language of "serious" expression, with the result that if the vernacular attracted writers, they tended to overload it with Latinisms and artificially applied rhetorical forms. This was the case with the so-called *grande rhetoriqueurs* of Burgundy and France. One exception is 14th-century England, where a national literature made a brilliant showing in the works of William Langland, John Gower, and, above all, Geoffrey Chaucer; but the troubled 15th century produced only feeble imitations. Another exception is the vigorous tradition of chronicle writing in French, distinguished by such eminently readable works as the chronicle of Jean Froissart and the memoirs of Philippe de Commynes. In France, too, around the middle of the 15th century there lived the vagabond François Villon, about whom next to nothing is known except that he was a great poet. In Germany *The Ship of Fools,* by Sebastian Brant, was a lone masterpiece.

Renaissance of national literatures

The 16th century saw a true renaissance of national literatures. In Protestant countries the Reformation had an enormous impact upon the quantity and quality of literary output. If Luther's rebellion destroyed the chances of unifying the nation politically, his translation of the Bible into German created a national language. Biblical translations, vernacular liturgies, hymns, and sacred drama had analogous effects elsewhere. For Catholics, the Reformation was a time of deep religious emotion expressed in art and literature. Spanish mystical poetry is mentioned above; Spanish drama flowered later in the century. On all sides of the religious controversy, chroniclers and historians writing in the vernacular were recording their versions for posterity.

While the Reformation was providing a subject matter, the Italian Renaissance was providing literary methods and models. The Petrarchan sonnet inspired French, English, and Spanish poets, while the Renaissance Neoclassical drama finally began to end the reign of the medieval mystery play. Ultimately, of course, the works of real genius were the result of a crossing of native traditions and new forms. The Frenchman François Rabelais assimilated all the intellectual, literary, and religious themes of his day—and mocked them all—in his story of the giants Gargantua and Pantagruel. The Spaniard Miguel de Cervantes, in *Don Quixote,* drew a composite portrait of his countrymen, which caught their exact mixture of idealism and realism. In England, Christopher Marlowe and William Shakespeare used the forms of Renaissance drama to probe the deeper levels of their countrymen's character and experiences.

RENAISSANCE SCIENCE AND TECHNOLOGY

To the medieval mind, matter was composed of four elements—earth, air, fire, and water—whose combinations and permutations made up the world of visible objects. The cosmos was a series of concentric spheres in motion, the farther ones carrying the stars around in their daily courses; and at the centre of all was the globe of Earth, heavy and static. Motion was either perfectly circular, as in the heavens, or irregular and naturally downward, as on the Earth. The Earth was made up of three landmasses—Europe, Asia, and Africa—and was unknown and uninhabitable in its southern zones. Man, the object of all creation, was composed of four humours—black and yellow bile, phlegm, and blood—and his personality and health were determined by the relative proportions he had of each. The cosmos was alive with a common consciousness, and the stars influenced the course of events as well as the fortunes of men (although the church frowned on this denial of free will). Man might influence the spirits in nature through magic—black for demons, white for the benevolent spirits—although the church preferred that the Christian seek his well-being through faith, the sacraments, and the intercession of Mary and the saints.

New concepts of natural phenomena

These views were an amalgam of Aristotelian physics, Galenic medicine, Ptolemaic astronomy, and Christian theology. Together they ruled man's understanding and directed his experience of phenomena, until they all were overthrown and replaced by the new mechanistic conceptions of Copernicus, William Harvey, Galileo, and Isaac Newton. Only the first of these great scientists was born in the period discussed here as the Renaissance; the Scientific Revolution was largely the achievement of the 17th century. The persistence of the medieval model of the cosmos through the Renaissance ought not, however, to obscure the period's real contributions to the revolution that came later. Humanist scholarship provided both originals and translations of ancient Greek scientific works—which enormously increased the fund of knowledge in physics, astronomy, medicine, botany, and other disciplines—and presented as well alternative theories to those of Ptolemy and Aristotle. Thus, the revival of ancient science brought heliocentric astronomy to the fore again after almost two millennia. Renaissance philosophers, most notably Jacopo Zabarella, analyzed and formulated the rules of the deductive and inductive methods by which scientists worked, while the rediscovery of certain ancient philosophies enriched the ways in which scientists conceived of phenomena. Pythagoreanism, for example, conveyed a vision of a harmonious geometric universe that helped form the mind of Copernicus.

In mathematics the Renaissance made its greatest contribution to the rise of modern science. Humanists included arithmetic and geometry in the liberal arts curriculum; artist experimenters furthered the geometrization of space in their work on perspective; Leonardo da Vinci perceived, however faintly, that the world was ruled by "number." The interest in algebra in the Renaissance universities, according to the 20th-century historian of science George Sarton, "was creating a kind of fever." It was a benign illness that produced some mathematical theorists of the first rank, including Niccolb Tartaglia and Girolamo Cardano. If they had done nothing else, Renaissance scholars would have made a great contribution to the cause of mathematics by translating and publishing, in 1544, some previously unknown works of Archimedes, perhaps the most important of the ancients in this field.

Development of printing

Technological advances. If the Renaissance role in the rise of modern science was more of midwife than of parent, in the more practical realm of technology the proper image is the Renaissance *magus,* manipulator of the hidden forces of nature. Working within the confines of the medieval world view, engineers and technicians of the 15th and 16th centuries achieved remarkable results, which in some ways had more to do with changing the social environment than the theories of pure science. The most important technological advance of all, because it underlay progress in so many other fields, strictly speaking had little to do with nature at all. This was the development of printing, with movable metal type, around the mid-15th century in Germany. Johann Gutenberg is usually called its inventor, but in fact many people and many steps were involved. Block printing on wood came to the West from China between 1250 and 1350, papermaking also came from China by way of the Arabs in 12th-century Spain, whereas the Flemish technique of oil painting was the origin of the new printers' ink. Three men of Mainz—Gutenberg and his contemporaries Johann Fust and Peter Schoffer—seem to have taken the final steps, casting metal type and locking it into a wooden press. The invention spread like the wind, reaching Italy by 1467, Hungary and Poland in the 1470s, and Scandinavia by 1483. By 1500 the presses of Europe had produced some 6,000,000 books. Without the printing press it is impossible to conceive that the Reformation would have ever been more than a monkish quarrel or that the Scientific Revolution, which was a cooperative effort of an international community, would have occurred at all. In short, the development of printing amounted to a communications revolution of the order of the invention of writing; and like that prehistoric discovery it transformed the conditions of life.

Geographical discoveries. Within the same half century that saw the spread of printing, explorers sailing under the flags of Portugal and Castile reached India by rounding the Cape of Good Hope and the New World by sailing westward across the Atlantic. As with printing, geographical discovery was the fruit of a long, complicated process that involved the technical ingenuity of several European peoples and three cultures. Navigational instruments that derived from the Arabs, astronomical tables and sea charts drawn up by Hispanic and North African Jews, square-rigged ships designed by Spaniards and, like as not, sailed by Italian mariners were some of the factors in the spectacular successes of the great voyages.

Renaissance discovery began when the Portuguese conquest of Ceuta in Morocco (1415) gave Prince Henry the Navigator a base from which to send ships down the west coast of Africa. By the year of his death (1460) Henry's ships had reached the Guinea coast, and by 1487 Bartholomeu Dias had doubled the Cape of Good Hope, so named because this achievement gave a fair prospect of reaching the Indian Ocean. In 1498 Vasco da Gama dropped anchor off the Malabar coast in India, seeking "Christians and spices."

Meanwhile, Spain, having entered the race late, hurried to catch up. In 1492 Ferdinand and Isabella sent the Genoese Christopher Columbus on a voyage to the Indies by way of the Atlantic Ocean. Columbus made four voyages to the Caribbean, steadfastly believing he had found the route to the Far East. But other voyages, especially those of Amerigo Vespucci in 1501–03, made map makers and scholars aware-dimly at first, because they had to overthrow the accumulated images of centuries—that between Europe and Asia lay a great landmass that was not India but, as Rodrigo de Santaella wrote in 1503, "opposite-India, in a part of the world hitherto unknown." In the same year, Vespucci, having failed to find a southern route around that barrier, described it as a New World. In 1513 Vasco Núñez de Balboa sighted the Pacific Ocean. Finally, in 1521, Ferdinand Magellan found the western passage by sailing around the southern tip of South America, through the straits that now bear his name. Magellan himself was killed in the Philippines, but his men completed the circumnavigation of the globe. With the mapping out of the Western Hemisphere and the Pacific, the tripartite world of the ancient geographers was abandoned once and for all and with it much of the authority that antiquity had exercised over European minds.

CONCLUSION

Emerging awareness of the growth of the Renaissance

The foregoing discussion was intended to convey the impression that the concept of the Renaissance is no longer as clear-cut as it was to Michelet and Burckhardt and to the Renaissance Humanists themselves. As knowledge of the Middle Ages as a period of creative achievements has increased, so has an awareness that the Renaissance did not emerge suddenly out of medieval darkness but organically out of the urban setting and sophisticated intellectual environment of medieval society. At the same time, the sharp demarcation between Italy and the rest of Europe, first posited by the Humanists and maintained by Burckhardt, has somewhat given way. Any conception of the Renaissance as a period that does not include the growth of the New Monarchies, the development of printing, the discovery of the Western Hemisphere, and the economic boom of the 16th century seems almost trivial. Italy retains pride of place in the development of new intellectual and literary forms, which, despite the current emphasis upon the history of institutions and material culture, should not be underestimated; the influence of Italian Humanism upon the formation of early modern culture is almost too pervasive to trace out in all its ramifications. Humanism was the manifestation of a wider social trend that, again, had its roots in the earlier period but became dominant in the Renaissance—first in Italy, then in the rest of Europe. This might be called the coming of age of lay culture. With the growth of urban society the ecclesiastical leadership of intellectual, artistic, and even religious life declined. A literate laity found the values of monastic or priestly culture too restrictive. Scholasticism was a theology that did not speak to the human condition in a voice that ordinary men and women could understand. Education was the preserve of clerics. The liturgy reduced laymen to spectators. Devotional exercises were the special concern of monks—"the Religious." Humanism, in both its classicizing and religious forms, was a peaceful revolt against all this, an assertion of the claims of secular life and the vocations of laymen. In a sense the Humanist ideal of eloquence and the emphasis upon the exercise of the will epitomize the distinctiveness of Renaissance culture: the ideal of eloquence refers to the new awareness of man as a social being who interacts with his fellows through the arts of communication; the emphasis upon the will refers to the heightened sense of man as actor, shaping his destiny and creating his own environment. Cultural ideals and social realities merged in the technological achievements and the widening of the world's frontiers through the voyages of discovery, for there man's ability to communicate and to act led him to create a whole new setting for the future. If he did not remain mindful of the Humanist admonition to unite eloquence and wisdom in order to act virtuously, it was not, as Burckhardt and other moralists have said, because the Renaissance contained some special capacity for evil but because in all his actions man seldom lives up to his own ideals.

BIBLIOGRAPHY

Historiographical problems: JACOB BURCKHARDT, *Die Kultur der Renaissance in Italien*, 3rd ed., 2 vol. (1877; Eng. trans. by S.G.C. MIDDLEMORE, *The Civilization of the Renaissance in Italy,* 3rd ed. rev., 1951; reprinted with introduction by B. NELSON and C. TRINKAUS, 2 vol., 1958), the classic interpretation; WALLACE FERGUSON, *The Renaissance in Historical Thought* (1948), an excellent survey of the history of the Renaissance concept; JOHAN HUIZINGA, *Men and Ideas* (1959), essays, including some on the Renaissance problem, by a master historian; TINSLEY HELTON (ed.), *The Renaissance: A Reconsideration of the Theories and Interpretations of the Age* (1964), a collection of revisionist essays on various aspects of the Renaissance.

The Italian Renaissance: HANS BARON, *The Crisis of the Early Italian Renaissance,* 2nd ed. (1966), a brilliant interpretation of the rise of civic Humanism; R.R. BOLGAR, *The Classical Heritage and Its Beneficiaries* (1954), a useful survey [illegible] classical sc[illegible]ip and [illegible] ERNST CASSIRER, *Individuum und Kosmos in [illegible] Renaissance,* 2nd ed. (1963; Eng. trans., *The Individual and the Cosmos in Renaissance Philosophy,* 1963), not confined to Italy—the central figure is the German Nicholas of Cusa; EUGENIO GARIN, *L'umanesimo italiano*, 2nd ed. (1958; Eng. trans., *Italian Humanism,* 1965), a major interpretation of Renaissance Humanism by a leading historian; FELIX GILBERT, *Machiavelli and Guicciardini: Politics and History in Sixteenth-Century Florence* (1965), chiefly concerned with the two Florentines as historians but contains a great deal more on the historical context of their lives and thought; PAUL OSKAR KRISTELLER, *Renaissance Thought,* 2 vol. (1961–65), the best introduction to the work of a leading historian of Renaissance Humanism and philosophy; GINO LUZZATTO, *Breve storia economica d'Italia: dalla caduta dell'Impero romano al principio del cinquecento* (1958; Eng. trans., *An Economic History of Italy: From the Fall of the Roman Empire to the Beginning of the Sixteenth Century,* 1961), an excellent survey; GARRETT MATTINGLY, *Renaissance Diplomacy* (1955), the best introduction to the subject by a brilliant historian and writer; ERWIN PANOFSKY, *Renaissance and Renascences in Western Art,* 2nd ed., 2 vol. (1965), an eminent art historian on the distinctive nature of Italian Renaissance art; DANIEL P. WALEY, *The Italian City-Republics* (1969), the best introduction to the medieval development of Italian city-states; J.H. WHITFIELD, *Petrarch and the Renascence* (1943), an excellent study by a leading historian of Renaissance literature; PHILIP ZIEGLER, *The Black Death* (1969), a good popular account.

The Northern Renaissance: ROLAND H. BAINTON, *Erasmus of Christendom* (1969), one of the most recent and readable of the many books on Erasmus; OTTO BENESCH, *The Art of the Renaissance in Northern Europe,* rev. ed. (1965), a major study by a leading scholar; KARL BRANDI, *Kaiser Karl V, 2* vol. (1937–41; Eng. trans., *The Emperor Charles V,* 1939), a superb biography of a pivotal 16th-century figure; J.M. CLARK, *The Great German Mystics: Eckhart, Tauler and Suso* (1949), a good introduction to the three founders of the German school; J.H. ELLIOTT, *Imperial Spain, 1469–1716* (1963), the best short account in English; JOHAN HUIZINGA, *Herfsttij der middeleeuwen* (1919; Eng. trans., *The Waning of the Middle Ages,* 1924, reprinted 1955), a classic—France and the Low Countries on the eve of their Renaissance; H.A. MISKIMIN, *The Economy of Early Renaissance Europe, 1300–1460* (1969), a brief, up-to-date survey; E. ALLISON PEERS, *Studies of the Spanish Mystics,* 3 vol. (1927–60), the classic work in the field; GUSTAVE REESE, *Music in the Renaissance* (1954), an excellent, comprehensive introduction to the subject.

Renaissance science and technology: PIERCE BUTLER, *The Origin of Printing in Europe* (1940), a standard work on the subject; HERBERT BUTTERFIELD, *The Origins of Modern Science, 1300–1800* (1949), the best non-technical introduction; CARLO M. CIPOLLA, *Guns, Sails and Empires: Technological Innovation and the Early Phases of European Expansion, 1400–1700* (1966), explores the technological origins of Western supremacy; J.H. PARRY, *The Age of Reconaissance,* 2nd ed. (1966), full of information on navigation, shipbuilding, and exploration.

(D.We.)

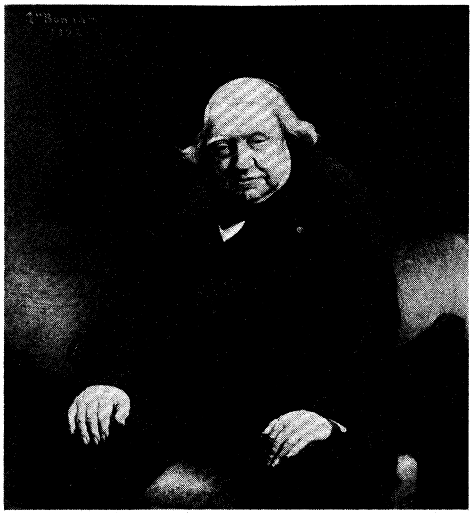
Renan, Ernest

Ernest Renan, a significant 19th-century philosopher, historian, and scholar of religion, anticipated in his life and writings the anxiety of the existential quest for authentic existence and the totalitarian messianism of the 20th century. Born of peasant stock on February 28, 1823, at Tréguier, Côtes-du-Nord, France, Renan was educated at the ecclesiastical college in his hometown. He began training for the priesthood, and in 1838 he was offered a scholarship at the seminary of Saint-Nicolas-du-Chardonnet. He later went on to the seminary of Saint-Sulpice, where he soon became involved in a crisis of faith that finally led him, reluctantly, to leave the Roman Catholic Church in 1845.

Ecclesiastical training

In his view, the church's teachings were incompatible with the findings of historical criticism; but he kept a quasi-Christian faith in the hidden God, revealed to him both in the 17th-century religious writings of Blaise Pascal and in experience, as well as in himself as a kind of intellectual messiah.

Early works. For Renan, the February revolution of 1848 in France and other parts of Europe was a religion in the making. Sometimes enthusiastic, sometimes critical, he participated in the revolution's messianic expectations and carried this ambiguous attitude over into *L'Avenir de la science (The Future of Science,* 1891)—a work that, in allusion to the works of the philosopher René Descartes, he called his "Discourse on Method." The main theme of this work, not published until 1890, is the importance of the history of religious origins, which he regarded as a human science having equal value to the sciences of nature. Though he was now somewhat anticlerical, the French government sent him in 1849 to Italy, where the papacy was still politically important, to help classify manuscripts previously inaccessible to French scholars. In Rome he was transformed, for a while, from



Renan, oil painting by Léon Bonnat, 1892. In the Musée Renan, Tréguier.
Archives Photographiques

an austere scholar into an artist alive to the pageantry and naïveté of popular religion.

Renan returned to Paris in 1850 to live with his sister, Henriette, on her savings and the small salary attached to his own post at the Bibliothèque Nationale. He began to make a name for himself with his doctoral thesis ***Averrobs et l'Averroïsme*** (1852; "Averroës and Averroism"), concerning the thought of that medieval Muslim philosopher. He continued his scholarly writings with two collections of essays, ***Études d'histoire religieuse*** (1857; ***Studies of Religious History,*** 1864) and ***Essais de morale et de critique*** (1859; "Moral and Critical Essays"), first written for the ***Revue des Deux Mondes*** and the ***Journal des Dkbats.*** The ***Etudes*** inculcated into a middle class public the insight and sensitivity of the historical, humanistic approach to religion. Many of the ***Essais*** denounce the materialism and intolerance of the Second Empire (1852–70) in the name of Renan's aristocratic ideal: the intellectuals, acting as "bastions of the spirit," must, he affirms, resist tyranny by intellectual and spiritual refinement; that is, by the humanism exemplified throughout these essays.

In 1856 Renan proposed marriage to Cornélie Scheffer, niece of the painter Ary Scheffer. When Henriette discovered that she herself meant less to her brother than his work, even less than another woman, her jealousy knew no bounds. As related by Renan with more art than truth in his tribute to her (1862), Henriette gave way to his benevolent egoism and agreed to live with him and his wife. Yet, the beginnings of his marriage were unhappy, the more so as he was still dependent on Henriette's savings.

Investigations concerning the life of Jesus

In October 1860 Renan was entrusted with an archaeological mission to the Lebanon. The Phoenician inscriptions that he discovered were published in his ***Mission de Phénicie*** (1864–74; "Phoenician Expedition"). They were later included in the ***Corpus Inscriptionum Semiticarum*** ("Corpus of Semitic Inscriptions"), which he helped to bring out through the Académie des Inscriptions et Belles-Lettres. But archaeology was not his main interest. In April 1861, with his wife and sister, he visited the Holy Land in search of materials and inspiration concerning a life of Jesus that he was bent on writing. He finished a first draft of it in the Lebanon but at tragic cost, for Henriette died of malaria at ʿAmshīt on September 24, 1861, while he himself fell desperately ill.

Religious controversies. Renan had counted on the writing of his life of Jesus to secure election to the chair of Hebrew at the College de France. He was elected, before the book was ready, on January 11, 1862. But in his opening lecture, on February 21, he referred to Jesus in the words of Jacques Bossuet, a French bishop and historian of the 17th and 18th centuries, as "an incomparable man." Though this was, in his eyes, the highest praise one could bestow on a man, it was not sufficient for the clericals, who took advantage of its implied atheism and the uproar caused by the lecture to have Renan suspended. Contemptuously refusing an appointment to the Bibliothèque Imperiale (June 1864), Renan decided to live by his pen for the next few years. He had to wait until 1870, however, before the chair was restored to him. He was thus pushed into opposition to the church but had already begun to frequent such dissident salons as that of Princess Mathilde, niece of Napoleon Bonaparte, and to associate with such literary notables as Gustave Flaubert, Charles Augustin Sainte-Beuve, Hyyyolyte-Adolyhe Taine, and the Goncourt brothers (Edmond and Jules).

When the ***Vie de Jésus (Life of Jesus,*** 1869) did appear in 1863 it was virulently denounced by the church. Though not Renan's best historical work, it can still claim the attention of 20th-century readers because it presents a "mythical" account of the making of Christianity by the popular imagination, and thus has a place, like his other historical works, in the literature of messianism. After a journey in Asia Minor in 1864–65 with his wife, he published ***Les Apôtres*** (1866; ***The Apostles,*** 1869) and ***Saint Paul*** (1869), to follow the ***Vie de Jésus*** as parts of a series, ***Histoire des origines du christianisme (The History of the Origins of Christianity).*** Both these volumes, containing brilliant descriptions of how Christianity spread among the rootless proletariat of the cities of Asia Minor, illustrate his preoccupation with the question: would the intellectuals of the 19th century lead the masses toward a new enlightenment?

Interest in politics. Renan began to interest himself increasingly in politics. In 1869, at the beginning of the "liberal" phase of the Second Empire, he stood unsuccessfully for Parliament. In the same year he defended constitutional monarchv in an article, "La Monarchie constitutionnelle en Fiance" ("Constitutional Monarchy in France"). Thus far he was a liberal. In the same spirit he tried, during the Franco-German War of 1870–71, to work across frontiers: he corresponded with David Friedrich Strauss, a German theologian, and tried to persuade the Prussian crown prince (later German emperor as Frederick III) to stop the war. But the bitterness of France's defeat and his anger with democracy caused him to become authoritarian. Thus, ***La Réforme intellectuelle et morale*** (1871), concerning intellectual and moral reform, argues that France, to achieve national regeneration, must follow the example set by Prussia after the Battle of Jena in 1806. By taking his advice, however, France would have become the sort of clerical monarchy that Renan soon found he did not want. He had to resign himself to accepting the Third Republic (1870–1940), but he withdrew from public life. Though he continued to travel zestfully all over Europe, visiting surviving Bonapartists, such as Prince Jérôme Napoléon, his life became more and more identified with his writings. He was elected to the Acadkmie Française in 1878.

Concerns for **apocalypticism** and death

Later writings. Renan's ironical yet imaginative vision of the "festival of the universe" found expression in ***L'Antéchrist*** (1873; ***The Antichrist,*** 1896; vol. iv of the ***Histoire des origines),*** with its satirical portrait of Nero and its apocalyptic atmosphere—replete with expectations of a cataclysmio consummation of history—assuredly the most impressive of his historical narratives. The "festival of the universe" provides a visionary end to the ***Dialogues et fragments philosophiques*** (1876; ***Philosophical Dialogues and Fragments,*** 1899). In the first of these, however, Renan is more ironically skeptical about the hidden God than he had been. In fact, the Epicureanism of his later years masks an anxiety about death and the hereafter. His more superficial side is illustrated in the "philosophic dramas" (collected edition 1888), which trace his acceptance of the Republic, especially ***Caliban*** (written 1877) and ***L'Eau de jouvence*** (written 1879; "The Water of Youth"). In the former, the aristocracy (Prospero and Ariel) loses to democracy (Caliban) because alchemical spells (traditional sanctions) are powerless against a people infected by positivism; scientific power politics would be an effective answer, but this is out of the question because in practice it would mean a clerical monarchy.

As to the remaining volumes of the ***Histoire des origines,*** if Renan's Epicureanism is hard to find in ***Les Évangiles*** (1877; ***The Gospels,* 1889**), it is present in *L'Église chrétienne* (1879; "The Christian Church") in the portrait of the Roman emperor Hadrian; but in ***Marc-Aurèle*** (1882; ***Marcus Aurelius,* 1904**), the study of Marcus Aurelius (a Roman emperor and philosopher), again a self-portrait, is dominated by the author's preoccupation with death. Since 1876 Renan had been working on his memoirs, ***Souvenirs d'enfance et de jeunesse*** (1883; ***Recollections of My Youth,* 1883**), in which he reconstructs his life so as to show that he was predestined to become a *prêtre manqué* (failed priest) and that, in spite of heavy odds, his wager on the hidden God had "paid off' in terms of happiness.

In the ***Souvenirs*** Renan is too serene for some tastes, though his irony keeps his complacency in check. In *L'Ecclésiaste* (1882; "Ecclesiastes") and two articles on Amiel (**1884**), he is above all an ironist combatting the Pharisees (religious legalists). On the other hand, in some of his speeches at the Académie Française, on Claude Bernard, a French physiologist (**1879**), and Emile Littré, a French philologist (**1882**), he reveals his anguish in moments of doubt. Thus, he manifests a baffling variety of characteristics, but the moral heart of the man is to be found in one of the later dramas, ***Le Prêtre de Némi*** (1885; "The Priest of Némi"), and above all in his ***Histoire du peuple d'Israël*** (1887–93; "History of the People of Israel," 1888–91). For him, the history of Jewish messianism bore witness to man's capacity for faith when the odds are against him. Thus, it revived his own faith. He could therefore hope that, though Judaism would disappear, the dreams of its prophets would one day come true, so that "without a compensatory Heaven justice will really exist on earth." Having exhausted himself in an effort to finish the work, he died shortly after its completion on October 2, 1892.

With his leanings toward liberalism and authoritarianism in politics and faith and skepticism in religion, Renan embodied the contradictions of the middle class of his time. Politically, his influence after his death was far-reaching, on nationalists, such as Maurice Barrès and Charles Maurras, on republicans, such as Anatole France and Georges Clemenceau. He succeeded in assuaging one of the great anxieties of his time, the antagonism between science and religion, but he very much felt this anxiety.

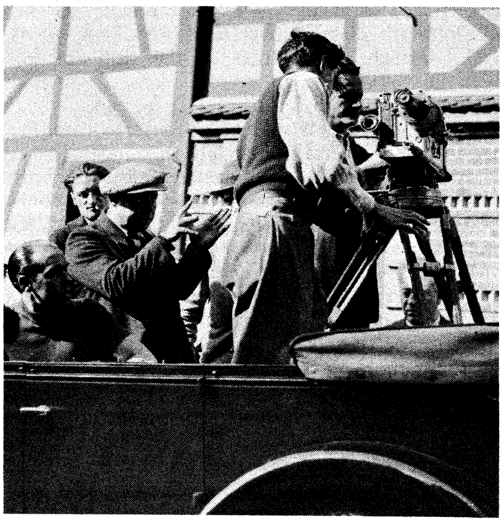
Henriette Psichari's definitive edition of Renan's works appeared in ten volumes in 1947–61.

BIBLIOGRAPHY. J.M. POMMIER, ***Renan, d'après des documents inédits*** (**1923**), may be regarded as the definitive hiography; the same author's ***La Jeunesse cle'ricale d'Ernesi Renan*** (**1933**), is indispensable for a detailed knowledge of Renan's seminary years. HENRIETTE PSICHARI, ***Renan d'après lui-même*** (1937) and ***Renan et la guerre de '70*** (**1947**), constitute an important appraisal of Renan by one of his granddaughters. R. DUSSAUD, ***L'Oeuvre scientifique &Ernest Renan*** (1951), is the most exhaustive assessment of Renan's scholarly work. R.M. CHADBOURNE, ***Ernest Renan As an Essayist*** (1957), breaks new ground in Renan criticism, and the same author's ***Ernest Renan*** (**1968**), condenses ***an*** impressive amount of analysis. H.W. WARDMAN, ***Ernest Renan*** (1964), is a critical biography that stresses Renan's interest in the psychology and politics of messianism. HENRI PEYRE (ed.), ***Sagesse de Renan*** (**1968**), is a discriminating anthology by a leading scholar. K. GORE, ***L'Idée de progrès dans la pensée de Renan*** (1970), is a sound scholarly study of his political philosophy.

(H.W.W.)

Renoir, Jean

Born virtually at the same time as motion pictures were invented, Jean Renoir was to achieve a place in the forefront of the creators of the art of the film. His fame is based largely on the deeply moving films he made in France in the 1930s, such as ***La Règle du jeu*** and ***La Grande Illusion,*** but he also wrote and directed films—in France, in Hollywood, and elsewhere—that are outstanding for their entertainment values alone. All his works are distinguished for their strong pictorial sense, and his best are of a quality equalling the most sublime art in their humanity, grace, and style.



Jean Renoir (centre) directing *Madame* Bovary in 1933.
H. Roger-Viollet

Artistic influences

The son of the Impressionist painter Auguste Renoir and the younger brother of the actor Pierre Renoir, Jean Renoir was born in the Montmartre section of Paris on September 15, 1894. In an environment in which art predominated, among painters and their models, he spent a happy childhood, which was richer in the carefree appreciation of beauty than in formal studies. Nevertheless, he received a degree in 1913 from the University of Aix-en-Provence, where he wrote poetry, and joined the cavalry to begin a military career.

World War I broke out in 1914, and Renoir was wounded in the leg. During his convalescence, he spent his time in Paris movie houses, where he discovered the serials and Charlie Chaplin. After he recovered, he rejoined the service in the air force and finished the war with the rank of lieutenant.

Undecided on a career, he studied ceramics with his brother at Cagnes-sur-mer, near Nice, where his family had settled. Early in 1920, he married one of his father's models, Andrée Heurschling, a few months after the painter's death, and went with her to live in the Paris area in Marlotte, a village in which his father had once painted.

Intending to set up a ceramics factory, Jean Renoir was joined by his friend Paul Cézanne, the son of the painter. Having come into contact with theatrical circles through his sister-in-law, the actress Vera Sergine, Renoir was attracted by the evolving art of the film and decided to write a screenplay. It was made into the the film ***Catherine,*** or ***Une vie sans joie,*** in 1923, with his wife appearing under the name of Catherine Hessling. Another film was made on Cézanne's property, ***La Fille de l'eau*** (released **1924**), which again starred Renoir's wife. All of his early films were produced in a makeshift way, with a technical clumsiness, a lack of means, and an amateurishness. Nevertheless, the instinctive genius of the film maker found expression in them. These early films, which reveal a strong pictorial influence, have taken on with time a particular charm. In the late 1920s, he found his inspiration in the writings of Emile Zola, Hans Christian Andersen, and others, but made of them personal films in the style of the French avant-garde of the period.

These films had no commercial success, and Renoir and his backers were almost ruined. The advent of sound in motion pictures brought new difficulties, but Renoir passed the test with ***On purge bkbk*** (1931) and proved himself with ***La Chienne*** (**1931**), a fierce and bitter film adapted from a comic novel by Georges de la Fouchardière.

Freedom of composition

During the 1930s, Jean Renoir produced many of his most notable works, but their freedom of composition was confusing to critics of the period, and the films achieved only middling success. These films include ***La Nuit du carrefour*** (**1932**), based on a novel by Georges Simenon; ***Boudu sauvk des eaux*** (**1932**), an anarchistic and unconstrained comedy; ***Madame Bovary*** (1934), based on Flaubert's classic novel; and ***Le Crime de M. Lange*** (1936), which, in contrast to the rather stilted manner of the first years of sound films, foretells a reconquest of the true moving-picture style, especially in use of improvisation and of montage—the art of editing, or cutting, to achieve certain associations of ideas.

In 1936, in sympathy with the social movements of the French Popular Front, Renoir directed the Communist propaganda film La ***Vie est à nous***, but the film was not commercially successful. The same year he recaptured the flavour of his early works with a short film, ***Une partie de campagne*** (released 1946), which he finished with great difficulty. A masterpiece of Impressionist cinema, this film presents all the poetry and all the charm of the pictorial sense that is, far more than is his technique, the basis of Renoir's art as a film maker. The late 1930s saw such major works as ***La Grande Illusion*** (1937), a moving story of World War I prisoners of war; La ***Bête humaine*** (1938), an admirable free interpretation of Zola; and especially La ***Règle du jeu*** (1939), his masterpiece. Cut and fragmented by the distributors, this classic film was also regarded as a failure until it was shown in 1965 in its original form, which revealed its astonishing beauty.

During World War II, at the time of the Nazi invasion of France in 1940, Jean Renoir, like many of his friends, went to Hollywood. His American period includes films of varying merit, which mark a departure from his previous style: ***Swamp Water*** (1941), ***The Southerner*** (1945), ***Diary of a Chambermaid*** (1946), and ***The Woman on the Beach*** (1947). In 1944, after being divorced from Catherine Hessling, he married Dido Freire, the daughter of the Brazilian film maker Alberto Cavalcanti. He made ***The River*** (1951) in India, his first colour film.

Mature style

Now in full command of a mature style that reflected the qualities of the man himself—sensitivity, fervour, and humanity—he returned to Europe by way of Italy, where he made ***Le Carrosse d'or*** (released 1953). A sumptuous work, combining the talents of both a painter and a dramatist, this film shows Renoir's love of actors and their profession. He often played roles in his own films, and he allowed his actors a great deal of initiative. Subsequently, he made ***French-Cancan*** (1955), a fabulous evocation of the Montmartre of the 18th century, and ***Eléna et les hommes*** (1956), a period fantasy swept along in a prodigious movement. His last works, of the 1960s, do not achieve the same beauty, nor does his television work.

A powerful personality, having been deeply impressed by the artistic environment of his youth, Renoir was also extremely open to later influences both in his art and in his ideas. A naturalized American citizen, settling in Los Angeles, he nevertheless kept his French nationality and maintained connections in Paris.

In addition to his films, Renoir also wrote a play, *Orvet* (first performed 1955), which was presented in Paris; a novel, ***Les Cahiers du capitaine Georges*** (1966); and an invaluable book of memories about his father, ***Renoir*** (1962). His autobiography, ***Ma Vie et mes films***, was published in 1974 (English translation, ***My Life and My Films***, also 1974). He died at Los Angeles on February 12, 1979.

MAJOR WORKS

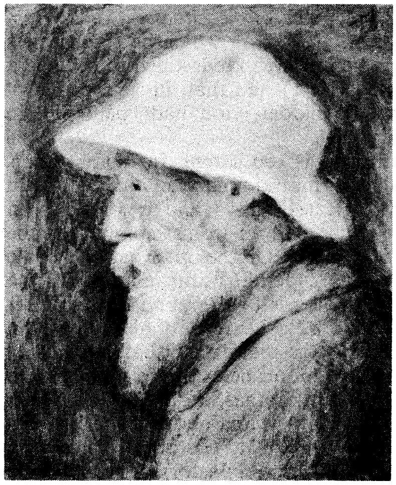
La Chienne (1931); Boudu sauvé des eaux (1932; Boudu Saved from Drowning); La Nuit du carrefour (1932); Toni (1934); Le Crime de M. Lange (1936; The Crime of Monsieur Lange); Une Partie de campagne (made in 1936, released in 1946; A Day in the Country); La Grande Illusion (1937; Grand Illusion); La Bête humaine (1938); La Règle du jeu (1939; The Rules of the Game); The River (1951); Le Carrosse d'or (1953; The Golden Coach); French-Cancan (1955).

(P.Le.)

Renoir, Pierre-Auguste

During the last quarter of the 19th century, Impressionism had the greatest effect on the methods and goals of painting, and to that Impressionism Pierre-Auguste Renoir brought an accent of pleasing sensuality that gave his work its particular force. His background (he belonged to a family of simple artisans), his subjects (much more than his contemporaries, he was drawn to the human figure), and lastly his technique (he did not use for long the divided colour and juxtaposition of small, separate strokes that characterized the other Impressionists) all set him apart from his fellow painters and at the same time helped make him one of the most brilliant members of that school.

Photograph Durand-Ruel permission S P A D E M 1971, by French Reproduction Rights, Inc



Renoir, self-portrait, oil painting, 1910. In the Durand-Ruel Collection, Paris.

Early work

Renoir was born at Limoges on February 25, 1841, into a family of small artisans. His father, a tailor who had seven children, moved with his family to Paris about 1845. Renoir demonstrated his talent for painting at a very early age. Quickly recognizing his talents, his parents apprenticed him, at the age of 13, to work in a porcelain factory, where he learned to decorate plates with bouquets of flowers. Shortly after that, he was painting fans and then cloth panels representing religious themes for missionaries to hang in their churches. His skill and the great pleasure he took in his work soon persuaded him to study painting in earnest. Having saved a little money, he decided, in 1862, to take evening courses in drawing and anatomy at the École des Beaux-Arts as well as painting lessons at the studio of Charles Gleyre, a Swiss painter who had been a student of the 19th-century Neoclassical painter J.-A.-D. Ingres. Although the academic style of his teacher did not suit Renoir, he nevertheless accepted its discipline in order to acquire the elementary knowledge needed to become a painter. He felt a much greater affinity with three students who entered the studio a few months later: Alfred Sisley, Claude Monet, and Frederic Bazille. All four students dreamed of an art that was closer to life and free from past traditions. The shared ideals of the four young men quickly led to a strong friendship. At the same time in another workshop at the Académie Suisse, the young artists Paul Cezanne and Camille Pissarro were preoccupied with the same problems. With Bazille as the intermediary, the two groups met frequently.

Fontainebleau period

Association with the Impressionists. Circumstances encouraged Renoir to attempt a new freedom and experimentation in his style. The tradition of the time was that a painting—even a landscape—had to be executed in the studio. In the spring of 1864, however, Gleyre's four students moved temporarily to the forest of Fontainebleau where they devoted themselves to painting directly from nature. The Fontainebleau forest had earlier attracted other artists, among them Theodore Rousseau and Jean-François Millet, who demanded that art represent the reality of everyday life, even though they had not yet completely renounced the constraints imposed by the traditional school. In 1863 Édouard Manet took a much bolder step: his picture "Le Déjeuner sur l'herbe" ("Luncheon on the Grass"; Louvre, Paris) provoked a resounding scandal because its subject and technique affirmed the need for a revival of painting through the observation of reality. Manet's daring made him the leader of a new movement in the eyes of these young artists.

Conditions were ripe for the birth of a new pictorial language; and Impressionism, bursting upon the scene, made quite a scandal in the first Impressionist exposition of 1874, held independently of the official Salon exhibition. It took ten years for the movement to acquire its definitive form, its independent vision, and its unique perceptiveness. But one can point to 1874 as the year of departure for the movement that subsequently spawned modern art.

Renoir's work is a perfect illustration of this new approach in thought and in technique. Better than any other artist, he suggested by small multicoloured strokes the vibration of the atmosphere, the sparkling effect of foliage, and especially the luminosity of a young woman's skin in the outdoors. Renoir and his companions stubbornly strove to produce light-coloured paintings from which black was excluded, but their pursuits led to many disappointments: their paintings, so divergent from traditional formulas, were constantly rejected by the juries of the Salon and were extremely difficult to sell. On the other hand, despite the continuing criticisms, some of the Impressionists were making themselves known, as much among art critics as among the lay public. Renoir, because of his interest in the human figure, separated himself from the others who were more tempted by landscape. Thus he obtained several orders for portraits and was introduced, thanks to the publisher Georges Charpentier, to an upper middle-class society, whose women and children he painted.

Renoir was now a master of his craft, and his paintings showed great vitality despite the grave financial worries that troubled him. Several of his masterpieces date from this period: "La Loge," "Le Moulin de la galette," "The Luncheon of the Boating Party," "Mme Charpentier and Her Children." Charpentier organized a personal exposition for the works of Renoir in 1879 in the gallery La Vie Moderne.

Rejection of Impressionism. In 1881 and 1882 Renoir made several trips to Algeria, Italy, and Provence, which eventually had a considerable effect on his art and on his life. He became convinced that the systematic use of the Impressionistic technique was no longer sufficient for him and that small brush strokes of contrasting colours placed side by side did not allow him to convey the satiny effects of the skin. He also discovered that black did not deserve the opprobrium given to it by his comrades and that, in certain cases, it had a striking effect and gave a great intensity to the other colours. During his journey to Italy, he discovered Raphael and the fascinations of classicism: the beauty of drawing, the purity of a clear line to define a form, and the expressive force of smooth painting to enhance the suppleness and modelling of a body. At this same time, he happened to read *Il libro dell'arte* (1437; *A Treatise on Painting*, 1844) by Cennino Cennini, which confirmed his new ideas. What he learned from all of these revelations was so powerful, brutal, and unexpected that it provoked a crisis, and he was tempted to break with Impressionism, which he had already begun to doubt. He felt that until now he had been mistaken.

The Ingres period

Most of his works executed from 1883–84 on are so marked by a new discipline that art historians have grouped them under the title "the Ingres period" to signify their vague similarity with the technique of Ingres. Renoir's experiments with Impressionism were not wasted, however, because he retained a palette that was bursting with colours.

His strong reaction against Impressionism continued until around 1890. During these years he made several trips to southern France: Aix-en-Provence, Marseille, and Martigues. The nature of this sunlit region gave greater encouragement to his separation from Impressionism, which to him was associated with the landscapes of the valley of the Seine. Southern France offered him scenes bursting with colour and sensuality. At the same time, the seemingly joyous spontaneity of nature gave him the desire to depart from the very strict rules of classicism. While in southern France, he recovered the instinctive freshness of his art; he painted women at their bath with the same healthful bloom he would give to bouquets of flowers.

Marriage and fame

His financial situation was appreciably improved; he was married in 1890 to Alice Charigot, and the exposition that was organized for him in 1892 by Durand-Ruel was a great success. Renoir's future was assured, and his work of that period reflected his new security and also his confidence in the future.

Later years. Two years later, he had his first attack of rheumatism, and, as the attacks became more and more frequent, he spent more and more time in southern France where the climate was better for his health. About 1899 he sought refuge in the small village of Cagnes; in 1907 he settled there permanently, buying the estate of Les Collettes, where he spent the rest of his life. In 1910 he was no longer able to walk. But in spite of his infirmity, which was more and more constraining, Renoir never ceased to paint; when his fingers were no longer supple, he continued by attaching his paintbrush to his hand.

In spite of his misfortune, his paintings still embodied a cheerful attitude toward life. He was no longer satisfied with topical themes or with smiling portraits of the Parisian bourgeoisie but turned instead to portraits of his wife, his children, and of Gabrielle his maid, who often also posed for his nude paintings. His still lifes were composed of flowers and fruits from his own garden, and the landscapes were those that surrounded him. The nudes, especially, reflect the serenity that he found in the joy of working throughout his life. He attempted to embody his admiration for the female form in sculpture, with the assistance of young Richard Guino. Since Renoir was no longer able to do sculpture himself, Guino became, about 1913, the skillful instrument who willingly followed his directions. He yielded before the personality of Renoir and succeeded so well in this difficult enterprise that the works born of this union have all the qualities of Renoir's style.

Mme Renoir died in 1915, after having returned from Gérardmer, where she had gone to see their son Jean, who had been seriously wounded in the war. Renoir survived her for four years, until his death on December 3, 1919, at Cagnes. Several months earlier—in August—he had been able to go to Paris to see his "Portrait de Mme Georges Charpentier," which had been recently acquired by the state. On that occasion, several friends had wheeled him for the last time to view the Louvre's masterpieces that he had venerated throughout his life.

MAJOR WORKS

"Le Carabet de la mère Anthony" (1866; Nationalmuseum, Stockholm); "Diana" (1867; National Gallery of Art, Washington, D.C.); "Portrait du peintre Bazille" (1867; Louvre, Paris); "Lise" (1867; Museum Folkwang, Essen); "The Painter Sisley and His Wife" (1868; Wallraf-Richartz-Museum, Cologne); "La Grenouillère" (1869; Nationalmuseum, Stockholm); "Chalands sur la Seine" ("Barges on the Seine," *c.* 1869; Louvre); "Odalisque" (1870; National Gallery of Art, Washington, D.C.); "Captain Darras" (1871; Gemäldegalerie, Dresden); "The Breakfast" (1872; Barnes Foundation, Merion, Pennsylvania); "Monet Painting in His Garden" (1873; Wadsworth Atheneum, Hartford, Connecticut); "Horsewoman in the Bois de Boulogne" (1873; Hamburger Kunsthalle); "Dancer" (1874; National Gallery of Art, Washington, D.C.); "La Loge" ("The Theatre Box," 1874; Courtauld Institute Galleries, London); "Mme Chocquet" (1875; Staatsgalerie, Stuttgart); "La Premiere Sortie" (1875–76; Tate Gallery, London); "Le Moulin de la galette" (1876; Louvre); "Mme Charpentier and Her Children" (1878; Metropolitan Museum of Art, New York); "Oarsmen at Chatou" (1879; National Gallery of Art, Washington, D.C.); "Little Blue Nude" ("Petit Nu Bleu," 1879–80; Albright-Knox Art Gallery, Buffalo, New York); "The Luncheon of the Boating Party" ("Le Déjeuner des canotiers," 1881; Phillips Collection, Washington, D.C.); "Bather" (1881: Sterling and Francine Clark Art Institute, Williamstown, Massachusetts); "Le Bal à Bougival" ("The Dance at Bougival," 1883; Museum of Fine Arts, Boston); "Les Parapluies" (c. 1883; National Gallery, London); "Bathers" (1884–87; Mrs. C.S. Tyson Collection, Philadelphia); "Mount Sainte-Victoire" (1889; Barnes Foundation, Merion, Pennsylvania); "Bather" (1892; Robert Lehman Collection, New York);

"Young Girl Reading" (1892; Durand-Ruel Collection, Paris); "Berthe Morisot and Her Daughter" (1894; Mme Ernest Rouart Collection, Paris); "After the Bath (*c.* 1895; E.C. Vogel Collection, New York); "The Artist's Family" (1896; Barnes Foundation, Merion, Pennsylvania); "Sleeping Bather" (1897; Sammlung Oskar Reinhart, Winterthur, Switzerland); "Le Lever" (1899; Barnes Foundation, Merion, Pennsylvania).

BIBLIOGRAPHY. The principal work on Renoir is the *Catalogue raisonné de l'oeuvre peint*, ed. by FRANCOIS DAULTE, vol. 1, *Figures, 1860–1890* (1971), the first volume of a projected four-volume work. Other important studies are: JULIUS MEIER-GRAEFE, *Auguste Renoir* (1911) and *Renoir* (1929). in German; ALBERT C. BARNES and VIOLETTE DE MAZIA, *The Art of Rerzoir* (1935); MICHEL DRUCKER, *Renoir* (1944), the most complete critical and biographical study in French; and HENRI PERRUCHOT, *La Vie de Renoir* (1964), a good and skillful use of the best documented sources. Among the numerous monographs, the principal ones are: AMBROISE VOLLARD, *La Vie et I'oeuvre de Pierre-Auguste Renoir* (1919; Eng. trans., 1925); ALBERT ANDRE, *Renoir,* new ed. (1923); THEODORE DURET, *Renoir* (1924; Eng. trans., 1937); GUSTAVE COQUIOT, *Renoir* (1925); MICHEL FLORISOONE, *Renoir* (1937); CLAUDE ROGER-MARX, *Renoir* (1937); and FRANCOIS FOSCA, *Renoir, I'homme et son oeuvre* (1961; Eng. trans., 1970). Works on more specialized aspects include: ALBERT ANDRE and MARC ELDER, *L'Atelier de Renoir* (1931); PAUL HAESAERTS, *Renoir, sculpteur* (1947; Eng. trans., 1947); CLAUDE RENOIR, "Souvenirs sur mou père," in AUGUSTE RENOIR, *Seize aquarelles et sanguines* (1948); JEAN RENOIR, *Renoir,* My *Father* (Eng. trans. 1962); CLAUDE ROGER-MARX, *Les Lithographies de Renoir* (1951); and JOHN REWALD (ed.), *Renoir Drawings* (1946, reprinted 1958). Recent studies in English include PARKER TYLER, *Renoir* (1969); and LAWRENCE HANSON, *Renoir: The Man, the Painter, and His World* (1968).

(R.Cog.)

Reproduction

In a general sense reproduction is one of the most important concepts in biology: it means making a copy, a likeness. Although reproduction is often considered solely in terms of the production of offspring in animals and plants, the more general meaning has far greater significance to living organisms. To appreciate this fact, the origin of life and the evolution of organisms must be considered. One of the first characteristics of life that emerged in primeval times must have been the ability of some primitive chemical system to make copies of itself. At its lowest level, therefore, reproduction is chemical replication. As evolution progressed, cells of successively higher levels of complexity must have arisen, and it was absolutely essential that they had the ability to make likenesses of themselves. In unicellular organisms, the ability of one cell to reproduce itself (along with its parts) means the reproduction of a new individual; in multicellular organisms, however, it means growth (the enlargement of the organism by cell multiplication) and regeneration (the replacement of lost parts). Multicellular organisms also reproduce in the strict sense of the term—that is, they make copies of themselves in the form of offspring.

LEVELS OF REPRODUCTION

Transmittal of genetic information

Molecular **replication.** The characteristics that an organism inherits are largely stored in cells as genetic information in very long molecules of deoxyribonucleic acid (DNA). In 1953, it was established that DNA molecules actually consist of two complementary strands, each of which can make copies of the other. The strands are like two sides of a ladder that has been twisted along its length in the shape of a double helix (spring). The lungs, which join the two sides of the ladder, are made up of two terminal bases. There are four bases in DNA: thymine, cytosine, adenine, and guanine. In the middle of each rung a base from one strand of DNA is linked by a hydrogen bond to a base of the other strand. But they can only pair in certain ways: adenine always pairs with thymine, and guanine with cytosine. This is why one strand of DNA is considered complementary to the other.

The double helices duplicate themselves by separating at one place between the two strands and becoming progressively unattached. As one strand separates from the other, each acquires new complementary bases until eventually each strand becomes a new double helix with a new complementary strand to replace the original one. Because adenine always falls in place opposite thymine and guanine opposite cytosine, the process is called a template replication—one strand serves as the mold for the other. It should be added that the steps involving the duplication of DNA do not occur spontaneously; they require catalysts in the form of enzymes that promote the replication process.

Genetic code

Molecular reproduction. The sequence of bases in a DNA molecule serves as a code by which genetic information is stored. Using this code, the DNA synthesizes one strand of ribonucleic acid (RNA), a substance that is so similar structurally to DNA that it is also formed by template replication of DNA. RNA serves as a messenger for carrying the genetic code to those places in the cell where proteins are manufactured. The way in which the messenger RNA is translated into specific proteins is a remarkable and complex process. (For more detailed information concerning DNA, RNA, and the genetic code, see the articles GENE; NUCLEIC ACID.)

The ability to synthesize enzymes and other proteins enables the organism to make any substance that existed in a previous generation. Proteins are reproduced directly; however, such other substances as carbohydrates, fats, and other organic molecules found in cells are produced by a series of enzyme-controlled chemical reactions, each enzyme being derived originally from DNA through messenger RNA. It is because all of the organic constituents made by organisms are derived ultimately from DNA that molecules in organisms are reproduced exactly by each successive generation.

Cell reproduction. The chemical constituents of cytoplasm (that part of the cell outside the nucleus) are not resynthesized from DNA every time a cell divides. This is because each of the two daughter cells formed during cell division usually inherits about half of the cellular material from the mother cell (see CELL AND CELL DIVISION), and is important because the presence of essential enzymes enables DNA to replicate even before it has made the enzymes necessary to do so.

Cells of higher organisms contain complex structures, and each time a cell divides the structures must be duplicated. The method of duplication varies for each structure, and in some cases the mechanism is still uncertain. One striking and important phenomenon is the formation of a new membrane. Cell membranes, although they are very thin and appear to have a simple form and structure, contain many enzymes and are sites of great metabolic activity. This applies not only to the membrane that surrounds the cell but to all the membranes within the cell. New membranes, which seem to form rapidly, are indistinguishable from old ones.

Thus, the formation of a new cell involves the further synthesis of many constituents that were present in the parent cell. This means that all of the information and materials necessary for a cell to reproduce itself must be supplied by the cellular constituents and the DNA inherited from the parent cell.

Binary fission. Of the various kinds of cell division, the most common mode is binary fission, the division of a cell into two separate and similar parts. In bacteria (prokaryotes) the chromosome (the body that contains the DNA and associated proteins) replicates and then divides in two, after which a cell wall forms across the elongated parent cell. In higher organisms (eukaryotes) there is first an elaborate duplication and then a separation of the chromosomes (mitosis), after which the cytoplasm divides in two. In the hard-walled cells of higher plants, a median plate forms and divides the mother cell into two compartments; in animal cells, which do not have a hard wall, a delicate membrane pinches the cell in two, much like the separation of two liquid drops. Budding yeast cells provide an interesting exception. In these fungi the cell wall forms a bubble that becomes engorged with cytoplasm until it is ultimately the size of the original cell. The nucleus then divides, one of the daughter nuclei passes into the bud, and ultimately the two cells separate.

Mitosis and meiosis

In some instances of binary fission, there may be an unequal cytoplasmic division with an equal division of the chromosomes. This occurs, in fact, in a large number of higher organisms during meiosis—the process by which sex cells (gametes) are formed: originally each chromosome of the cell is in a pair (diploid); during meiosis these diploid pairs of chromosomes are separated so that each sex cell has only one of each pair of chromosomes (haploid). During the two successive meiotic divisions involved in the production of eggs, a primordial diploid egg cell is converted into a haploid egg and three small haploid polar bodies (minute cells). In this instance the egg receives far more cytoplasm than the polar bodies.

Multiple fission. Some algae, some protozoans, and the true slime molds (Myxomycetes) regularly divide by multiple fission. In such cases the nucleus undergoes several mitotic divisions, producing a number of nuclei. After the nuclear divisions are complete, the cytoplasm separates, and each nucleus becomes encased in its own membrane to form an individual cell. In the Myxomycetes, the fusion of two haploid gametes or the fusion of two or more diploid zygotes (the structures that result from the union of two sex cells) results in the formation of a plasmodium—a motile, multinucleate mass of cytoplasm. The nuclei are in a syncytium, that is, there are no cell boundaries, and the nuclei flow freely in the motile plasmodium. As it feeds, the plasmodium enlarges, and the nuclei divide synchronously about once every 24 hours. The plasmodium may become very large, with millions of nuclei, but, ultimately, when conditions are right, it forms a series of small bumps, each of which becomes a small, fruiting body (a structure that bears the spores). During this process the nuclei undergo meiosis, and the final haploid nuclei are then isolated into uninucleate spores (reproductive bodies).

Development of a plasmodium

Many algae (*e.g.*, the Siphonales and related groups) are multinucleate. In most instances the nuclei are in one common cytoplasm within a large and elaborate organism surrounded by a hard cell wall. As the wall becomes extended, the nuclei, which wander freely in the central cavity, undergo repeated mitoses. Again, either during the formation of zoospores (asexual reproductive cells) or after meiosis during gamete formation, a massive progressive division occurs. The most unusual of such organisms is the marine alga Acetabularia; many nuclei stay clumped together in one compound nucleus in the rootlike base, which often is as much as two inches (five centimetres) away from the tip of the plant. The compound nucleus breaks up just before gamete formation, and the minute individual nuclei undergo meiosis and wander to the elaborate tip structures, where they are released as uninucleate gametes.

Syncytial organisms raise the question of whether or not cells, in the strict sense, are necessary for the development of large organisms. Syncytia are also found in animals—*e.g.*, in the early stages of development of fishes and insects—and in the voluntary muscles of man. The proposal of the 19th-century botanist Julius von Sachs is generally considered a satisfactory answer to this question; he suggested that the important matter was the existence not of a cell membrane but of a certain amount of cytoplasm surrounding a nucleus and acting as a unit of metabolism, which he called an energid. Cell reproduction, therefore, might be considered a special case of energid reproduction.

Reproduction of organisms. In single-celled organisms (*e.g.*, bacteria, protozoans, many algae, and some fungi), organismic and cell reproduction are synonymous, for the cell is the whole organism. Details of the process differ greatly from one form to the next and, if the higher ciliate protozoans are included, can be extraordinarily complex. It is possible for reproduction to be asexual, by simple division, or sexual. In sexual unicellular organisms the gametes can be produced by division (often multiple fission, as in numerous algae) or, as in yeasts, by the organism turning itself into a gamete and fusing its nucleus with that of a neighbour of the opposite sex, a process that is called conjugation. In ciliate protozoans (*e.g.*, Paramecium), the conjugation process involves the exchange of haploid nuclei; each partner acquires a new nuclear apparatus, half of which is genetically derived from its mate. The parent cells separate and subsequently reproduce by binary fission. Sexuality is present even in primitive bacteria, in which parts of the chromosome of one cell can be transferred to another during mating.

Vegetative reproduction

Multicellular organisms also reproduce asexually and sexually; asexual, or vegetative, reproduction can take a great variety of forms. Many multicellular lower plants give off asexual spores, either aerial or motile and aquatic (zoospores), which may be uninucleate or multinucleate. In some cases the reproductive body is multicellular, as in the soredia of lichens and the gemmae of liverworts. Frequently, whole fragments of the vegetative part of the organism can bud off and begin a new individual, a phenomenon that is found in most plant groups. In many cases a spreading rhizoid (rootlike filament) or, in higher plants, a rhizome (underground stem) gives off new sprouts. Sometimes other parts of the plant have the capacity to form new individuals; for instance, buds of potentially new plants may form in the leaves; even some shoots that bend over and touch the ground can give rise to new plants at the point of contact (see REPRODUCTIVE SYSTEMS, PLANT).

Among animals, many invertebrates are equally well endowed with means of asexual reproduction. Numerous species of sponges produce gemmules, masses of cells enclosed in resistant cases, that can become new sponges. There are many examples of budding among coelenterates, the best known of which occurs in freshwater Hy*dra.* In some species of flatworms, the individual worm can duplicate by pinching in two, each half then regenerating the missing half; this is a large task for the posterior portion, which lacks most of the major organs—brain, eyes, and pharynx. The highest animals that exhibit vegetative reproduction are the colonial tunicates (*e.g.*, sea squirts), which, much like plants, send out runners in the form of stolons, small parts of which form buds that develop into new individuals. Vertebrates have lost the ability to reproduce vegetatively; their only form of organismic reproduction is sexual (see REPRODUCTIVE SYSTEMS, ANIMAL).

In the sexual reproduction of all organisms except bacteria, there is one common feature: haploid, uninucleate gametes are produced that join in fertilization to form a diploid, uninucleate zygote. At some later stage in the life history of the organism, the chromosome number is again reduced by meiosis to form the next generation of gametes. The gametes may be equal in size (isogamy), or one may be slightly larger than the other (anisogamy); the majority of forms have a large egg and a minute sperm (oogamy). The sperm are usually motile and the egg passive, except in higher plants, in which the sperm nuclei are carried in pollen grains that attach to the stigma (a female structure) of the flower and send out germ tubes that grow down to the egg nucleus in the ovary. Some organisms, such as most flowering plants, earthworms, and tunicates, are bisexual (hermaphroditic, or monoecious)—*i.e.*, both the male and female gametes are produced by the same individual. All other organisms, including some plants (*e.g.*, holly and the ginkgo tree) and all vertebrates, are unisexual (dioecious) : the male and female gametes are produced by separate individuals.

Some sexual organisms partially revert to the asexual mode by a periodic degeneration of the sexual process. For instance, in aphids and in many higher plants the egg nucleus can develop into a new individual without fertilization, a kind of asexual reproduction that is called parthenogenesis.

Parthenogenesis

Liie-cycle reproduction. Although organisms are often thought of only as adults, and reproduction is considered to be the formation of a new adult resembling the adult of the previous generation, a living organism, in reality, is an organism for its entire life cycle, from fertilized egg to adult, not for just one short part of that cycle. Reproduction, in these terms, is not just a stage in the life history of an organism but the organism's entire history. It has been pointed out that only the DNA of a cell is capable of replicating itself, and even that process requires specific enzymes that were themselves formed from DNA. Thus, the reproduction of all living forms must be considered in relation to time; what is reproduced is a series of copies that, like the frames of a motion picture, change through time in an exact and orderly fashion.

A few examples serve to illustrate the great variety of life cycles in living organisms. They also illustrate how different parts of the life cycle can change, and the fact that these changes are not confined solely to adult structures. One variation is that of minimum size—that is to say, the differences in the sizes of gametes (mature sex cells) and asexual bodies. An even greater variation in life cycles, however, involves maximum size; there is an enormous difference between a single-celled organism that divides by binary fission and a giant sequoia. Size is correlated with time. A bacterium requires about 30 minutes to complete its life history and divide in two (generation time); a giant sequoia bears its first cones and fertile seeds after 60 years. Not only is the life cycle of the sequoia 10,000,000 times longer than that of the bacterium, but the large difference in size also means that the tree must be elaborate and complex. It contains different tissue types that must be carefully duplicated from generation to generation.

Life cycles of plants. Most life histories, except perhaps for the simplest and smallest organisms, consist of different epochs. A large tree has a period of seed formation that involves many cell divisions after fertilization and the laying down of a small embryo in a hard resistant shell, or seed coat. There then follows a period of dormancy, sometimes prolonged, after which the seed germinates, and the adult form slowly emerges as the shoots and roots grow at the tips and the stem thickens. In some trees the leaves of the juvenile plant have a shape that is quite different from that of the taller, more mature individuals. Thus, even the growth phase may be subdivided into epochs, the final one being the flowering or gamete-bearing period. Some of the parasitic fungi have much more complex life histories. The wheat rust parasite, for example, has alternate hosts. While living on wheat, it produces two kinds of spores; it produces a third kind of spore when it invades its other host, the barberry, on which it winters and undergoes the sexual part of its life cycle.

In plants, variations in the epochs of the life cycle are often centred around the times of fertilization and meiosis. After fertilization the organism has the diploid number of chromosomes (diplophase); after meiosis it is haploid (haplophase). The two events vary in time with respect to each other. In some simple algae (*e.g.*, *Chlamydomonas*), for example, most of the cycle is haploid; meiosis occurs immediately after fertilization. Yet in other algae, such as the sea lettuce (Ulva), two equal haploid and diploid cycles alternate. The outward morphological structures of mature Ulva are indistinguishable; the two cycles can be differentiated only by the size of the cell or nucleus, those of the haploid stage being half the size of those of the diploid stage.

In many of the higher algae, there is a progressive diminution of the haplophase and an increase in the importance of the diplophase, a trend that is especially noticeable in the evolution of the vascular plants (*e.g.*, ferns, conifers, and flowering plants). In mosses, the haplophase, or gametophyte, is the main part of the green plant; the diplophase, or sporophyte, usually is a spore-bearing spike that grows from the top of the plant. In ferns, the haplophase is reduced to a small, inconspicuous structure (prothallus) that grows in the damp soil; the large spore-bearing fern itself is entirely diploid. Finally, in higher plants the haploid tissue is confined to the ovary of the large diploid organism, a condition that is also prevalent in most animals.

Metamorphosis

Life cycles of animals. Invertebrate animals have a rich variety of life cycles, especially among those forms that undergo metamorphosis, a radical physical change. Butterflies, for instance, have a caterpillar stage (larva), a dormant chrysalis stage (pupa), and an adult stage (imago). One remarkable aspect of this development is that, during the transition from caterpillar to adult, most of the caterpillar tissue disintegrates and is used as food, thereby providing energy for the next stage of development, which begins when certain small structures (imaginal disks) in the larva start growing into the adult form. Thus, the butterfly undergoes essentially two periods of growth and development (larva and pupa–adult) and two periods of small size (fertilized egg and imaginal disks). A somewhat similar phenomenon is found in sea urchins; the larva, which is called a pluteus, has a small, wartlike bud that grows into the adult while the pluteus tissue disintegrates. In both examples it is as if the organism has two life histories, one built on the ruins of another.

Another life-cycle pattern found among certain invertebrates illustrates the principle that major differences between organisms are not always found in the physical appearance of the adult but in differences of the whole life history. In the coelenterate Obelia, for example, the egg develops into a colonial hydroid consisting of a series of branching Hydra-like organisms called polyps. Certain of these polyps become specialized (reproductive polyps) and bud off from the colony as free-swimming jellyfish (medusae) that bear eggs and sperm. As with caterpillars and sea urchins, two distinct phases occur in the life cycle of Obelia: the sessile (anchored), branched polyps and the motile medusae. In some related coelenterates the medusa form has been totally lost, leaving only the polyp stage to bear eggs and sperm directly. In still other coelenterates the polyp stage has been lost, and the medusae produce other medusae directly, without the sessile stage. There are, furthermore, intermediate forms between the extremes.

NATURAL SELECTION AND REPRODUCTION

The significance of biological reproduction can be explained entirely by natural selection (see EVOLUTION: Natural selection). In formulating his theory of natural selection, Charles Darwin realized that, in order for evolution to occur, not only must living organisms be able to reproduce themselves but the copies must not all be identical; that is, they must show some variation. In this way the more successful variants would make a greater contribution to subsequent generations in the number of offspring. For such selection to act continuously in successive generations, Darwin also recognized that the variations had to be inherited, although he failed to fathom the mechanism of heredity (*q.v.*). Moreover, the amount of variation is particularly important. According to what has been called the principle of compromise, which itself has been shaped by natural selection, there must not be too little or too much variation: too little produces no change; too much scrambles the benefit of any particular combination of inherited traits.

Of the numerous mechanisms for controlling variation, all of which involve a combination of checks and balances that work together, the most successful is that found in the large majority of all plants and animals—*i.e.*, sexual reproduction. During the evolution of reproduction and variation, which are the two basic properties of organisms that not only are required for natural selection but are also subject to it, sexual reproduction has become ideally adapted to produce the right amount of variation and to allow new combinations of traits to be rapidly incorporated into an individual.

The evolution of reproduction. An examination of the way in which organisms have changed since their initial unicellular condition in primeval times shows an increase in multicellularity and therefore an increase in the size of both plants and animals. After cell reproduction evolved into multicellular growth, the multicellular organism evolved a means of reproducing itself that is best described as life-cycle reproduction. Size increase has been accompanied by many mechanical requirements that have necessitated a selection for increased efficiency; the result has been a great increase in the complexity of organisms. In terms of reproduction this means a great increase in the permutations of cell reproduction during the process of evolutionary development.

Size increase also means a longer life cycle, and with it a great diversity of patterns at different stages of the cycle. This is because each part of the life cycle is adaptive in that, through natural selection, certain characteristics have evolved for each stage that enable the organism to survive. The most extreme examples are those forms with two or more separate phases of their life cycle separated by a metamorphosis, as in caterpillars and butterflies; these phases may be shortened or extended by natural selection, as has occurred in different species of coelenterates.

Efficiency of reproduction

To reproduce efficiently in order to contribute effectively to subsequent generations is another factor that has evolved through natural selection. For instance, an organism can produce vast quantities of eggs of which, possibly by neglect, only a small percent will survive. On the other hand, an organism can produce very few or perhaps one egg, which, as it develops, will be cared for, thereby greatly increasing its chances for survival. These are two strategies of reproduction; each has its advantages and disadvantages. Many other considerations of the natural history and structure of the organism determine, through natural seiection, the strategy that is best for a particular species; one of these is that any species must not produce too few offspring (for it will become extinct) or too many (for it may also become extinct by overpopulation and disease). The numbers of some organisms fluctuate cyclically but always remain between upper and lower limits. The question of how, through natural selection, numbers of individuals are controlled is a matter of great interest; clearly, it involves factors that influence the rate of reproduction.

The evolution of variation control. Because inherited variation is largely handled by genes in the chromosomes, organisms that reproduce sexually require a single-cell stage in their life cycle, during which the haploid gamete of each parent can combine to form the diploid zygote. This is also often true in organisms that reproduce asexually, but in this case the asexual reproductive bodies (*e.g.*, spores) are small and hence are effectively dispersed.

The amount of variation is controlled in a large number of ways, all of which involve a carefully balanced set of factors. These factors include whether the organism reproduces asexually or sexually; the mutation (gene change) rate; the number of chromosomes; the amount of exchange of parts of chromosomes (crossing over); the size of the individual (which correlates with complexity and generation time); the size of the population; the degree of inbreeding versus outbreeding; and the relative amounts and position of haploidy and diploidy in the life cycle. It is clear, therefore, that the mode of reproduction influences the amount of variation and vice versa; the two together permit natural selection to operate, and selection in turn modifies the mechanisms of reproduction and variation.

BIBLIOGRAPHY. JAMES WATSON, The *Molecular Biology of the Gene,* 2nd ed. (1970), an up-to-date summary of molecular replication by one of the pioneers in the field; HAROLD C. BOLD, *The Plant Kingdom,* 2nd ed. (1964), a brief, general botany textbook that clearly describes the different modes of plant reproduction; ROBERT D. BARNES, *Invertebrate Zoology* (1963), the reproduction of each major invertebrate group is discussed and illustrated; ROBERT T. ORR, *Vertebrate Biology,* 2nd ed. (1966), contains a good general discussion of vertebrate reproduction; J.T. BONNER, *Size* and *Cycle* (1965), a very general discussion of the evolutionary significance of life cycles in animals and plants.

(J.T.Bo.)

Reproductive Behaviour

Reproductive behaviour in animals includes all the events and actions that are directly involved in the process by which an organism generates at least one replacement of itself. In an evolutionary sense, the goal of an individual in reproduction is not to perpetuate the population or the species; rather, relative to the other members of its population, it is to maximize the representation of its own genetic characteristics in the next generation. The dominant form of reproductive behaviour for achieving this purpose is sexual rather than asexual, although it is easier mechanically for an organiam simply to divide into two or more individuals. Even many of the organisms that do exactly this—and they are not all the so-called primitive forms--every so often intersperse their normal asexual pattern with sexual reproduction.

This article considers the modes of behaviour directly involved in the asexual and sexual reproduction of invertebrates and vertebrates; it does not include man. For a discussion of man's behaviour as it relates to his reproductive activities, the reader should refer to the article SEXUAL BEHAVIOUR, HUMAN. The articles BEHAVIOUR, ANIMAL and SOCIAL BEHAVIOUR, ANIMAL also contain information related to the subject of reproductive behaviour. For more complete details concerning the anatomical, physiological, neurological (nervous), and mechanical aspects of reproduction, see REPRODUCTION; and REPRODUCTIVE SYSTEMS, ANIMAL.

GENERAL FEATURES

The dominance of sexual reproduction. Two explanations have been given for the dominance of sexual reproduction. Both are related to the fact that the environment in which an organism lives changes in location and through time; the evolutionary success of the organism is determined by how well it adapts to such changes. The physiological and morphological aspects of an organism that interact with the environment are governed by the organism's germ plasm—the genetic materials that determine hereditary characteristics. Unlike asexual methods, sexual reproduction allows the reshuffling of the genetic material, both within and between individuals of one generation, resulting in an array of offspring, each with a genetic makeup different from that of its parents.

According to proponents of the so-called long-term theory for the dominance of sexual reproduction, sexual reproduction will replace asexual reproduction in the evolutionary development of an organism because it assures greater genetic variability, which is necessary if the species is to keep pace with its changing environment. According to proponents of the short-term theory, however, the above argument implies that natural selection acts on groups of organisms rather than on individuals, which is contrary to the Darwinian concept of natural selection (see EVOLUTION: Natural selection). They prefer to view the advantages of sexual reproduction on a more immediate and individual level: an organism employing sexual reproduction has an advantage over one employing asexual means because the greater variety of offspring produced by the former results in a larger number of genes being transmitted to the next generation. The latter view is probably more nearly correct, especially in violently fluctuating and unpredictable environments. The former theory is probably correct when viewed in terms of its advantage to individuals that are spreading in geographic range, thereby increasing the likelihood of encountering different environments.

Natural selection and reproductive behaviour. Natural selection places a premium on the evolution of those physiological, morphological, and behavioral adaptations that will increase the efficiency of the exchange of genetic materials between individuals. Organisms will also evolve mechanisms for sensing whether or not the environment is always permissive for reproduction or if some times are better than others. This involves not only the evolution of environmental sensors but also the concurrent evolution of mechanisms by which this information can be processed and acted upon. Because all seasons are not usually equally conducive, individuals whose genetic backgrounds result in their reproducing at a more favourable rather than less favourable period will eventually dominate succeeding generations. This is the basis for the seasonality of reproduction among most animal species.

Natural selection also results in the evolution of systems for transmitting and receiving information that will increase the efficiency of two individuals' finding each other. These attraction systems are usually, but not always, species specific (see SPECIES AND SPECIATION).

Need for genetic variability

Once the proper individuals have found each other, ıt is clearly important that they are both in a state of reproductive readiness. That their sensory receptors are tuned to the same environmental stimuli is usually sufficient to achieve this synchrony (proper timing) in the lower organisms. Apparently, however, this is not enough in the more complex organisms, in which the fine tuning for reproductive synchrony is accomplished chiefly by a process called courtship. Another evolutionary necessity is a mechanism that will guide the partners into the proper orientation for efficient copulation. Such mechanisms are necessary for both internal and external fertilization, especially the latter, where improper orientation could result in a complete waste of the eggs and sperm.

In most organisms, the period of greatest mortality occurs between birth or hatching and the attainment of maturity. Thus, it is not surprising that some of the most elaborate evolutionary adaptations of an organism are revealed during this period. Natural selection has favoured an enormous variety of behaviour in both parents and offspring that serves to ensure the maximum survival of the young to maturity. In some animals this involves not only protecting the young against environmental vicissitudes and providing them with adequate nutrition but also giving them, in a more or less active manner, the information they will need to reproduce in turn.

External and internal influences **on** reproductive behaviour. As mentioned at the beginning of this article, the anatomical, physiological, and neurological aspects of reproduction and behaviour are dealt with in other articles. It is useful here, however, to consider briefly the external and internal factors that initiate reproductive behaviour.

Environmental influences. Light, usually in the form of increasing day length, seems to be the major environmental stimulus for most vertebrates and many invertebrates, especially those living in areas away from the Equator. That this should be such an important factor is quite reasonable in an evolutionary sense: increasing day length signifies the onset of a favourable period for reproduction. In equatorial regions, where changes in day length are usually insignificant throughout the year, other environmental stimuli, such as rain, predominate.

Superimposed on day length are usually several other factors, which, if lacking, often override the stimulating effect of light. Many insects, for example, will not initiate a reproductive cycle if they lack certain protein foods. Many animal groups have an internal cycle of cellular activity that must coincide with the external factors before reproduction can occur; a familiar example is the estrous cycle in most mammals except primates. Females display sexual behaviour (are "in heat") only during a brief period when they have ovulated an egg.

Hormonal influences. Although the exact way by which light affects the reproductive cycle is still disputed, it undoubtedly varies from group to group. In birds, light passes either through the eyes or through the bony tissue of the skull and stimulates the development of certain cells in the forepart of the brain. These cells then secrete a substance that stimulates the anterior pituitary gland, which is located at the base of the brain, to produce an array of regulatory substances (hormones), called gonadotropins, that are carried by the blood to the gonads (ovaries and testes), where they directly stimulate the development of eggs and sperm. The gonads, in turn, produce the sex hormones—estrogen in the female and testosterone in the male—that directly control several overt aspects of reproductive behaviour.

Insect pheromones

Unlike the higher animals, the gonads of insects apparently do not themselves secrete hormones. Instead, stimulation by the corpus allatum, an organ in insects that corresponds in function to the pituitary gland, causes the secretion of liquid substances on the body surface. These substances are transmitted as liquids, or, even more significantly, as gases, to the recipient, in which they are usually detected by olfaction or taste. Such substances, which are called ectohormones or pheromones, may serve as the major regulation and communication system for reproduction as well as other behaviour in insects.

In the absence of all other stimuli, many types of sexual behaviour can be induced simply by an injection of the appropriate gonadal hormone. Conversely, removal of the gonads usually inhibits most sexual behaviour. The apparent failure of complete hormonal control over reproductive behaviour has been a subject of much investigation and dispute. There is much evidence that many types of reproductive behaviour are or can be controlled solely by neural mechanisms, bypassing the hormonal system and any effect that it might exert on the nervous system to produce behaviour. Several types of reproductive behaviour controlled solely or almost solely by neural mechanisms are involved in or triggered by the processes that are initiated by courtship.

MODES OF SEXUAL **ATTRACTION**

The chief clues by which organisms advertise their readiness to engage in reproductive activity are visual, auditory, and olfactory in nature. Most animals use a combination of two modes; sometimes all three are used.

Visual clues. The appearance of many higher vertebrates changes with the onset of reproductive activity. The so-called prenuptial molt in many male birds results in the attainment of the nuptial plumage, which often differs radically from that possessed by the bird at other times of the year or from that possessed by a nonreproductive individual. The hindquarters of female baboons become bright red in colour, which indicates, or advertises, the fact that she is in estrus and sexually receptive. Such changes in appearance are less common in the lower animals but do occur in many fishes, crabs,, and cephalopods (*e.g.*, squids and octopuses).

The role of aggressive behaviour

Often associated with changes in appearance are changes in behaviour, particularly the increase in aggressive behaviour between males, often a prime feature in attracting females; such changes have interesting evolutionary implications. In certain grouse, for example, females are most attracted to males that engage in the greatest amount of fighting. No doubt, fighting in some groups of mammals also serves this function as well as others (see AGGRESSIVE BEHAVIOUR).

In many animals the rise in aggression takes the form of territoriality, in which an individual, usually a male, defends a particular location or territory by excluding from it all other males of his own kind. Occasionally, other species are also excluded when it is to the advantage of the defending individual to do so. Territorial behaviour involves many functions, not all of which are directly concerned with reproduction. For purposes of advertising, however, aggression probably reduces the amount of interference between males and also makes it easier for females to find males at the proper time.

Auditory clues. The fact that sound signals can travel around barriers, whereas visual signals cannot, accounts for their widespread use in indicating sexual receptiveness, especially in frogs, insects, and birds. Like visual signals, a sound for advertising purposes usually encodes several pieces of information; for example, the signals usually reveal to the receiver the caller's species, its sex, and, in some cases, whether or not it is mated. The vocalizations of one type of frog also reveal the number of other males located nearby. This information, a critical clue for females, is a measure of how good the habitat is for depositing eggs. The sounds produced by the wings of mosquitoes attract females and are species specific. Humans have taken advantage of this signal by using artificial sound generators to eradicate certain mosquitoes. Advertising signals also serve to repel other males; a classical example is the territorial song of many songbirds.

Olfactory clues. Researchers have now become aware of the enormous amount of information that is passed between animals bv chemical means. Well-known are the urine, feces, anh scent markings employed by most mammals to delimit their breeding territories and to advertise their sexual state. Males of a number of mammals are capable of determining if a female will be sexually receptive simply by smelling her urine markings. A substance in the urine of male mice, on the other hand, actually induces and accelerates the estrous cycle of females. A female gypsy moth is able to attract males thousands of metres downwind of it simply by releasing minute quantities of its sex pheromone each second. It has been calculated that one female silkworm moth carries only about 1.5 micrograms (0.0015 gram) of its sex attractant, called bombykol, at any given moment; theoretically, this is enough to activate more than 1,000,000,000 males, surely more than exist in any one place at any time. The sex attractant of barnacles, which are rather sessile (sedentary) organisms, causes individuals to aggregate during the breeding period.

The potency of pheromones

One other possible channel of communication occurs in a few fishes, namely electric discharge. Evidence now accumulating suggests that weak electrical fields and discharges in the Mormyridae of Africa and the Gymnotidae of South America represent the major mode of social interaction in these families.

COURTSHIP

Synchrony is the major factor in achieving fertilization in the lower animals, particularly in aquatic forms. In most of these groups, the eggs and sperm are simply discharged into the surrounding water, and fertilization occurs externally. The parents may never meet, so to speak. It might be assumed that this procedure would be roughly the same in the higher animals, with perhaps more overt behaviour to achieve synchrony, and that, after the two individuals found each other, fertilization would proceed fairly quickly. This is usually not the case, however. Although fertilization in the higher terrestrial forms involves contact during copulation, it has been suggested that all of the higher animals may have a strong aversion to bodily contact. This aversion is no doubt an antipredator mechanism: close bodily contact signifies being caught. Since females are in an especially helpless situation during copulation, they are particularly wary about bodily contact. In addition, males are particularly aggressive during the breeding period, which further increases the uncertainty of both individuals. These difficulties were solved by the evolution of a collection of behaviours called courtship. Courtship has been defined as the heterosexual reproductive communication system leading to the consummatory sexual act.

Courtship behaviour has many advantages and functions, not the least of which is the reduction of hostility between the potential sex partners, especially in species in which the male actively defends a territory. The major aspects of such behaviour seem to be appearance, persistence, appeasement, persuasion, and even deception. Because courtship behaviour involves the transmission of information by means of signals, it is useful to define at this point a peculiar and important group of social signals called displays.

A social signal may be considered any behavioral pattern that effectively conveys information from one individual to another. The term display has been restricted by some authorities to social signals that not only convey information but that, in the course of evolution, have also become "ritualized." In other words, such signals have become so specialized and exaggerated in form or function that they expressly facilitate a certain type of communication. The visual, auditory, olfactory, tactile, or other patterns by which organisms advertise their readiness to engage in reproductive activity provide examples of displays. Clearly, the kinds of displays utilized by organisms depend on the sensory receptors of the receiver. Whereas higher vertebrates tend to use visual and auditory displays, insects tend toward olfactory and tactile displays.

In animals in which the male takes on a wholly different appearance during the breeding period, natural selection has eliminated from the female's appearance the "aggressive badges" of males that provoke fighting. It is not without significance that the appearance of the adult female in many species is much like that of the juvenile; this implies to the male a friendly, nonaggressive relationship. When one male approaches another that has intruded into the former's territory, the outsider may either return the aggressive display or flee. Females, however, usually quietly back up slightly and then slowly move forward again. With each approach the male's hostility lessens toward this appeasing, increasingly familiar individual. Often, as in many birds, the females resort to displays that resemble the food-begging behaviour normally seen in the young. Males frequently respond to this display by actually regurgitating food. Male spiders of some species offer the larger and more aggressive females food as bait, and copulation occurs while the female is eating the food rather than her potential mate. Mutual feeding displays, often with nonedible items, are engaged in by a number of insects and birds. In the courtship behaviour of several birds, extremely elaborate displays are utilized to hide the bill from the potential partner, because the bills of these birds are their chief weapons. Some aspects of nest building have been incorporated into the displays of such birds as penguins. Early in the relationship between the individuals, one or both may offer the other stones that are placed in a pile. The actual nest is not constructed until much later, however.

The common denominator in courtships is that the displays resemble functional behaviours that are appropriate to friendly, bonded situations, such as those between parents and between parents and their offspring. The degree of elaborateness of the display is governed by a number of factors. One is to prevent cross-mating between different species, an occurrence that usually results in the waste of the eggs and sperm. Any specific aspect—*i.e.*, one or more displays—used by an organism in species discrimination is called an isolating mechanism. In many species, the majority of the displays between individuals are a series of identity checks.

Species discrimination

Another factor governing the complexity of displays is the length of time that the pair bond will endure. Brief relationships are usually, but not always, associated with rather simple courtship activity. In a number of insects, birds, and mammals, the males display on a common courtship ground called a lek or an arena. Females visit these courtship areas, copulate, and leave. The males do not participate in any aspect of parental care; the bond lasts but a few seconds. Yet despite the brevity of this relationship, in no other courtship system is there the development of such elaborate and almost fantastic displays in both the movements and appearances of the males.

POST-FERTILIZATION BEHAVIOUR

Various types of behaviour ensure that a maximum number of fertilized eggs or young will survive to become reproductive adults. Clearly, the number of eggs produced and their size represents a balance achieved by natural selection. This balance conforms to some optimum compromise between producing many eggs containing little food for the development of young or fewer eggs with more provisions.

There has been considerable controversy about the factors that limit the number of offspring an organism can produce. It has been suggested that, among animals in which the offspring are dependent on the parents for varying lengths of time, clutch or litter size has been adjusted through natural selection to the maximum number of offspring that the parents, on the average, can feed. There are, on the other hand, organisms that do not practice parental care and produce millions of eggs. According to one popular school of thought, these species have such a high fecundity (productivity) because the eggs and larvae suffer a very high mortality rate. Hence, it is necessary for such animals to produce thousands, even millions, of eggs just to obtain a few reproductive adults. An opposing school of thought, however, says that such species have high mortality rates because of their great fecundities. Similarly, low death rates would be the consequence of low fecundity.

Protective adaptations. Several adaptations have evolved to protect the eggs and larvae of species not attended by adults. In one such adaptation, the eggs or larvae are distasteful, inedible, or apparently harmful to potential enemies. The eggs of the jellyfish *Bougainvillia*, for example, contain stinging cells on the surface that deter predators. Many female butterflies deposit their eggs on plants that contain poisonous compounds, which the larvae incorporate into their bodies, making them distasteful. When disturbed many insect larvae, especially those that are camouflaged, give a so-called startle display; several caterpillars, for example, raise their heads as if to bite or their hindparts, in the manner of a wasp, as if to sting. Others suddenly present striking colour patterns previously hidden. Most of these displays have been shown experimentally to be effective deterrents against predators.

Caring for offspring. Animals that do not care for their young must provide for the nutritional needs of their offspring. One way of doing so is by producing an egg with a sufficiently large yolk supply that the young, when hatched, are already at an advanced, almost independent state. **A** peculiar example of this is found in the incubator birds (Megapodiidae), which cover their large eggs with soil and debris to create a mound of considerable depth, effectively providing heat for the developing eggs. After a very long incubation period, the young emerge as fully feathered miniature adults and are capable of flying in 24 hours. Before sealing the nest that they make for their eggs, many insects, such as certain solitary wasps, stock the nest with food. In a more bizarre manner, other solitary wasps place one egg in the body of an insect or spider previously paralyzed by the wasp. Upon hatching, the larva eats the still living host.

Social parasitism

Social parasitism, another fascinating aspect of post-fertilization behaviour, is found in certain insects and birds. In this case, the true parents do not care for their eggs or offspring; rather, they place them under the foster care of other species, often, but not always, to the detriment of the foster parents' offspring. In certain parasitic species of cuckoos, the females are divided into groups, or gentes, each of which lays eggs with a colour and pattern unlike those of the other groups. The females of each group usually select a particular species as the host, and, more often than not, the eggs of the parasite closely resemble those of the potential foster parent. This mimicry has evolved because many host species throw eggs not resembling their own out of the nest. Some young cuckoos also exhibit a behaviour called backing, in which they push out the other nestlings and monopolize the food supply.

Parental *care*. Among the organisms that remain with the eggs or offspring, one particular behaviour is striking —that of nest construction to keep the eggs and larvae in one spot and to protect them against predators as well as such environmental factors as sun and rain. The placement of a nest usually serves an antipredatory purpose, as in birds that put their nests near those of social wasps or stinging ants. Although they are not normally thought to do so, many mammals, particularly rodents and carnivores, construct special nests, dens, or burrows solely for reproductive purposes.

A number of fishes build nests made of bubbles that not only hold the eggs together but also provide the oxygen necessary for the developing embryos. Other fishes, particularly those that live in oxygen-poor waters, display elaborate fanning behaviour to keep the water moving around the eggs. In some fishes, the female incubates the egg in her mouth, thus providing protection against predators as well as constant aeration. The fry (young) of some of these mouthbreeders travel in a school near the parent. When danger approaches, they flee into the parent's mouth and later swim out after the danger passes.

Birds have the problem of keeping the eggs at an optimum temperature for development of the embryo. With the onset of egg laying in many species, the feathers of the lower abdomen are lost, and the skin in that area becomes thickened and highly vascularized (filled with blood vessels), forming the so-called brood patches. Usually the female develops these patches, which serve to transfer more effectively to the eggs the warmth from the adult's body. It has been shown that, like much of parental behaviour in the higher vertebrates, brood patches and "broodiness" are controlled by several hormones, combined with visual and tactile stimuli. Chief among these hormones is prolactin, which also controls the production of pigeon milk, a cheeselike substance produced only in the crops of adult doves and pigeons and fed to the nestlings by regurgitation.

Although there are some outstanding exceptions, most young mammals are completely helpless at birth. This helplessness is most striking in the marsupials (*e.g.*, opossums and kangaroos), in which the young are born at a very early stage of development; they crawl through the mother's hair to the brood pouch, where they attach themselves to a nipple and their development continues for many more months.

Licking of newborn

An early characteristic behaviour in mammals following birth is that of the mother licking the newborn. This serves at least two functions--one is general cleanliness to avoid infections or the attraction of parasites; the other would appear to be purely social. If a newborn mammal is removed from its mother and cleaned elsewhere before she can lick it, she usually will not accept it. Thus, licking behaviour also serves, in some manner, to establish a unique relationship between the mother and her offspring. Another characteristic mammalian behaviour is the suckling response of the newborn. Although this behaviour has been claimed to be the perfect instinctive response, it apparently is not so in many species; the trial-and-error period during which the newborn discovers the nipple, however, is quite short.

In birds, especially those that nest on the ground, one of the first adult responses to the hatching of the eggs is to remove the conspicuous eggshells from the area of the nest. It has been shown experimentally that, in gulls at least, this is an important antipredatory measure. When birds hatch, they have the ability to stretch their heads and to gape for food in response to any mechanical disturbance, such as that produced when the parent lands on the nest. Later in development, they stretch and gape only when the parents appear. This is another type of adaptive, antipredatory behaviour, as it would be dangerous for the nestlings to gape and vocalize in response to any environmental disturbance.

Group care. The ability of an animal to identify its own offspring at an early stage is apparently not important in animals that nest or are solitary breeders; offspring in the nest belong to that parent. In colonially breeding species or in those where the offspring of different parents are likely to become mixed, however, natural selection has favoured the evolutionary development of behaviour that makes possible the recognition by the parent of its own offspring, thereby avoiding the danger of expending energy on offspring that do not possess the parent's genes.

There is, on the other hand, the situation in which the offspring are cared for by individuals who are not the parents. This phenomenon occurs among the social insects in particular and also among several groups of birds and mammals; future investigations may show it to be even more widespread. In such birds as the anis, the effective breeding group consists of several females and males. One nest is constructed in which all the females deposit their eggs, and all individuals participate in the care of the resulting offspring. In certain jays (Corvidae), the offspring of one generation participate in the care of the offspring of the next or another generation, but the exact family relationships among the participants are not clear.

In the social insects, this type of parental behaviour apparently results from the peculiar genetic relationships between the individuals in most social-insect colonies (termites are among the exceptions). The female and, in the termites, both the male and the female can control by chemical means the kinds (called castes in ants and termites) and sexes of the offspring. An outstanding feature of such colonial insects as the honeybee is that the majority of the individuals produced by the queen are sterile; these are the workers, the individuals who care for and feed both the queen and her offspring, the sibs of the workers.

The queen is diploid in genetic makeup; that is to say, half of her genes are derived from her mother and half from her father. The males (drones) are haploid; that is, they have only half the genes possessed by the queen, all of them derived from the mother. A queen produces eggs fertilized by sperm she has retained in her body from the mating flight; thus the individuals produced are diploid, but, unlike the queen, they are sterile. This sterility results indirectly from a chemical secreted by the queen, called the queen substance. It inhibits the workers from building special brood cells that give rise to sexually developed individuals. If the queen fails to secrete this substance because of age or death, the workers immediately construct special brood cells with a substance they secrete; called royal jelly, it is necessary for the development of a larva then destined to be a queen.

How can the evolution of sterility in workers and their care of offspring not their own be accounted for? One possible explanation concerns the coefficient of relationship (the number of genes on the average shared in common) among the individuals of a colony. Because of the peculiar haplodiploid mode of sex determination, the workers (sisters) share all the genes from their father and, on the average, half of those from their mother. Since each worker receives half of its genes from the father and half from the mother, the average genes shared between any two workers (sisters) is three-fourths. But between mother (the queen) and daughter (a worker) this average is only one-half. The offspring (the sterile workers), therefore, may contribute more to their fitness (the maximum representation of their genes in the next generation) by caring for their sisters than by providing an equal amount of care to their "own" offspring, had they been fertile rather than sterile. A drone, on the other hand, has a coefficient of relationship with one of his sterile sisters of only one-fourth, but retains a relationship of one-half with his mother and daughters (future sterile workers). This explains why workers provide more care for their sisters than for their brothers, and why the workers eventually drive off the almost useless drones, which are relatively scarce (having resulted from unfertilized eggs), from the colony. Because sisters share more genes with each other than with their brothers, they maximize the chances of these genes surviving into the next generation by providing more care for their sisters.

This explanation of group care and extreme sociality does not account for all cases. Indeed, termites are perhaps the most extreme among animals in these respects but lack the haplodiploid sex determination mechanism. In addition, several groups having this mechanism have not evolved extreme brood care and sociality. Other factors have to interact for these systems to evolve, but it is not yet clear what they are.

REPRODUCTIVE BEHAVIOUR IN INVERTEBRATES

Protozoans and sponges. Most protozoans (one-celled organisms) reproduce asexually, usually by fission (splitting in two); in some species, however, sexual as well as asexual reproduction occurs and may be complex. The colonial organism *Volvox*, which may be either of one "sex" or composed of cells of both sexes, produces true eggs and sperm. **A** chemical substance released by "females" induces the production of sperm packets; following the union of the egg and sperm, the parent colony dissolves, and the zygote (fertilized egg) is released.

Conjugation

Another form of reproduction in protozoans is conjugation, in which organisms such as *Paramecium* fuse together briefly to exchange nuclear products. This results in a reshufflingof hereditary characteristics just as occurs in true sexual reproduction in higher animals. In some species of *Paramecium*, there are mating types, and an individual is of one type or the other. Opposite types apparently recognize each other by a chemical (pheromone) that is released on their body.

In the lower metazoans (multicellular organisms), reproduction is also by both asexual and sexual means. As befits their sessile life-style and low population densities, sponges that reproduce sexually are usually hermaphroditic; that is, each individual is capable of producing both sperm and eggs, but often at different times to prevent self-fertilization. The sperm are swept by water currents into another sponge, where they are picked up by specialized cells called choanocytes and carried to the egg. Fertilization takes place when a choanocyte fuses with the egg. The free-swimming larval stage that is produced is of short duration, after which the organism settles on the bottom and becomes a new adult sponge.

Coelenterates. Hydroids, jellyfishes, sea anemones, and corals of the phylum Coelenterata, or Cnidaria, reproduce by a variety of mechanisms. A familiar coelenterate animal, the freshwater *Hydra*, usually reproduces asexually by budding, a process by which small portions of the adult structure become new, but genetically identical, individuals. Hydras are also dioecious; that is, each individual produces either sperm or eggs. In many temperate-zone species of *Hydra*, sexual reproduction occurs during the autumn; the fertilized eggs enable the species to survive the winter.

Most of the other hydrozoans are colonial organisms, often occurring in polyp and medusal (umbrella-shaped) forms. In a colony, reproductive individuals called gonophores develop into free-swimming organisms (medusae) that reproduce sexually. Fertilization can be either external or internal; if external, the eggs are shed directly into the water. Internal fertilization results in larvae that swim out of the parent and soon settle on a surface, where they develop into another hydroid colony.

Sea anemones and the polyps of corals reproduce asexually by budding or sexually. In the sexual mode, sea anemones have both dioecious and hermaphroditic species. One interesting aspect of sea anemones, which undergo internal fertilization, is that they are among the first lower animals to provide parental care. The larvae remain inside the adult until they are ready to metamorphose (change in form), at which time they swim from the parent's mouth and settle on its base, remaining there until they develop tentacles, after which they move away from the parent's protection.

Flatworms **and** rotifers. The reproductive structures of flatworms (phylum Platyhelminthes) resemble those found in the higher groups. Such flatworms as the land and freshwater planarians are hemaphrodites. Although some species can reproduce asexually by splitting in two, most engage in copulation. Some freshwater planarians can produce both thin-shelled summer eggs, which hatch in a short time, and thick-shelled winter eggs, which are resistant to freezing and hatch in the spring. An apparently unique situation in many planarians is that nutrition for the embryo is supplied by the addition of separate cells to the zygote, after which the entire mass is enclosed in the shell; more commonly, the yolk is incorporated within the structure of the zygote itself.

In the rotifers (phylum Aschelminthes), small but abundant freshwater animals, reproduction is usually sexual, and the sexes are separate. Copulation occurs by injection of sperm anywhere in the body wall of the female. Many species found in temporary ponds and streams exhibit a peculiar reproductive behaviour that is well adapted to their transient environment: they produce different kinds of eggs at different times of the year. One egg type, called amictic, is produced in the early spring. These eggs apparently cannot be fertilized, and the embryo develops without fertilization (parthenogenesis); the result is females with a life-span no longer than two weeks. When the population reaches a peak in the early summer, a second type of egg is produced. If unfertilized, this egg, which is called mictic, results in males. As the male population increases, most mictic eggs become fertilized, resulting in the production of a heavy-shelled dormant egg with much yolk. The dormant egg survives the winter and gives rise to the amictic females of the next spring. Thus, despite the many generations produced in the summer by so-called sexual means, the reshuffling and recombination of genetic material occurs only once a year.

Parthenogenesis

Segmented worms. The marine worms of the class Polychaeta (*e.g.*, clam worms and lugworms of the phylum Annelida) provide the first examples of a kind of courtship behaviour involving both visual and chemical displays initiated by some rather subtle environmental stimuli. Most polychaetes reproduce sexually, and there are two distinct sexes in most species. Either by transformation or budding, many polychaetes produce a reproductive form (epitoke). At a certain time of the year, the epitokes swarm to the ocean surface and engage in mass shedding of eggs and sperm. Some female epitokes of clam worms (Nereis) produce a chemical substance called fertilizin that attracts the male epitokes and stimulates the shedding of sperm. Male epitokes of a polychaete found in the Atlantic Ocean emit a flashing light; females emit a steady light. The light may serve to attract male and female and to aid in species discrimination. The swarming of the palolo worm Eunice in parts of the South Pacific is apparently triggered by an annual and a lunar cycle; the epitokes separate from the parent (atoke) in October or November, during the last part of the lunar cycle.

The class Oligochaeta (phylum Annelida) contains a diversity of both aquatic and terrestrial worms, among which is the familiar earthworm, Lumbricus. Although some aquatic oligochaetes reproduce asexually, the majority are sexual, and all of these are hermaphrodites. At mating, two oligochaetes lie side by side so that the head of one is opposite the tail of the other. Sperm then pass reciprocally into small sacs, where they are temporarily stored. This transfer is more complex in the earthworms, however, because the respective male pores are not in direct opposition; each individual forms a temporary skin canal through which the sperm flow to their respective sacs for storage. The body of oligochaetes has a swollen girdle-like structure, the clitellum, which serves an important function in reproduction. After the eggs have matured, a mucous tube, secreted from the clitellum, slides along the body as the worm moves backward. The stored sperm are discharged into this tube, as are the eggs when the tube slides along the section containing them. As the worm literally passes out of the tube, a mucous, lemon-shaped cocoon forms around the now-fertilized eggs. This cocoon serves as a kind of primitive nest, in which the young hatch.

Many leeches (class Hirudinea), all of which are hermaphrodites, have copulatory behaviour much like that of earthworms. Cocoons are formed in a manner similar to that described above, but in some leeches the cocoon is transparent and remains attached to the parent in which the eggs were developed. After hatching, the young leeches remain attached to the "mother" until they become independent. One African leech gives birth to live young and even possesses a special incubating chamber in its body for the developing embryos.

Mollusks. The animals in the phylum Mollusca (*e.g.*, clams, snails, and squid) display a diversity of reproductive behaviour. The majority of the amphineurans (chitons) and pelecypods (*e.g.*, clams, oysters) are dioecious —*i.e.*, individuals are either male or female. Because most species simply shed their eggs and sperm directly into the sea, individuals tend to form dense aggregations during the breeding period. The environmental factor that triggers the release of eggs and sperm has not yet been established with certainty, but, at least in a few species, after one individual has shed its sex products, the others follow in a kind of chain reaction that is clearly chemical in nature. In some mollusks, however, such as the European oyster, the eggs are retained and brooded.

The snails and slugs include hermaphroditic as well as dioecious species. Copulation in the hermaphroditic land snail Helix is preceded by a curious courtship involving a bizarre tactile stimulation. When the two partners come together, each drives a calcareous dart (the so-called love dart) into the body wall of the other with such force that it is buried deep in the other's internal organs.

The love dart

To avoid predators, some arboreal slugs copulate in mid. air while each partner is suspended by a viscous thread, In the slipper-shell snails (Crepidula), which are rather sessile, all the young are males; their subsequent sex, however, is determined by their nearest neighbour. They remain males as long as they are near a female but change into females if isolated or placed near another male.

Remarkably advanced courtship behaviour in the cephalopods, particularly the squids, involves complex visual displays of movement and changes in colour pattern. Males signify that they are ready for breeding by assuming a distinctive zebra-striped pattern, displaying their fourth arm in a flattened manner, and approaching other individuals with a jerky motion. This fourth arm in squids and the third arm in octopods, called a hectocotylus, is structurally modified for carrying spermatophores, or balls of sperm. The male cuttlefish (Sepia) places the spermatophores in a pocket near the female's mouth, from which the sperm subsequently make their way to the tubes that carry eggs (oviducts). In no squid studied thus far do either of the sexes care for the fertilized eggs, which are laid on vegetation. This is not the case with octopuses, however; at least in Octopus vulgaris, the female broods her large number of eggs (about 150,000) for as long as six weeks. During this period she aerates the egg clusters and keeps them free of detritus, exhibiting remarkable behaviour for an animal that produces so many eggs. Brood care such as this is usually associated only with organisms that produce a small number of eggs.

From *California Fish and Game (1954)*

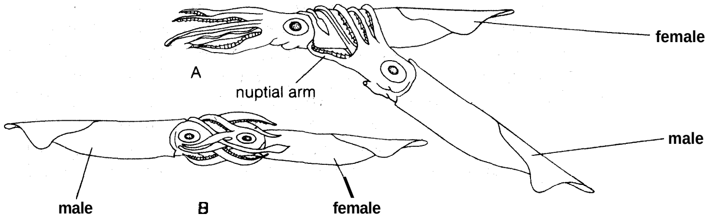


Figure 1: Copulating squid *(Loligo vulgaris).*
(A) Position when spermatophores are transferred to mantle cavity. (B) Position when spermatophores are transferred to sperm receptacle.

Arthropods. Crustaceans. With a few exceptions, barnacles are the only hermaphroditic members of the class Crustacea in the phylum Arthropoda. This is in agreement with the theory that a sessile mode of life tends to be correlated with hermaphroditism. Thus, it is not important for the organism to be near an individual of the opposite sex, but simply to be near any individual of the same species.

Some barnacles are parasitic and have undergone a radical degeneration in form. One, Sacculina, is an example of the way in which the reproductive necessities of one species can profoundly affect the reproductive behaviour of another—in this case, the host. Several cells from a larval barnacle penetrate a crab's body and migrate through the bloodstream until they reach the lower portion of its stomach. The cells then send rootlike projections throughout the crab's body. When the crab molts, the barnacle protrudes a large bulbous portion of its body through the ventral (bottom) surface of the crab. If the crab is a female, its broad shell protects this structure, which contains the barnacle's reproductive organs. The body shape of the male crab, however, is much narrower and does not provide such protection. If the host is a male, therefore, the barnacle first consumes the host's testes; at its next molt, the crab assumes the shape of a female. Should the parasite be removed, the crab regains a male appearance and regenerates its testes.

In the copepods (*e.g.*, sea lice, Cyclops) and the amphipods (*e.g.*, beach fleas), the sexes are mostly separarate, copulation is brief and without elaboration, and the female of both groups broods the fertilized eggs. The eggs of copepods are usually attached in two clusters to the rear of the female; many amphipods have a special pouch on their ventral surface for brooding the eggs. Many copepods and some amphipods are parasitic on fish and on such marine mammals as whales.

In the crustacean order Decapoda, which includes shrimp, crayfish, lobsters, and crabs, the sexes are separate, fertilization is mostly internal, and egg laying usually occurs shortly after copulation. In terrestrial crabs, however, the females of which migrate to salt water to expel the eggs, the sperm are stored, and fertilization and egg laying are delayed for several months after copulation.

Fiddler crabs of the genus ***Uca*** and several other decapods show territorial behaviour, an act that is not very common among invertebrates. As in many groups in which males defend territories, male crabs often differ in appearance from the females. Males are much more brightly coloured than the females, and one of their front claws is greatly enlarged; the mostly dull-coloured females have two small front claws. Depending on the species, males perform either simple or complex rhythmic dances in front of their sand burrows. The waving and vertical movement of the large claw is apparently species specific.

The necessity of spermatophores

As in squids and octopuses, the sperm of primitive terrestrial arthropods—millipedes, centipedes, springtails, and silverfish—are often transferred from males to females in structures called spermatophores. During the transition from an aquatic to a terrestrial mode of life, spermatophores became necessary, particularly for those species that had not developed copulatory organs for direct transmission of sperm. Because sperm transfer in these animals is often complicated and takes considerable time, the delicate sperm would be in danger of drying up, were it not for the moisture contained in the spermatophores. It would appear, therefore, that all species that exhibit indirect sperm transfer in which spermatophores are utilized have not achieved complete independence of water.

Males of most primitive soil-dwelling arthropod species place sperm drops on threads in damp locations or use threads or chemical products to guide females to externally placed spermatophores. Most male millipedes have secondary genital appendages called gonopods, by which they transfer the spermatophore directly to the genital opening of the female. One millipede actually uses a "tool" in sperm transfer; the male rounds a fecal pellet, places a drop of sperm on it, and, using its legs, passes the pellet back along its body to a point opposite the female's genital pore. Paired body projections then are used to inject the sperm into the female, and the pellet is dropped. Males of the common bark-inhabiting millipede ***Polyxenus*** transfer sperm by spinning thin threads on which they place sperm drops; they then construct two parallel thicker threads on which they place a pheromone to attract the female. This chemical and tactile guidance system causes the sperm to become attached to the female's vulva (the external part of the female's genital organs). Males eat the sperm not picked up and replenish it with fresh sperm.

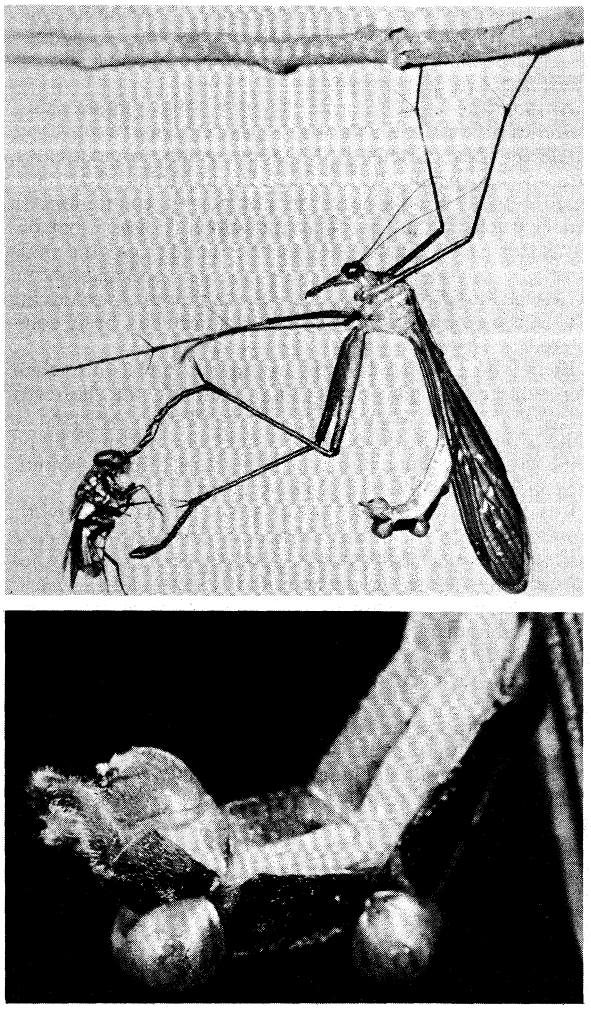
Arachnids. The arachnids (*e.g.,* spiders and scorpions) exhibit the earliest pattern of classical courtship behaviour during which rather ritualized movements are involved. In the true scorpions this behaviour takes the form of the ***promenade à deux,*** in which the male holds the female by her front claws and apparently stings her in a joint near the base of the claw. The ensuing dancelike pattern apparently results from the male seeking a suitable surface upon which to deposit his spermatophore. After he deposits the spermatophore, the male drags the female over it, releasing her after the spermatophore has passed into her genital pore.

Behaviour of male spiders

As mentioned above, many male spiders have a particular problem in approaching the aggressive and predatory female in order to deposit a spermatophore. The hunting behaviour of most spiders is adapted to react to the slightest movement or vibration of the web, causing the spider to rush forward and bite its prey as quickly as possible. Thus, it is not surprising that male spiders have evolved fairly elaborate display movements and patterns to convey their identity. Many males are quite strikingly coloured, providing additional information about their identity. Some males approach the female only at night and vibrate her web in a highly characteristic manner, different from that caused by the struggling of a trapped animal.

Insects. One puzzling aspect about the courtship behaviour of insects is its sporadic nature. Most insects should exhibit behaviour involving approach, identification, and copulation. Yet, whereas male fruit flies (*Drosophila*) often have elaborate displays preceding copulation, male houseflies and blowflies ***(Musca)*** simply fly at any object of the proper size and attempt to copulate with it. The reason for these differences in behaviour may be that some insects do not require courtship. Males of some butterflies and moths, for example, simply wait by the pupa and copulate with the female immediately after she emerges.

It is more likely, however, that the majority of insects have fairly elaborate displays, but man is unable to sense them. The pheromones are, in fact, rather elaborate displays used as sex attractants by many insects; such sensory mechanisms are not usually perceived by man. It has been experimentally demonstrated that the reproductive behaviour of some butterfly species depends heavily on visual clues; similar experiments with other species have failed to show such behaviour. It must be realized, however, that insect vision is quite different from that of vertebrates. Most insects have vision that is sensitive to ultraviolet light, which man and the other vertebrates cannot normally perceive. Butterflies may appear to have identical wing colour patterns under normal light, but, when viewed under ultraviolet light, the patterns differ drastically. Thus, insects that mimic each other in order



BY courtesy of the Commonwealth Scientific and Industrial Research Organization., Division of Entomology

Figure 2: Reproductive behaviour in mecopterans.
(Top) With his hind tarsi a male mecopteran, Harpobittacus australis, captures an insect. Suspending himself he everts two abdominal sacs that release a pheromone that "calls" females. The female, who approaches upwind, is presented with the prey, and copulation follows. (Bottom) Protruding sacs as they appear in magnified view of male abdomen.

to appear identical to a vertebrate predator actually possess an unbreakable code by which each species is able to distinguish its own kind.

A reproductive behaviour that is usually misunderstood by those who have observed it is the copulation process in dragonflies. The actual copulatory organ of the male is located close to the thorax, not, as in most insects, near the tip of the abdomen. After a male alights on a plant and transfers sperm from the terminal genital opening to the copulatory organ, he seeks out a female and grasps her behind the head with claspers on his abdomen. Although the two fly in a tandem position, actual copulation occurs only when they alight, and the female bends her abdomen to receive the sperm from the male's organ. Colour, pattern, and movement are important in species recognition. In experiments, it has been found that artificial models acceptable to male *Platycnemis* dragonflies must consist of a female head, thorax, and one wing; the model also must be moved from side to side about once every four seconds to be effective. Complete aerial mating in insects is rare, but it does occur in mayflies, houseflies, ants, wasps, and bees.

Sound displays

Among the cicadas, crickets, and some grasshoppers, females normally mate after they have been attracted to a male by vocalizations of the latter, which, in most cases, are species specific. It has been demonstrated that deafened female grasshoppers do not permit copulation. In many crickets, the specific stridulations (noises) that occur after each copulation keep the female near the male until he is ready to produce another spermatophore. These stridulations also prevent the female from removing the spermatophores before insemination has been completed.

Even some butterflies incorporate sounds into their reproductive displays; in some manner, the butterfly Ageronia makes a loud cracking sound when engaged in courtship. Many other insects may incorporate sound into their reproductive displays, perhaps utilizing sounds beyond the sensitivity of the human ear.

Research has revealed that olfactory displays are widespread in insects. The sex attractants for this purpose are usually volatile pheromones. Among certain species of butterflies, such as the queen butterfly (Danaus *gilippus*), the males possess "hair pencils" that project from the end of the abdomen and emit a scent when swept over the female's antennae during courtship behaviour. Copulation does not occur in the absence of this chemical display.

During some stage of their development, a number of insects are either external or internal parasites on a wide variety of animals, including other insects. A particularly bizarre pattern is found in the stylopids, which belong to the order Strepsiptera. Though seldom seen, these insects may be common internal parasites of wasps and bees. The abdomen of the adult females, which never leave their hosts, consists of a bag of eggs that is concealed in the host. The forepart of the parasite, which projects from between abdominal segments of the host, is usually concealed by the host's wings. The females of one stylopid group are apparently unique among animals in having two genital openings—both in the head—in the form of membranous windows. The larvae emerge through these openings, crawl onto a plant, and seek another host. When the host molts its old cuticle (hard skin), the larvae penetrate the soft body. Females extend their heads through the host's abdomen and mature within the host. The males, however, leave the host, pupate in the host's cast-off cuticle, and emerge several days later as adults. The male stylopid then seeks out a host insect and taps it on the side of the abdomen. If no female is present, the male leaves; if a female is present, she somehow signals her presence. The male then inserts his abdomen under the host's wing and enters the genital window of the female.

It is in the orders Isoptera (termites) and Hymenoptera (bees, wasps, and ants), however, that the reproductive behaviour of insects attains its highest level of sophistication. Although dung beetles and some other insect species brood their eggs and care for the young, extreme insect sociality, with its peculiar brood-care system, is found only among the isopterans and the hymenopterans. The principal criterion for such behaviour would appear to be that the female must remain with her brood until after they begin to hatch. Although the phenomenon has been intensively studied, the explanation for the evolution of extreme brood care in ants, many wasps and bees, and termites remains one of the more challenging problems in biology.

Most colonies of social insects reproduce in two ways: either sexual individuals are produced that mate and start new colonies, or the colony breaks up after reaching a certain size. Some species reproduce in both ways. In the first case, the chances of finding new sites are maximized by providing as many individuals of different sexes as possible, each equipped with appropriate guidance mechanisms. In the second, members of the parent colony explore the environment and establish a new colony where suitable.

Another example of reproduction in social insects is that practiced by many ants. Most larvae in an ant colony develop into wingless, sterile workers. Some, however, may get more food (a point that is controversial) and grow more rapidly. These do not pupate when the other larvae do; instead, they become king-sized individuals that eventually metamorphose into sexually mature males or females with wings. Their sex, like that of the wasps and bees, depends upon whether or not the egg was fertilized by the queen.

Mating swarms

The winged sexual forms, or alates, are produced at certain times during the year and swarm in mating flights to establish a new colony, which may actually be no more than a few hundred feet from the old colony. Actual copulation may occur either during flight or after landing on a surface. For most species of ants, it is not known whether a male will copulate with more than one female or if a female will copulate with more than one male. After copulation, the female seeks a location for a new nest and loses her wings within three to five days. Generally, two months are required to rear the first daughter workers. Some females carry a live mealybug with them on the mating flight and take it to the new colony site, where the mealybug's offspring provide the honeydew to feed the ant's initial offspring. Generally, however, the female ant does not provide food for her first offspring; instead, the larvae eat many of the first 100 or so eggs. This egg cannibalism decreases when there are sufficient workers to feed the larvae.

REPRODUCTIVE BEHAVIOUR IN VERTEBRATES

Fishes. The reproductive behaviour of fishes is remarkably diversified: they may be oviparous (lay eggs), ovoviparous (retain the eggs in the body until they hatch), or viviparous (have a direct tissue connection with the developing embryos and give birth to live young). All cartilaginous fishes—the elasmobranchs (*e.g.*, sharks, rays, and skates)--employ internal fertilization and usually lay large, heavy-shelled eggs or give birth to live young. The most characteristic features of the more primitive bony fishes is the assemblage of polyandrous (many males) breeding aggregations in open water and the absence of parental care for the eggs. Many of the species in this group, such as herrings, make what appear to be completely chaotic migrations to their breeding areas. Actually, however, each of these huge spawning aggregations is made up of small, coordinated parties consisting of one female and one or more males. On the other hand, a number of fishes are monogamous, form pairs, and care for the eggs or young. In courtship behaviour, in which they utilize all potential stimuli including sound, chemical, and electrical stimuli, the range and complexity of their displays are not exceeded by any other vertebrate group.

Hermaphroditism in fishes

Although the sexes are usually separate, hermaphroditism is much more common among the bony fishes than in any other group of vertebrates. The reasons for this condition are both physiological and ecological. Whereas the developing gonads of all other vertebrates have an outer and inner layer of tissue, those of bony fishes have a simple origin that lacks any male or female elements. In terms of the evolutionary process, this type of development is likely to be more adaptable to pressures that favour hermaphroditism. When, because of one or several interacting factors, a population density reaches a low point in some species, reproduction may be limited to a low probability of contact with another sexually active individual. In such situations (*e.g.*, very deep sea habitats, tide or stream pools) the evolution of even temporary self-fertilizing hermaphrodites would have the greatest advantage.

From W. Wickler, *Mimicry in Plants and Animals* copyright (1968); Weidenfeld and Nicolson Co . Ltd.

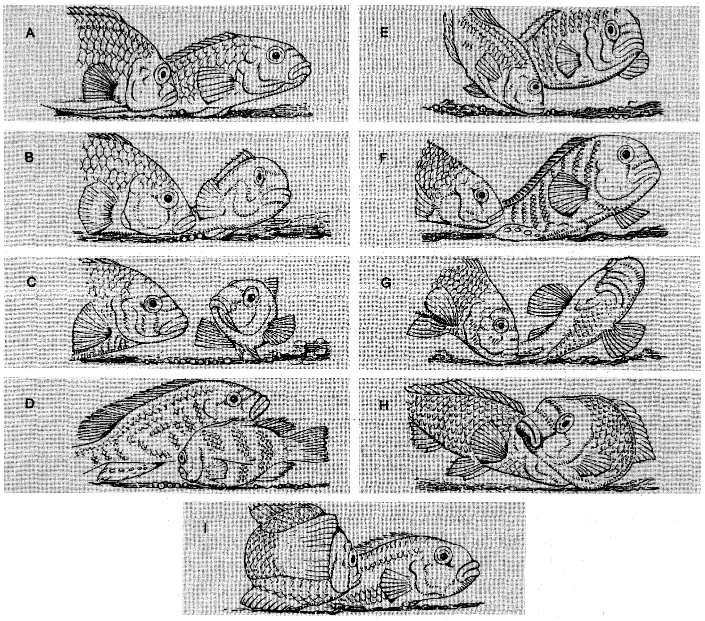


Figure 3: Phases in the spawning sequence of Haplochromis burtoni.
(A–C) Female on right deposits batch of eggs; (D,E) female collects eggs in her mouth, where she broods them; (F–H) male emits sperm onto bare surface while female attempts to take up dummy eggs (conspicuous spots near the base of the anal fin of the male) into her mouth; semen enters her mouth in the process and fertilizes the eggs; (I) female deposits another batch of eggs.

One form of hermaphroditism fairly common in bony fishes is the protogynous type, in which the individual functions first as a female and later as a male; it is much more frequent than the reverse situation (protandrous hermaphroditism). The selective reasons for the predominance of the former are presumably associated with the relationship between smaller body size in females and the greater energy requirements needed to produce eggs. In addition, in some promiscuous mating systems, it may be selectively advantageous to be a male when the body size is large and the individual experienced, rather than small and young. Most sea basses, parrot fishes, and wrasses have this sort of hermaphroditism.

Amphibians. Although true viviparity has been described in the African frog Nectophrynoides, most amphibians lay eggs. Some salamanders, however, retain the eggs within their body and give birth to live young. Courtship displays in frogs are almost entirely vocal, although in salamanders they may involve tactile, visual, and chemical stimuli. In the European newt *Triturus,* for example, in which mating takes place in the water, the male places himself in front of a female with his back to her. Suddenly, he executes a leap, directs a current of water at her, faces her, and bends his tail forward alongside his body; by waving his tail, he sends toward her a gentle current of water that probably carries a chemical stimulant. If the female responds by approaching the male, he turns and faces away, whereupon she touches his tail and he deposits a spermatophore, which she takes into her cloaca, a common passageway into which waste products and reproductive cells are discharged.

Lack of
brood
care

Most frogs and salamanders do not show brood care, but there are exceptions. In the European midwife toad the male rather than the female carries the sticky eggs on its hindlimbs. In a number of Neotropical frogs, the male carries the eggs under a flap of skin on its back. In some species, the young (tadpoles) cling to the back of the male by using their sucker-like mouths.

Reptiles. Reptiles are the first vertebrates that, in an evolutionary sense, have evolved an egg that is truly independent of water. Indeed, many snakes and lizards have even gone beyond this stage and have attained complete viviparity. It is difficult to generalize about reproductive behaviour in the reptiles because the various groups differ from each other in the sensitivity of their receptor organs. In many turtles, for example, the males are territorial and are very aggressive during the breeding period. Courtship behaviour involves mainly tactile stimuli, but olfactory clues are also important. It has been recorded that the wood turtle (*Clemmys*) actually emits a low whistle during courtship. Turtles usually bury their eggs and do not brood them.

Lizards appear to use almost every sensory mechanism in their reproductive activities. The nocturnal geckos employ vocalizations, in addition to tactile and olfactory stimuli. Skinks such as Eumeces rely heavily on olfactory clues. Lizards of the large family Iguanidae, on the other hand, are almost entirely diurnal creatures and utilize, in the main, visual displays, some of which are the equal in complexity to any known among the vertebrates. Many, such as the anoles, are equipped with a throat flap (dewlap) that is often brightly coloured and specifically marked; it is utilized both in courtship and territorial defense. The skinks and a number of other lizards are known to guard their eggs.

In general, the reproductive behaviour of snakes is not well known. The tongue is apparently an important sense organ for receiving olfactory and other chemical stimuli. The males of some snakes have characteristic skin papillae (nipple-like projections) on the throat; the fact that they rub the papillae over the female's body suggests that tactile stimuli are also important to reproduction. In boas, the rudimentary pelvic bones serve as "claws" for lifting the hind end of the female and for producing a vibration that is said to be important in the process of copulation. Some snakes, the pythons in particular, incubate and guard their eggs.

The bellowing roars of male alligators serve to establish breeding territories and apparently also to attract the females. Female crocodiles remain in the vicinity of their nest and will defend it vigorously.

Birds. Although all birds lay eggs, it is curious that they do so, because the time of highest mortality in most birds usually occurs during the egg-laying period. Apparently, birds lack some adaptation that would permit them to become viviparous.

Number
of eggs
laid

Most birds build a nest and incubate their eggs, but the incubator birds and such brood parasites as cuckoos are among the exceptions to this rule. Many females that lay a fixed number of eggs are referred to as determinant layers. The pigeons and doves are outstanding examples of this behaviour; for some as yet unknown reason, they never lay more than one or two eggs. Other species are often referred to as indeterminate layers because, in the absence of a suitable stimulus, they continue to produce eggs. More often than not, this stimulus is the presence in the nest of a certain number of eggs. Such behaviour is clearly adaptive—if eggs are lost for some reason and if other environmental stimuli are present, the missing eggs are replaced. The distinction between determinate and indeterminate layers is often blurred, for many indeterminate layers will not replace more than one or two missing eggs.

The duration of egg incubation varies from as little as nine days in some tropical perching birds to as long as 80 days in some albatrosses. In most species that form pairs, both individuals incubate and feed the young, but the female usually has the greater share of the burden. Among the exceptions to this behaviour pattern are the tinamou (partridge-like game birds), ostriches, some gallinaceous species (*e.g.*, pheasant, grouse, turkeys), and phalaropes. In the phalaropes, the role of the sexes is largely reversed: the females are more brightly coloured than the males and, not surprisingly, are the aggressive ones in courtship and in territorial defense; incubation is carried out solely by the male, but the female aids in feeding the young.

Because many birds begin incubation with the laying of the first egg in the clutch, the eggs hatch at different times. This strategy is often employed by species whose food supply for the young may vary in abundance over a fairly short period. Hence, should food suddenly become scarce, only the smallest chick or chicks will starve rather than the entire clutch. Species in which the young hatch in a relatively well developed, almost independent state tend to have very large clutches, as in many gallinaceous birds. In this case, it might be said that the ultimate size of the clutch is regulated by the abundance and quality of the food available to the female as she produces eggs. The same explanation also accounts for clutch size in parasitic birds—*i.e.*, those that lay eggs in the nests of other species. The breeding densities of birds vary from one pair in many square miles, as in some birds of prey, to such species as the fulmar, which forms colonies numbering as many as 250,000. Some colonies of the African weaverbird (Quelea) have been estimated to exceed 1,000,000 individuals.

One interesting aspect of reproductive behaviour in birds, possibly peculiar to them and to some mammals, is that many courtship displays are learned, or at least perfected through practice, from the parents. An example is the learning of birdsongs. It has been shown in some cases that when chicks are switched from the nest of one species to that of another, they learn some and perhaps all of the songs of the foster parents and do not develop their own species' vocalizations. When mature, such birds often prefer to choose as mates individuals of the same species as their foster parents' rather than those of their own species.

Courtship stimuli in birds are mostly visual and auditory, but it is possible that odour may he important in some petrels and shearwaters. As previously mentioned, most birds form pairs. In these and in many that do not, the males engage in communal, or lek-type, displays on a common courtship ground, such as the familiar strutting grounds of turkeys and many grouse. In addition, there are the incredibly bizarre communal dances of the birds of paradise; the jungle-floor dancing of the cock of the rock; the pasture display grounds of the shorebird, the ruff; and the forest arenas cleared for displaying purposes by the tiny manakins. Many of these display areas are used for many years; in some manakins, for example, certain cleared arenas have existed continuously for at least 30 years. In most lek species, the males are usually brightly coloured, and the females are rather dull in appearance. An exception occurs in some hummingbirds, the so-called hermits, in which both sexes are rather dull in coloration and in which the males group together in singing assemblies.

From J. Sparks. *Bird Behaviour*, Copyright © 1969 by Hamlyn Publishing Group, Ltd. Copyright © 1970 by Grosset & Dunlap, Inc.

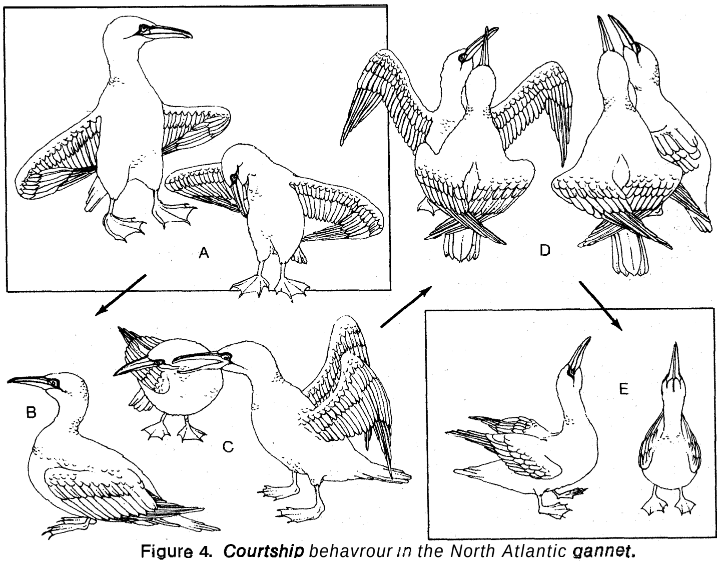


Figure 4. *Courtship behavrour in the North Atlantic gannet.* (A) *Bowing;* (B) *male advertising;* (C) *facing away;* (D) *mutual fencing;* (E) *sky pointing.*

Mammals. Most mammals give birth to live young. The outstanding exceptions are the egg-laying monotremes of Australia, the platypus (*Ornithorhynchus*) and the echidnas (spiny anteaters). In the duckbill platypus, a brief courtship involving a chase in the water precedes copulation. The two eggs that are produced are placed in a burrow and hatch in eight to ten days. In the reproductive behaviour of the spiny anteater (*Tachyglossus*), the female apparently lays her single egg directly into her pouch.

Estrous cycle

As already mentioned, another general aspect of reproductive behaviour in mammals is the estrous cycle, knowledge of which is essential to an understanding of the mechanisms involved in the reproduction of any mammalian species. Females are usually responsive to males only during that portion of the estrous cycle when they are in heat; that is, when one or more eggs have broken out of the ovary and are in the process of descending to the uterus. The factors causing this event vary greatly, hut in some such as rabbits and cats, copulation itself is the main stimulus. In general, however, those mammals, particularly the large ones, that live in temperate areas—*e.g.*, bears, dogs, wolves, foxes, seals, and some deer and antelopes—have one estrous cycle per year. Mammals that live in warmer zones, such as some areas of the tropics, tend to have more than one estrous cycle per year. The sexual cycle in males, the height of which in some forms is referred to as the rut, is, not surprisingly, usually correlated with that of the females. The males of many species of domestic mammals, however, seem to be capable of copulating at almost any time of the year.

Another general aspect of mammalian reproductive behaviour is that they do not normally form pairs. Exceptions occur in certain carnivores and in some primates, in which parental care is divided between the sexes. As in many insects, the courtship behaviour of most mammals does not appear to be elaborate; but, just as in the former group, most mammals (humans are an exception) have an acute sense of smell. It is possible, therefore, that many of the chemical attractants wafted into the air by receptive females are actually courtship displays that are more complex than has been realized. This is not to say, of course, that visual, auditory, and tactile displays do not occur. Many deer and antelopes, for example, have rather complex ritualized visual displays employing such movements as strutting and arching of the heads, as well as conspicuous colour patterns. Males in many species discharge urine on females as a preliminary to copulation. Tactile and auditory displays have been shown to be important in aquatic mammals, such as porpoises and whales.

Additional effects of odours

In addition to a number of mammalian pheromones, other odour effects occur in mammals that, aside from their simple advertising value, have an important influence on reproductive behaviour. It has been shown that, when a recently impregnated female mouse is exposed to the odour of a male other than the one with which she has mated, implantation of the egg in the uterus often fails; as a result, there is a rapid return to estrus. The odour of a strange male may signify to a female rodent an unfavourable situation in which to raise young, inasmuch as a number of male rodents attempt to attack offspring not their own. Although it is not yet certain, there might he an adaptive explanation for this behaviour. The population fluctuations of rodents have attracted much attention, and, perhaps correctly, studies have focussed on the ecological parameters of these fluctuations; for example, it has been demonstrated in the laboratory that certain behavioral mechanisms involving odours exercise profound control over the reproduction and population levels of rodents. It has also been shown that the odour of mice can stimulate the production of hormones that cause a decrease in the reproductive capacity of other mice. In another study, estrus was suppressed, and many pseudopregnancies developed when four or more female mice were grouped together in the absence of a male. These results offer a partial explanation for the well-known phenomenon of reduction of population growth in rodent colonies that have high population densities.

THE EVOLUTION OF REPRODUCTIVE BEHAVIOUR

There is a popular tendency to think of primitive animals (in a phylogenetic or descent sense) as lacking "elaboration"; *i.e.,* that the animals of earlier geological periods had simpler displays or perhaps lacked crests or pheromones or elaborate communal displays in comparison with the reproductive systems of their present-day counterparts. There is no a priori reason for this belief. The fossil record indicates that the societies of which these animals were a part were as diverse and complex as those in which their relatives now live; certainly their display repertoires should have been equally complete. This is not to say, however, that the primitive forms of reproductive behaviour used the same displays for courtship as do the modem forms.

Displays. It has been pointed out that, in general, animals have relatively few displays; in addition, it has been deduced that the relative stability of displays is a dynamic equilibrium—that is, new ones are gained and old ones are lost at about the same frequency. Displays are lost when they no longer convey a selective advantage to the individual animals using them; that is, when the displays are no longer effective in promoting the behaviour that seeks to maximize gene survival in the next generation.

New displays, on the other hand, generally arise by ritualization of previously existing behaviours or functions; that is, when a selective advantage accrues to those individuals who, to convey information, use certain behaviours or functions in a manner that is either partly or totally different from their original purpose. Pheromones, for example, are usually derived from compounds that are natural breakdown products of body metabolism, such as the compounds in urine. Thus, urine, as the precursor of these chemical sex attractants in mammals, functions for display purposes, which is far removed from its basic excretory function.

Darwin's theory of sexual selection

Darwin proposed a theory of sexual selection to account for the presence in animals of displays and functions that apparently were not related to survival. He pointed out that two general concepts were involved. First, the evolution of such characteristics as the larger size of males in many species and the development of horns and antlers in mammals could be accounted for by their usefulness in fights between males for the sexual possession of females. This concept has been termed intrasexual selection. For such colourful male structures as the plumes of birds of paradise and the tails of peacocks, Darwin suggested that they resulted from the cumulative effects of sexual preference exerted by the females of the species at the time of mating. This second concept has been termed epigamic selection.

A displaying male has been known to convey information about his relative fitness; that is, his ability, with respect to other displaying males, to maximize the survival of his genes into the next generation. Both the brightness of his coloration and the frequency with which he struts say something about the effectiveness of his genes to produce a "healthy" individual. Once this correlation takes place, selection favours those females who perceive the differences between males and choose the "most fit" ones. Correspondingly, sexual selection intensifies the signals up to the point at which any further elaboration of those signals would result in a loss of fitness. When selection goes beyond this point, the male, because of his elaborate ornamentation and other displays, is more likely to suffer from predation before he has the opportunity to reproduce.

Sexual selection. The discussion concerning courtship displays leads naturally to the concept of sexual selection. Why do the males of some species possess elaborate displays? Why, in fact, do some species "elect" to utilize one mating system, say a monogamous one, while others "choose" a polygamous one? It has been suggested that many courtship displays and mating systems, particularly those involving polygamous systems with communal displays in a common courtship area, have an epideictic function—that is, they provide information as to the number of like individuals in a locality. The animals then act according to the information received, often by reducing their reproductive output. Because this concept implies that natural selection is acting for the good of the species rather than for the good of the individual, it has been called group selection. This concept has provoked considerable controversy for two reasons: first, there is no known mechanism by which group selection can function; second, as mentioned earlier, the pertinent behaviours involved can be more simply explained in terms of Darwinian selection dealing with individuals rather than groups.

Polygynous mating systems

In a number of polygynous (mating of one male with more than one female) and promiscuous species, adult females outnumber adult males, sometimes by a factor of five or more. It has been erroneously suggested that this sexual imbalance is the cause of the polygynous mating system, in which one male has several female partners. It has been demonstrated, however, in all polygynous species so far studied, that the ratio of males to females is 50:50 at the time of birth; in many cases, this ratio persists until the cessation of parental care. Therefore, it is the polygynous relationship that causes the imbalance, not vice versa: because sexual selection is the dominant factor in a polygynous and promiscuous species, it results in a greater (but not intolerably greater) mortality of males than of females.

Because one male can impregnate many females, thus lowering the selective value of an individual male, females are more valuable than males in an evolutionary sense. It can be seen, therefore, that sexual selection always favours a polygynous and promiscuous system unless it is disadvantageous to the females, as is the case with most birds. In most mammals, however, polygyny is the dominant mating system because the male is not needed for parental care. Therefore, monogamy is favoured over polygamy only when some environmental resource (food, for example) is limited and when the maximum survival of the young requires that they receive the care of both parents. As in all other aspects of reproductive behaviour, the type of mating system that is employed by a given species is the result of natural selection.

BIBLIOGRAPHY. MARGARET BASTOCK, ***Courtship: An Ethological Study*** (1967), an excellent survey; DESMOND MORRIS, ***Patterns of Reproductive Behaviour*** (1970), a compilation of some classical papers, all by Morris; N. TINBERGEN and the EDITORS OF LIFE, ***Animal Behavior*** (1965), a good introduction; ARI VAN TIENHOVEN, ***Reproductive Physiology of Vertebrates*** **(1968), *a*** good survey with an emphasis on the hormonal and neurophysiological aspects of reproductive behaviour, especially ch. 9, *11, 13,* and ***14;*** S.A. ASDELL, ***Patterns of Mammalian Reproduction*** (1964), a comprehensive survey stressing anatomical and physiological aspects; C.M. BREDER and D.E. ROSEN, ***Modes of Reproduction in Fishes*** (1966), the best modern survey of reproductive behaviour in fishes; E.O. WILSON, ***The Insect Societies*** (1971), the best general treatment of this group; JOHN SPARKS, ***Bird Behaviour*** (1969), an excellent introduction to the reproductive behaviour of birds, with excellent illustration.

(N.G.S.)

Reproductive System, Human

Man belongs to that group of mammals characterized by the bearing of live offspring that have attained considerable development within the uterus, or womb. Provided all organs are: present, normally constructed, and functioning properly, the essential features of human reproduction are (1) liberation of an egg from the ovary at the right time in the reproductive cycle; (2) internal fertilization by spermatozoa (sperm, or male sex cells) of the ovum in the uterine tube; (3) transport of the fertilized ovum along the uterine tube to the uterus; (4) implantation of the blastocyst, the early embryo that develops from the fertilized ovum, in the wall of the uterus; (5) formation of a placenta and maintenance of the intra-uterine existence of the unborn child; *(6)* birth of the child and expulsion of the placenta; and (7) suckling and care of the child, with an eventual return of the maternal organs virtually to their original state.

For this biological process to be carried out, certain organs and structures are required in both male and female bodies. The source of the ova, the female germ cells, is the female gonad or ovary; that of spermatozoa is the testis. In human beings, the two ovaries are situated in the pelvic cavity, and the two testes are enveloped in a sac of skin, the scrotum, lying below and outside the abdomen. Besides producing the germ cells, or gametes, the ovaries and testes are the source of hormones that cause full development of secondary sexual characteristics and also the proper functioning of the genital (reproductive) tracts. These tracts comprise the uterine tube, the uterus, vagina, and associated structures in females, and, in males, the penis, the sperm channels—epididymis, ductus deferens, and ejaculatory ducts—and other related structures and glands. The function of the uterine tube is to convey an ovum, which is fertilized in the tube, to the uterus, where gestation (development before birth) takes place. The function of the male ducts is to convey spermatozoa from the testis, to store them, and, when ejaculation occurs, to eject them with secretions from the male glands through the penis.

At copulation the erect penis is inserted into the vagina and spermatozoa contained in the seminal fluid are ejaculated into the female genital tract. Spermatozoa then pass from the vagina through the uterus to the uterine tube to fertilize the ovum in the outer part of the tube. Human females exhibit a periodicity in the activity of their ovaries and uterus, which starts at puberty and ends at the menopause. The periodicity is manifested by menstruation at intervals of about 28 days; important changes occur in the ovaries and uterus during each reproductive or menstrual cycle. Periodicity is suppressed during pregnancy and lactation.

This article describes the component parts of the reproductive system in humans, with an indication of the functions of the various organs. Some comparisons are made with arrangements in other mammals, and, where helpful to understand relationships, a brief account of the embryology of a particular region is given. Endocrine and other physiological factors in human reproduction are mentioned briefly.

THE SEXES: MALE AND FEMALE

The sex of a human child is determined at the time of fertilization of the ovum by the spermatozoon. The differences between a man and a woman are genetically determined by the chromosomes that each possesses in the nuclei of the cells. This stage in the development of the individual is detailed in the article EMBRYOLOGY, HUMAN.

Development of sexual differences in embryo

Once the genetic sex has been determined there normally follows a succession of changes that will result, finally, in the development of an adult male or female. There is, however, no external indication of the sex of a human embryo during the first eight weeks of its life within the uterus. This is a neutral or indifferent stage during which the sex of an embryo can be ascertained only by examination of the chromosomes in its cells. The next phase, one of differentiation, begins first in gonads that are to become testes, and a week or so later in those destined to be ovaries. Embryos of the two sexes are initially alike in possessing similar duct systems linking the undifferentiated gonads with the exterior and in having similar external genitalia, represented by three simple protuberances. The embryos each have four ducts, the subsequent fate of which is of great significance in the eventual anatomical differences between men and women. Two ducts closely related to the developing urinary system are called mesonephric, or wolffian, ducts. In males, each mesonephric duct becomes differentiated into four related structures: a duct of the epididymis, a ductus deferens, an ejaculatory duct, and a seminal vesicle (see below). In females, the mesonephric ducts are largely suppressed. The other two ducts, called the paramesonephric or müllerian ducts, persist, in females, to develop into the uterine tubes, the uterus, and part of the vagina; in males they are largely suppressed. Differentiation also occurs in the primitive external genitalia, which in males become the penis and scrotum, and in females the clitoris and labia.

At birth the organs appropriate to each sex have developed and are in their adult positions but are not functioning. Various abnormalities can occur during development of sex organs in human embryos, leading to hermaphroditism, pseudohermaphroditism, and other chromosomally induced conditions (see BIRTH DEFECTS AND CONGENITAL DISORDERS). During childhood until puberty there is steady growth in all reproductive organs and a gradual development of activity. Puberty marks the onset of increased activity in the sex glands and the steady development of secondary sexual characteristics. Puberty is brought about primarily by increased output of hormones from the anterior lobe of the pituitary gland at the base of the brain, but other endocrine organs are also involved.

Changes at puberty

In males at puberty the testes enlarge and become active, the external genitalia enlarge, and the capacity to ejaculate develops. Marked changes in height and weight occur as hormonal secretion from the testes increases. The larynx, or voice box, enlarges, with resultant deepening of the voice. Certain features in the skeleton, as seen in the pelvic bones and skull, become accentuated. The hair in the armpit and the pubic hair becomes abundant and thicker. A beard, a moustache, and cheek hair develop, as well as hair on the chest, abdomen, and limbs. Hair at the temple recedes. Skin glands become more active, especially apocrine glands (a type of sweat gland that is found in the armpit and groin and around the anus).

These secondary sex characteristics do not develop in individuals castrated before puberty, but the administration of androgens (male sex hormones) to such persons and to males having poorly developed testes can correct, in large measure, some of the poorly developed secondary characteristics. Large amounts of androgen, however, by preventing production of the hormone gonadotrophin by the pituitary, suppress testicular activity, thus depressing formation and release of sperm. Some derivatives of the male sex hormone testosterone can promote general bodily development.

In females at puberty, the external genitalia enlarge and the uterus commences its periodic activity with overt menstruation. The mammary glands develop, and there is a deposition of body fat in accordance with the usual contours of the mature female. Growth of axillary (armpit) and pubic hair is more abundant, and the hair becomes thicker. In a female receiving androgens, the typical male secondary sex characteristics may develop, menstruation may be suppressed, and the mammary glands may atrophy.

MALE REPRODUCTIVE SYSTEM

The male gonads are the testes; they are the source of spermatozoa and also of male sex hormones called androgens. The other genital organs are the epididymides, the ductus or vasa deferentia, the seminal vesicles, the ejaculatory ducts, and the penis, as well as certain accessory structures, the prostate and the bulbourethral (Cowper's) glands. The principal functions of these structures are to transport the spermatozoa from the testes to the exterior, to allow their maturation on the way, and to provide certain secretions that help form the seminal fluid.

The penis. The penis, the male organ of copulation, is partly inside and partly outside the body. The inner part, attached to the bony margins of the pubic arch (that part of the pelvis directly in front and at the base of the trunk), is called the root of the penis. The second, or outer, portion is free, pendulous, and enveloped all over

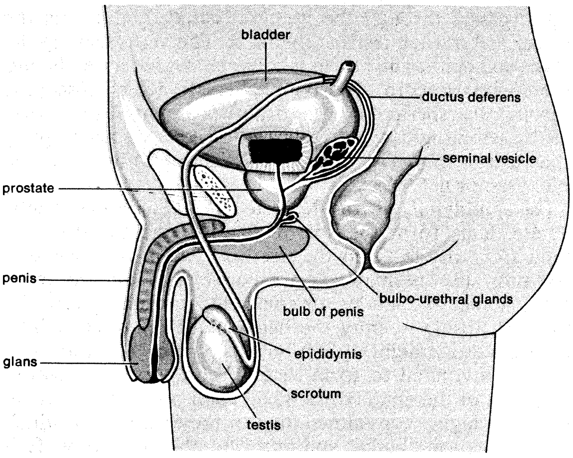


Figure 1: Male reproductive organs.
Adapted from H. Gray. *Anatomy of the Human Body, 28th ed, by* C.M. Goss (1966); Lea & Febiger

in skin; it is termed the body of the penis. The organ is composed chiefly of cavernous or erectile tissue that becomes engorged with blood to produce considerable enlargement and erection. The penis is traversed by a tube, the urethra, which serves as a passage both for urine and for semen.

The corpus spongiosum and the corpora cavernosa

The body of the penis, sometimes referred to as the shaft, is cylindrical in shape when flaccid but when erect is somewhat triangular in cross section, with the angles rounded. This condition arises because the right and left corpora cavernosa penis, the masses of erectile tissue, lie close together in the dorsal part of the penis, while a single body, the corpus spongiosum penis, which contains the urethra, lies in a midline groove on the under surface of the corpora cavernosa. The dorsal surface of the penis is that which faces upward and backward during erection.

The slender corpus spongiosum reaches beyond the extremities of the erectile corpora cavernosa, and at its outer end is enlarged considerably to form a soft, conical, sensitive structure called the glans penis. The base of the glans has a projecting margin, the corona, and the groove where the corona overhangs the corpora cavernosa is referred to as the neck of the penis. The glans is traversed by the urethra, which ends in a vertical, slitlike, external opening. The skin over the penis is thin and loosely adherent and at the neck is folded forward over the glans for a variable distance to form the prepuce or foreskin. A median fold, the frenulum of the prepuce, passes to the under surface of the glans to reach a point just behind the urethral opening. The prepuce can usually be readily drawn back to expose the glans.

The root of the penis comprises two crura, or projections, and the bulb of the penis. The crura and the bulb are attached respectively to the edges of the pubic arch and to the perineal membrane (the fibrous membrane that forms a floor of the trunk). Each crus is an elongated structure covered by the ischiocavernosus muscle, and each extends forward, converging toward the other, to become continuous with one of the corpora cavernosa. The oval bulb of the penis lies between the two crura and is covered by the bulbospongiosus muscle. It is continuous with the corpus spongiosum. The urethra enters it on the flattened deep aspect that lies against the perineal membrane, traverses its substances, and continues into the corpus spongiosum.

The two corpora cavernosa are close to one another, separated only by a partition in the fibrous sheath that encloses them. The erectile tissue of the corpora is divided by numerous small fibrous bands into many caverncus spaces, relatively empty when the penis is flaccid but engorged with blood during erection. The structure of the tissue of the corpus spongiosum is similar to that of the corpora cavernosa, but there is more smooth muscle and elastic tissue. A deep fascia, or sheet of connective tissue, surrounding the structures in the body of the penis, is prolonged to form the suspensory ligament, which anchors the penis to the pelvic bones at the mid point of the pubic arch.

The penis has a rich blood supply from the internal pudendal artery, a branch of the internal iliac artery, which supplies blood to the pelvic structures and organs, the buttocks, and the inside of the thighs. Erection is brought about by distension of the cavernous spaces with blood, which is prevented from draining away by compression of the veins in the area.

The penis is amply supplied with sensory and with autonomic (involuntary) nerves. Of the autonomic nerve fibres the sympathetic fibres cause constriction of blood vessels, and the parasympathetic fibres cause their dilation. It is usually stated that ejaculation is brought about by the sympathetic system, which at the same time inhibits the desire to urinate and also prevents the seminal fluid from entering the bladder.

Septum and raphe

The scrotum. The scrotum is a pouch of skin lying below the pubic symphysis and just in front of the upper parts of the thighs. It contains the testes and lowest parts of the spermatic cord. **A** scrotal septum or partition divides the pouch into two compartments and arises from a ridge, or raphe, visible on the outside of the scrotum. The raphe turns forward onto the under surface of the penis and is continued back onto the perineum (the area between the legs and as far back as the anus). This arrangement indicates the bilateral origin of the scrotum from two genital swellings that lie one on each side of the base of the phallus, the precursor of the penis or clitoris in the embryo. The swellings are also referred to as the labioscrotal swellings because in females they remain separate to form the labia majora, while in males they unite to form the scrotum.

The skin of the scrotum is thin, brown, devoid of fatty tissue, and more or less folded and wrinkled. There are some scattered hairs and sebaceous glands on its surface. Below the skin is a layer of involuntary muscle, the dartos, which can alter the appearance of the scrotum. On exposure of the scrotum to cold air or cold water, the dartos contracts and gives the scrotum a shortened, corrugated appearance; warmth causes the scrotum to become smoother, flaccid, and less closely tucked in around the testes. Beneath the dartos muscle are layers of fascia continuous with those forming the coverings of each of the two spermatic cords, which suspend the testes within the scrotum and contain each ductus deferens, the testicular blood and lymph vessels, the artery to the cremaster muscle (which draws the testes upward), the artery to each ductus deferens, the genital branch of the genitofemoral nerve, and the testicular network of nerves.

The scrotum is supplied with blood by the external pudendal branches of the femoral artery, which is the chief

Adapted from H. Gray, *Anatomy of the Human Body*, 28th ed. by C.M. Goss (1966); Lea & Febiger

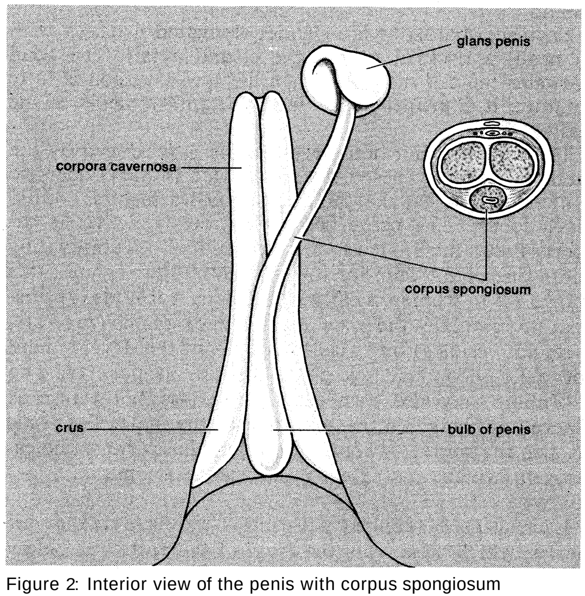


Figure 2: Interior view of the penis with corpus spongiosum detached and turned to one side.

artery of the thigh, and by the scrotal branches of the internal pudendal artery. The veins follow the arteries. The lymphatic drainage is to the lymph nodes in the groin.

The testes. The two testes, which usually complete their descent into the scrotum from their point of origin on the back wall of the abdomen in the seventh month after conception, are suspended in the scrotum by the spermatic cords. Each testis is four to five centimetres (about one to one and one-half inches) long and is enclosed in a fibrous sac, the tunica albuginea. This sac is lined internally by the tunica vasculosa, containing a network of blood vessels, and is covered by the tunica vaginalis, which is a continuation of the membrane that lines the abdomen and pelvis. The tunica albuginea has extensions into each testis that act as partial partitions to divide the testis into approximately 250 compartments, or lobules.

Adapted from H. Gray, *Anatomy of the Human Body*, 28th ed. by C.M. Goss (1966); Lea & Febiger

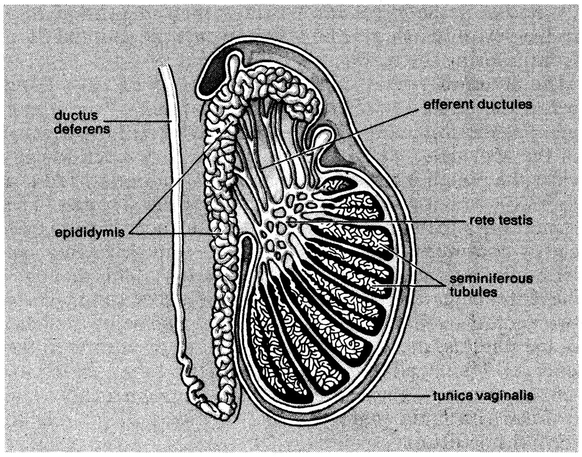


Figure 3: Longitudinal cross section of testis.

Each lobule contains one or more convoluted tubules, or narrow tubes, where sperm are formed. The tubules, if straightened, would extend about 70 centimetres (about 30 inches). The multistage process of sperm formation, which takes about 60 days, goes on in the lining of the tubules, starting with the spermatogonia, or primitive sperm cells, in the outermost layer of the lining. Spermatozoa (sperm) leaving the tubules are not capable of independent motion, but undergo a further maturation process in the ducts of the male reproductive tract; the process may be continued when, after ejaculation, they pass through the female tract. Maturation of the sperm in the female tract is called capacitation; little is known about it.

Each spermatozoon is a slender elongated structure with a head, a neck, a middle piece, and a tail. The head contains the cell nucleus. When the spermatozoon is fully mature, it is propelled by the lashing movements of the tail.

Production of hormones

The male sex hormone testosterone is produced by the cells of Leydig. These cells are located in the connective (interstitial) tissue that holds the tubules together within each lobule. The tissue becomes markedly active at puberty under the influence of the interstitial-cell-stimulating hormone of the anterior lobe of the pituitary gland; this hormone in women is called luteinizing hormone. Testosterone stimulates the male accessory sex glands (prostate, seminal vesicles) and also brings about the development of male secondary sex characteristics at puberty. The hormone may also be necessary to cause maturation of sperm and heighten the sex drive of the male. The testis is also the source of some of the female sex hormone estrogen, which may exert an influence on pituitary activity.

Each testis is supplied with blood by the testicular arteries, which arise from the front of the aorta just below the origin of the renal (kidney) arteries. Each artery crosses the rear abdominal wall, enters the spermatic cord, passes through the inguinal canal, and enters the upper end of each testis at the back. The veins leaving the testis and epididymis form a network, which ascends into the spermatic cord. The lymph vessels, which also pass through the spermatic cord, drain to the lateral and preaortic lymph nodes. Nerve fibres to the testis accompany the vessels; they pass through the renal and aortic nerve plexuses, or networks.

The epididymis, ductus deferens, and ejaculatory ducts. These ducts form the sperm canal. Together they extend from the testis to the urethra, where it lies within the prostate. Spermatozoa are conveyed from the testis along some 20 ductules, or small ducts, which pierce the fibrous capsule to enter the head of the epididymis. The ductules are straight at first but become dilated and then much convoluted to form distinct compartments within the head of the epididymis. They each open into a single duct, the highly convoluted duct of the epididymis, which constitutes the "body" and "tail" of that structure. It is held together by connective tissue but if unravelled would be nearly six metres (20 feet) long. The duct enlarges and becomes thicker walled at the lower end of the tail of the epididymis, where it becomes continuous with the ductus deferens.

The ductules from the testis have a thin muscular coat and a lining that consists of alternating groups of high columnar cells with cilia (hairlike projections) and low cells lacking cilia. The cilia assist in moving spermatozoa toward the epididymis. In the duct of the epididymis the muscle coat is thicker and the lining is thick with tufts of large nonmotile cilia. There is some evidence that the ductules and the first portion of the duct of the epididymis remove excess fluid and extraneous debris from the testicular secretions entering these tubes. The blood supply to the epididymis is by a branch from the testicular artery given off before that vessel reaches the testis.

The ductus deferens is the continuation of the duct of the epididymis. It commences at the lower part of the tail of the epididymis and ascends along the back border of the testis to its upper pole. Then, as part of the spermatic cord, it extends to the deep inguinal ring. Separating from the other elements of the spermatic cord—the blood vessels, nerves, and lymph vessels—at the ring, the ductus deferens makes its way through the pelvis toward the base of the prostate, where it is joined by the seminal vesicle to form the ejaculatory duct. A part of the ductus that is dilated and rather tortuous, near the base of the urinary bladder, is called the ampulla.

The ductus deferens has a thick coat of smooth muscle that gives it a characteristic cordlike feel. The longitudinal muscle fibres are well developed, and peristaltic contractions (contractions in waves) move the spermatozoa toward the ampulla. The mucous membrane lining the interior is in longitudinal folds and is mostly covered with nonciliated columnar cells, although some cells have nonmotile cilia. The ampulla is thinner walled and probably acts as a sperm store.

The prostate, seminal vesicles, and bulbourethral glands. These structures are the male accessory reproductive organs and provide secretions to form the bulk of the seminal fluid of an ejaculate. The prostate is in the lesser or true pelvis, centred behind the lower part of the pubic arch. It lies in front of the rectum. The prostate is shaped roughly like an inverted pyramid; its base is directed upward and is immediately continuous with the neck of the urinary bladder. The urethra traverses its substance. The two ejaculatory ducts (see below) enter the prostate near the upper border of its posterior surface. The prostate is of a firm consistency, surrounded by a capsule of fibrous tissue and smooth muscle. It measures about 4 centimetres across, 3 centimetres in height, and 2 centimetres front to back (about 1.6 by 1.2 by 0.8 inch) and consists of glandular tissue contained in a muscular framework. It is imperfectly divided into three lobes. Two lobes at the side form the main mass and are continuous behind the urethra. In front of the urethra they are connected by an isthmus of fibromuscular tissue devoid of glands. The third, or median, lobe is smaller and variable in size and may lack glandular tissue. There

Structure of prostate

are three clinically significant concentric zones of prostatic glandular tissue about the urethra. A group of short glands that are closest to the urethra and discharge mucus into its channel are subject to simple enlargement. Outside these is a ring of submucosal glands (glands from which the mucosal glands develop), and farther out is a large outer zone of long branched glands, composing the bulk of the glandular tissue. Cancer of the prostate is almost exclusively confined to the outer zone. The glands of the outer zone are lined by tall columnar cells that secrete prostatic fluid under the influence of androgens from the testis. The fluid is thin, milky, and slightly acid.

The seminal vesicles are two structures, about 5 centimetres (2 inches) in length, lying between the rectum and the base of the bladder. Their secretions form the bulk of the seminal fluid. Essentially, each vesicle consists of a much-coiled tube with many diverticula or outpouches that extend from the main tube, the whole being held together by connective tissue. At its lower end the tube is constricted to form a straight duct or tube that joins with the corresponding ductus deferens to form the ejaculatory duct. The vesicles are close together in their lower parts but are separated above where they lie close to the deferent ducts. The seminal vesicles have longitudinal and circular layers of smooth muscle, and their cavities are lined with mucous membrane, which is the source of the secretions of the organs. These secretions are ejected by muscular contraction during ejaculation. The activity of the vesicles is dependent on androgen production by the testes; castration causes atrophy of the seminal vesicles. The secretion is thick, sticky, and yellowish: it contains the sugar fructose and is slightly alkaline.

The bulbourethral glands, often called Cowper's glands, lie on the underside of the urethra between the prostate and the bulk of the penis. They are hardly larger than a pea. Each has a slender duct that runs forward and toward the centre to open on the floor of the spongy portion of the urethra. These glands are poorly developed in man. Their secretion is liberated during sexual excitement and may help to lubricate and coat the urethra to assist passage of the ejaculate.

Ejaculatory ducts

The two ejaculatory ducts lie on each side of the midline and are formed by the union of the duct of the seminal vesicle, which contributes secretions to the seminal fluid, with the end of the ductus deferens at the base of the prostate. Each duct is about 2 centimetres (about 0.8 inch) long and passes between a lateral and the median lobe of the prostate to reach the floor of the prostatic urethra. This part of the urethra has on its floor (or posterior wall) a longitudinal ridge called the urethral crest. On each side is a depression, the prostatic sinus, into which open the prostatic ducts. In the middle of the urethral crest is a small elevation, the colliculus seminalis, on which the opening of the prostatic utricle is found. The prostatic utricle is a short diverticulum or pouch lined by mucous membrane; it may correspond to the vagina or uterus in the female. The small openings of the ejaculatory ducts lie on each side of or just within the opening of the prostatic utricle. The ejaculatory ducts are thin walled and lined by columnar cells.

THE FEMALE REPRODUCTIVE SYSTEM

The female gonads or sexual glands are the ovaries; they are the source of ova and also of the female sex hormones estrogens and progestogens. The uterine tubes conduct ova to the uterus, which lies within the lesser or true pelvis. The uterus connects through the cervical canal with the vagina. The vagina opens into the vestibule about which lie the external genitalia, collectively known as the vulva.

The external genitalia. The female external genitalia include the structures placed about the entrance to the vagina and external to the hymen, the membrane across the entrance to the vagina. They are the mons pubis (also called the mons veneris), the labia majora and minora, the clitoris, the vestibule of the vagina, the bulb of the vestibule, and the greater vestibular glands.

The mons pubis is the rounded eminence, made by fatty tissue beneath the skin, lying in front of the pubic symphysis. A few fine hairs may be present in childhood; later, at puberty, they become coarser and more numerous. The upper limit of the hairy region is horizontal across the lower abdomen.

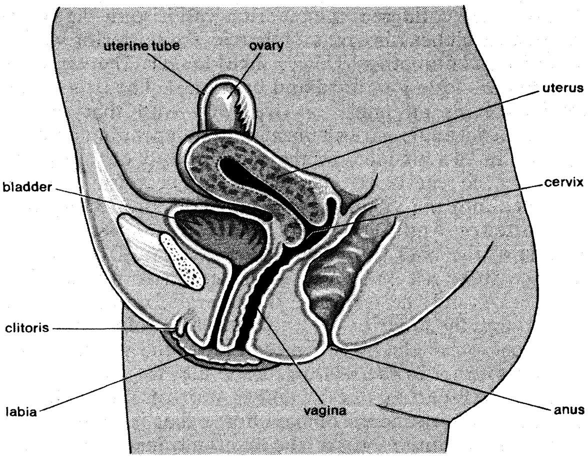


Figure 4: Female reproductive system.
Adapted from H. Gray, *Anatomy of the Human Body*, 28th ed. by C.M. Goss (1966); Lea & Febiger

The labia majora are two marked folds of skin that extend from the mons pubis downward and backward to merge with the skin of the perineum. They form the lateral boundaries of the vulval or pudendal cleft, which receives the openings of the vagina and the urethra. The outer surface of each labium is pigmented and hairy; the inner surface is smooth but possesses sebaceous glands. The labia majora contain fat and loose connective tissue and sweat glands. They correspond to the scrotum in the male and contain tissue resembling the dartos muscle. The round ligament (see below ***The uterus)*** ends in the tissue of the labium. The labia minora are two small folds of skin, lacking fatty tissue, that extend backward on each side of the opening into the vagina. They lie inside the labia majora and are some 4 centimetres (about 1.6 inches) in length. In front, an upper portion of each labium minus passes over the clitoris—the structure, in the female, corresponding to the penis (excluding the urethra) in the male—to form a fold, the prepuce of the clitoris, and a lower portion passes beneath the clitoris to form its frenulum. The two labia minora are joined at the back across the midline by a fold that becomes stretched at childbirth. The labia minora lack hairs but possess sebaceous and sweat glands.

Clitoris and hymen

The clitoris is a small erectile structure composed of two corpora cavernosa separated by a partition. Partially concealed beneath the forward ends of the labia minora, it possesses a sensitive tip of spongy erectile tissue, the glans clitoridis. The external opening of the urethra is some 2.5 centimetres (about 1 inch) behind the clitoris and immediately in front of the vaginal opening.

The vestibule of the vagina is the cleft between the labia minora into which the urethra and vagina open. The hymen vaginae lies at the opening of the vagina: it is a thin fold of mucous membrane that varies in shape. After rupture of the hymen, the small rounded elevations that remain are known as the carunculae hymenales. The bulb of the vestibule, corresponding to the bulb of the penis, is two elongated masses of erectile tissue that lie one on each side of the vaginal opening. At their posterior ends lie the greater vestibular glands, small mucous glands that open by a duct in the groove between the hymen and each labium minus. They correspond to the bulbourethral glands of the male.

The blood supply and nerve supply of the female external genital organs are similar to those supplying corresponding structures in the male.

The vagina. The vagina (the word means "a sheath") is the canal that extends from the cervix (outer end) of the uterus within the lesser pelvis down to the vestibule between the labia minora. The orifice of the vagina is guarded by the hymen. The vagina lies behind the bladder and urethra and in front of the rectum and anal canal.

Its walls are collapsed; the anterior wall is some 7.5 centimetres (3 inches) in length, whereas the posterior wall is about 1.5 centimetres (over ½ inch) longer. The vagina is directed obliquely upward and backward. The axis of the vagina forms an angle of over 90° with that of the uterus. This angle varies considerably, depending on conditions in the bladder, in the rectum, and during pregnancy. The cervix of the uterus projects for a short distance into the vagina and is normally pressed against its posterior wall. There are, therefore, recesses in the vagina at the back, on each side, and at the front of the cervix. These are known as the posterior fornix (behind the cervix and the largest), the lateral fornices (at the sides), and the anterior fornix (at the front of the cervix). The position of the uterus in relation to the vagina is described further in the section on the uterus.

The upper part of the posterior wall of the vagina is covered by peritoneum or membrane that is folded back onto the rectum to form the recto-uterine pouch. The lower part of the posterior vaginal wall is separated from the anal canal by a mass of tissue known as the perineal body.

Mucous membrane and muscle coat of vagina

The vagina has a mucous membrane and an outer smooth muscle coat closely attached to it. The mucous membrane has a longitudinal ridge in the midline of both the anterior and posterior walls. The ridges are known as the columns of the vagina; many rugae or folds extend from them to each side. The furrows between the rugae are more marked on the posterior wall and become especially pronounced before birth of a child. The membrane undergoes little change during the menstrual cycle (except in its content of glycogen, a complex starchlike carbohydrate) ; this is in contradistinction to the situation in many mammals in which marked exfoliation (shedding of the surface cells) can occur. No glands are present in the vaginal lining, and mucus present has been secreted by the glands in the cervical canal of the uterus. The smooth muscle coat consists of an outer longitudinal layer and a less developed inner circular layer. The lower part of the vagina is surrounded by the bulbospongiosus muscle, a striped muscle attached to the perineal body.

The blood supply to the vagina is derived from several adjacent vessels, there being a vaginal artery from the internal iliac artery and also vaginal branches from the uterine, middle rectal, and internal pudendal arteries, all branches of the internal iliac artery. The nerve supply to the lower part of the vagina is from the pudendal nerve and from the inferior hypogastric and uterovaginal plexuses.

The uterus. The uterus, or womb, is shaped like an inverted pear. It is a hollow, muscular organ with thick walls, and it has a glandular lining called the endometrium. The human uterus is normally a single structure, termed unicornuate; in the majority of mammals there are two horns or pouches to the uterus (bicornuate). In an adult virgin the uterus is 7.5 centimetres (3 inches) long, 5 centimetres (2 inches) in breadth, and 2.5 centimetres (1 inch) thick, but it enlarges to four to five times this size in pregnancy. The narrower, lower end is called the cervix; this projects into the vagina. The cervix is made of fibrous connective tissue and is of a firmer consistency than the body of the uterus. The two uterine tubes enter the uterus at opposite sides, near its top. The part of the uterus above the entrances of the tubes is called the fundus; that below is termed the body. The body narrows towards the cervix, and a slight external constriction marks the juncture between the body and the cervix.

Adapted from H. Gray Anatomy *of* the Human *Body*, 28th ed. by C.M. Goss (1966); Lea & Febiger

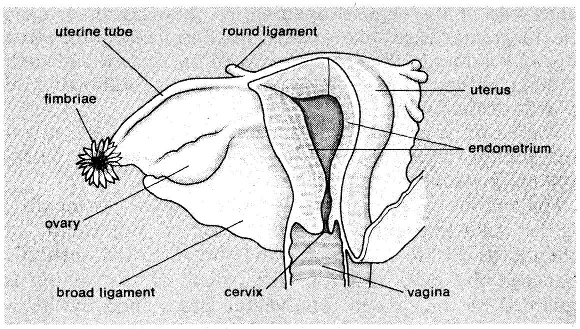


Figure 5: The uterus.

The uterus does not lie in line with the vagina but is usually turned forward (anteverted) to form approximately a right angle with it. The position of the uterus is affected by the amount of distension in the urinary bladder and in the rectum. Enlargement of the uterus in pregnancy causes it to rise up into the abdominal cavity, so that there is closer alignment with the vagina. The nonpregnant uterus also curves gently forward; it is said to be anteflexed. The uterus is supported and held in position by the other pelvic organs, by the muscular floor or diaphragm of the pelvis, by certain fibrous ligaments, and by folds of peritoneum. Among the supporting ligaments are two double-layered broad ligaments, each of which contains a uterine tube along its upper free border and a round ligament, corresponding to the gubernaculum testis of the male, between its layers. Two ligaments—sometimes called Mackenrodt's ligaments—at each side of the cervix are also important in maintaining the position of the uterus.

The cavity of the uterus is remarkably small in comparison with the size of the organ. Except during pregnancy, the cavity is flattened, with front and rear walls touching, and is triangular. The triangle is inverted, with its base at the top, between the openings of the two uterine tubes, and with its apex at the internal os, the opening into the cervix. The canal of the cervix is flattened from front to back and is somewhat larger in its middle part. It is traversed by two longitudinal ridges and has oblique folds stretching from each ridge in an arrangement like the branches of a tree. The cervical canal is 2.5 centimetres (about 1 inch) in length; its opening into the vagina is called the external os of the uterus. In virgins the external os is small, almost circular, and often depressed. After childbirth, the external os becomes bounded by lips in front and in back and is thus more slitlike. The cervical canal is lined by a mucous membrane containing numerous glands that secrete a clear, alkaline mucus. The upper part of this lining, in the region called the isthmus, undergoes cyclical changes resembling, but not as marked as, those occurring in the body of the uterus. Numerous small cysts (Naboth's follicles) are found in the cervical mucous membrane. It is from this region that cervical smears are taken in order to detect early changes indicative of cancer.

The tissue layers in the uterus

The uterus is composed of three layers of tissue. On the outside is a serous coat of peritoneum (a membrane exuding a fluid like blood minus its cells and the clotting factor fibrinogen) which partially covers the organ. In front it covers only the body of the cervix; behind it covers the body and the part of the cervix that is above the vagina and is prolonged onto the posterior vaginal wall; from there it is folded back to the rectum. At the side the peritoneal layers stretch from the margin of the uterus to each side wall of the pelvis, forming the two broad ligaments of the uterus.

The middle layer of tissue is muscular (the myometrium) and comprises the greater part of the bulk of the organ. It is very firm and consists of densely packed, unstriped, smooth muscle fibres. Blood vessels, lymphatics, and nerves are also present. The muscle is more or less arranged in three layers of fibres running in different directions. The outermost fibres are arranged longitudinally. Those of the middle layer run in all directions without any orderly arrangement; this layer is the thickest. The innermost fibres are longitudinal and circular in their arrangement.

The innermost layer of tissue in the uterus is the mucous membrane, or endometrium. It lines the uterine cavity as far as the internal os, where it becomes continuous with the lining of the cervical canal. The endometrium contains numerous uterine glands that open into the uterine cavity and that are embedded in the cellular framework or stroma of the endometrium. Numerous blood vessels and lymphatic spaces are also present. The appearances of the endometrium vary considerably at the different stages in reproductive life. It begins to reach full development at puberty and thereafter exhibits dramatic changes during each menstrual cycle. It undergoes further changes before, during, and after pregnancy; during the menopause; and in old age. These changes are for the most part hormonally induced and controlled by the activity of the ovaries.

To understand the nature of the changes in the endometrium during each menstrual cycle it is usual to consider the endometrium to be composed of three layers. They blend imperceptibly but are functionally distinct: the inner two layers are shed at menstruation; the outer or basal layer remains in position against the innermost layer of the myometrium. The three layers are called, respectively, the stratum compactum, the stratum spongiosum, and the stratum basale. The stratum compactum is nearest to the uterine cavity and contains the lining cells and the necks of the uterine glands; its stroma is relatively dense. Superficial blood vessels lie beneath the lining cells. The stratum spongiosum is the large middle layer. It contains the main portions of uterine glands and accompanying blood vessels; the stromal cells are more loosely arranged and larger than in the stratum compactum. The stratum basale lies against the uterine muscle; it contains blood vessels and the bases of the uterine glands. Its stroma remains relatively unaltered during the menstrual cycle.

The menstrual cycle

The menstrual cycle extends over a period of about 28 days (normal range 21–34 days), from the first day of one menstrual flow to the first day of the next. It reflects the cycle of changes occurring in the ovary, which is itself under the control of the anterior lobe of the pituitary. The menstrual cycle is divided into four phases; menstrual, postmenstrual, proliferative, and secretory phases.

In the menstrual, or bleeding, phase the stratum compactum and most of the stratum spongiosum are shed, become liquefied, and, with a quantity of blood, flow out of the uterus through the cervical opening. During the postmenstrual phase, a process of repair takes place, and the lining, or endometrium, is re-established. In the proliferative phase the endometrium thickens, and its glands begin their secretion of mucus. The secretory, or premenstrual, phase is a period of continued endometrial growth and of active secretion by the endometrial glands. These states are discussed in more detail in the article MENSTRUATION.

The secretory phase reaches its climax about a week after ovulation. Ovulation occurs in midcycle, about 14 days before the onset of the next menstrual flow. The endometrium has been prepared and has been stimulated to a state of active secretion for the reception of a fertilized ovum. The stage has been set for the attachment of the blastocyst, derived from a fertilized ovum, to the endometrium, and for its subsequent embedding. This process is called implantation; its success depends on the satisfactory preparation of the endometrium in both the proliferative and secretory phases. When implantation occurs, a hormone from certain cells of the blastocyst causes prolongation of the corpus luteum and its continued activity. This causes suppression of menstruation and results in the maintenance of the endometrium and its further stimulation by progesterone, with consequent increased thickening. The endometrium of early pregnancy is known as the decidua.

In a cycle in which fertilization of the ovum has not taken place, the secretory phase terminates in menstruation.

The phenomenon of menstruation occurs also in the great apes and in the Old World monkeys, but in New World monkeys, uterine changes are not as dramatic and bleeding is slight. It does not occur in other mammals (except perhaps in modified form in the elephant shrew); and the slight bleeding that may occur in some mammals at about the time of ovulation, caused by high levels of estrogen, is distinct from menstruation. Menstruation may seem a wasteful process, with its loss of tissue and of blood and the iron contained in the blood. If not excessive or abnormally frequent, this loss can be readily made good by a healthy woman. The endometrium needs to be in a certain state of preparedness before implantation can occur. When this stage has been passed, menstruation occurs. Repair then re-establishes an endometrium capable of being stimulated again to the critical stage when implantation can occur.

The uterus is supplied with blood by the two uterine arteries, which are branches of the internal iliac arteries, and by ovarian arteries, which connect with the ends of the uterine arteries and send branches to supply the uterus. The nerves to the uterus include the sympathetic nerve fibres, which produce contraction of uterine muscle and constriction of vessels, and parasympathetic (sacral) fibres, which inhibit muscle activity and cause dilation of blood vessels.

The uterine tubes. The uterine tubes, often called the fallopian tubes, carry ova from the ovaries to the cavity of the uterus. Each opens into the abdominal cavity near an ovary, at one end, and into the uterus, at the other. Three sections of the tubes are distinguished: the funnel-shaped outer end, or infundibulum; the expanded and thin-walled intermediate portion, or ampulla; and the cordlike portion, the isthmus, that opens into the uterus. The infundibulum is fringed with irregular projections called fimbriae. One fimbria, somewhat larger than the others, is usually attached to the ovary. The opening into the abdomen is at the bottom of the infundibulum and is small. Fertilization of the ovum usually occurs in the ampulla of the tube. Normally the fertilized egg is transported to the uterus, but occasionally it may adhere to the tube and start developing as an ectopic or tubal pregnancy. The tube is unable to support this pregnancy, and the conceptus may either be extruded through the abdominal opening or may cause rupture of the tube, with ensuing hemorrhage.

The uterine tube is covered by peritoneum except on its border next to the broad ligament. There are inner circular and outer longitudinal layers of smooth muscle fibres continuous with those of the uterus. The inner lining has numerous longitudinal folds that are covered with ciliated columnar and secretory cells. Muscular contraction, movement of the hairlike cilia, and the passage of the watery secretions all probably assist in the transport of spermatozoa to the ampulla and of a fertilized ovum toward the uterus.

The ovaries. The female gonads, or primary sex organs, corresponding to the testes in a male, are the two ovaries. Each is suspended by a mesentery, or fold of membrane, from the back layer of the broad ligament of the uterus. In a woman who has not been pregnant, the almond-shaped ovary lies in a vertical position against a depression, the ovarian fossa, on the side wall of the lesser pelvis. This relationship is altered during and after pregnancy. Each ovary is somewhat over 2.5 centimetres (1 inch) in length, 1.25 centimetres (½ inch) across and slightly less in thickness, but the size varies much with age and with the state of activity.

Shape, position, structure, of ovaries

The mesentery of the ovary helps to keep it in position, and within this membrane lie the ovarian artery and vein, lymphatics, and nerve fibres. The uterine tube arches over the ovary and curves downward on its inner or medial surface.

Except at its hilum, the point where blood vessels and the nerve enter the ovary and where the mesentery is attached, the surface of the ovary is smooth and is covered by cubical cells. Beneath the surface, the substance of the ovary is divided into an outer portion, the cortex, and an inner portion, or medulla. The outermost part of the cortex, immediately beneath the outer covering, forms a thin connective tissue zone, the tunica albuginea. The rest of the cortex consists of stromal or framework cells, contained in a fine network of fibres, and also the follicles and corpora lutea.

The ovarian follicles, sometimes called graafian follicles, are rounded enclosures for the developing ova in the cortex near the surface of the ovary. At birth and in childhood they are present as numerous primary or undeveloped .ovarian follicles. Each contains a primitive ovum, or oocyte, and each is covered by a singie iayer of

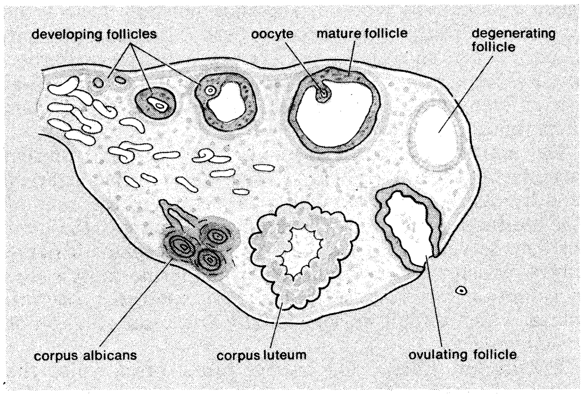


Figure 6: Cross section of the ovary.

flattened cells. As many as 700,000 primary follicles are contained in the two ovaries of a young child. Most of these degenerate before or after puberty.

Cyclical development of follicles

During the onset of puberty and thereafter until the menopause (except during pregnancy), there is a cyclic development of one or more follicles each month into a mature follicle. The covering layer of the primary follicle thickens and can be differentiated into an inner membrana granulosa and an outer vascularized theca interna. The cells of these layers (mostly the theca interna) produce estrogenic steroid hormones that exert their effects on the endometrium of the uterus and on other tissues. The maintenance and growth of the follicle to maturity is brought about by a follicle-stimulating hormone (FSH) from the anterior lobe of the pituitary gland. Another hormone, called luteinizing hormone (LH), from the anterior lobe, assists FSH to cause the maturing, now fluid-filled follicle to secrete estrogens. LH also causes a ripe follicle (1.0–1.5 centimetres in diameter) to rupture, causing the liberation of the oocyte into the peritoneal cavity and thence into the uterine tube.

This liberation of the oocyte is called ovulation; it occurs at about the midpoint of the human reproductive cycle, on the 13th–14th day of a 28-day cycle as measured from the first day of the menstrual flow.

Corpus luteum

After ovulation the ruptured follicle collapses because of loss of its follicular fluid and rapidly becomes transformed into a soft, well-vascularized glandular structure known as the corpus luteum. The human corpus luteum ("yellow body") is not yellow but is a creamy gray on section. It develops rapidly, becomes vascularized after about four days, and is fully established by nine days. The gland produces the steroid hormone progesterone and some estrogens. Its activity is both stimulated and maintained by luteinizing hormone. Progesterone stimulates glandular proliferation and secretion in an endometrium primed by estrogens.

While the ovarian follicle matures, the primary oocyte divides into a secondary oocyte and a small rudimentary ovum called the first polar body. This occurs at about the time when the follicle develops its cavity; the oocyte also gains a translucent, acellular covering or envelope, the zona pellucida. The secondary oocyte is liberated at ovulation: it is 120–140 microns in diameter and is surrounded by the zona pellucida and a few layers of cells known as the corona radiata. The final maturation of the oocyte, with the formation of the rudimentary ovum called the second polar body, occurs at the time of fertilization.

If fertilization does not occur, then the life of the corpus luteum is limited to about 14 days. Degeneration of the gland starts toward the end of this period and menstruation occurs. The corpus luteum shrinks, fibrous tissue is formed, and it is converted into a scarlike structure, called a corpus albicans, which persists for a few months. Should fertilization occur and be followed by implantation of the blastocyst, hormones (particularly chorionic gonadotrophin) are produced by cells of the blastocyst to prolong the life of the corpus luteum. It persists in an active state for at least the first two months of pregnancy until the placental tissue has taken over its endocrine (hormone-producing) function. The corpus luteum of pregnancy then also retrogresses and is becoming a fibrous scar by the time of parturition.

The ovarian arteries arise from the front of the aorta, in a manner similar to the testicular arteries, but at the brim of the lesser pelvis they turn down into the pelvic cavity. Passing in the suspensory ligament of the ovary, each artery reaches the broad ligament below the uterine tube and then passes into the mesovarium to divide into branches distributed to the ovary. One branch continues in the broad ligament to anastomose with the uterine artery. The ovarian veins emerge from each ovary as a network that eventually becomes a single vein; the terminations are similar to those of the testicular veins. The nerves are derived from the ovarian nerve network on the ovarian artery.

BIBLIOGRAPHY. R.J. HARRISON, *Reproduction and Man* (1967), a concise review of the characteristics of human reproduction and reproductive organs, and with W. MONTAGNA, *Man* (1969), a general text comparing reproduction in man with that in primates and other mammals; W. INGIULLA and R.B. GREENBLATT (eds.), *Endocrinologic and Morphologic Correlations of the Ovary* (1969), a review of modern research on hormonal activities in the ovary; A.S. PARKES (ed.), *Marshall's Physiology of Reproduction,* 3rd ed., 3 vol. (1960–66), a comprehensive reference work on comparative aspects of reproduction; R.M. WYNN, *Cellular Biology of the Uterus* (1967), a useful summary of modern research on many aspects of uterine anatomy and physiology; W C. YOUNG (ed.), *Sex* and *Internal Secretions,* 3rd ed., *2* vol. (1961), a useful reference work on basic reproductive patterns and on the endocrine control of reproductive organs; S. ZUCKERMAN (ed.), *The Ovary*, 2 vol. (1962), a standard work on structure and function in the ovary.

(R.J.Ha.)

Reproductive System Diseases

The human reproductive system is made up of the organs, glands, and secretions required for the development and maintenance of the male and female characteristics and the production of the sex cells, or gametes (spermatozoa and ova), responsible for reproduction.

The system in the male includes (1) the testes (testicles), gonads, or sex glands, that produce spermatozoa (sperm) and male sex hormones and are enclosed in the pouch of skin called the scrotum; (2) the penis, the shaft of which is made up of erectile tissue called the corpora cavernosa and the structure known as the corpus spongiosum, capped by the sensitive glans penis; (3) the sperm channels—the epididymides, the ductus deferentus, or vasa deferentia, the ejaculatory ducts, and the urethra, which also carries urine; and (4) the glands that secrete the greater part of the seminal fluid; these are the prostate, which encircles the urethra close to the bladder, the seminal vesicles, and the bulbourethral glands.

The reproductive system in the female includes (1) the ovaries, which are the source of ova and of female sex hormones; (2) the uterine (Fallopian) tubes, which carry the ova to the uterus, and normally are the site for the fertilization of the ova; (3) the uterus, or womb, in which the fertilized ovum is normally implanted and the embryo and the fetus develop; and (4) the vagina, or pouch, in which semen is deposited during sexual intercourse. The lower end of the uterus, the cervix, is a projection that opens into the vagina.

The reproductive system in the human male and female is described at more length and illustrated in the article REPRODUCTIVE SYSTEM, HUMAN. The menstrual cycle is set forth in MENSTRUATION and the development of the offspring in the uterus in DEVELOPMENT, HUMAN. Pregnancy is covered in the article PREGNANCY and childbirth in PARTURITION, HUMAN.

The human reproductive system may be affected by abnormal hormone production coming from the ovaries or the testes or from other endocrine glands, such as the pituitary, thyroid, and adrenals. Reproductive-system diseases can also be caused by genetic abnormalities, congenital anomalies (abnormalities), infections, tumours, and disorders of unknown cause.

The main sections of this article are concerned with (1) genetic and congenital abnormalities; *(2)* functional genital disorders; (3) infections; (4) structural changes of unknown cause; and (5) tumours.

GENETIC AND CONGENITAL ABNORMALITIES

In the male. Congenital anomalies of the prostate and seminal vesicles are rare; they consist of absence, hypoplasia (underdevelopment), or the presence of fluid- or semisolid-filled sacs, called cysts. Cysts of the prostatic utricle (the uterine remnant found in the male) are often found in association with advanced stages of hypospadias (a defect in the urethra, see below) and pseudohermaphroditism (in which sex glands are present but bodily appearance is ambiguous as to sex; *i.e.*, the secondary sexual characteristics are underdeveloped). Cysts may also cause urinary obstructive symptoms through local pressure on the bladder neck.

Abnormalities of the penis

Severe anomalies of the penis are rare and are generally associated with urinary or other systemic defects that are incompatible with life. Anomalies are those of absence, transposition, torsion (twisting), and reduplication of the penis. An abnormally large penis frequently is present in boys affected by precocious puberty, in congenital imbeciles, in dwarfs, in men with overactive pituitaries, and in persons affected by adrenal tumours. A small penis is seen in infantilism and in underdevelopment of the genitals, or undersecretion of the pituitary or pineal gland, and failure of development of the corpora cavernosa.

The only anomaly of the foreskin of grave concern is congenital phimosis, characterized by a contracture of the foreskin, or prepuce, sufficient to prevent its retraction over the glans; the preputial opening may be pinhole in size and may impede the flow of urine. The condition is easily remedied by circumcision, a permanent cure.

There is a considerable variety of urethral anomalies. Stenosis (contracture) of the external opening (meatus) is the most common, but congenital stricture of the urethra occasionally occurs at other points. Valves (or flaps) across the anterior or posterior part of the urethra may cause congenital urethral obstruction in boys. Posterior urethral valves are more common than anterior valves and consist of deep folds of mucous membrane, often paper-thin and usually attached at one end to the verumontanum, a small prominence in the back wall of that part of the urethra that is surrounded by the prostate gland. If too tight, the valves may obstruct the urethra and destroy the kidneys.

There are various defects associated with incomplete closure of the urethra. One of the commonest is hypospadias, in which the underside (ventral side) of the urethral canal is open for a distance at its outer end. Frequently the hypospadiac meatus is narrowed, and the penis also has a downward (ventral) curvature beyond the meatus. The posterior part of the urethra is never involved; therefore, the muscle that closes the urethra, the sphincter, functions normally, and urinary control exists. Although the condition occurs in both sexes, it is seen predominantly in the male. There is a high incidence of partial or complete failure of the testes to develop, cryptorchism (failure of the testes to descend into the scrotum), and small external and internal genitalia; variable male–female admixtures may be associated with this deficiency. Epispadias, an opening in the upper (dorsal) side of the penis, is considerably less common than hypospadias. Dorsal curvature may also be present, but the disabling aspect is that the defect usually extends through the urinary sphincter and causes urinary incontinence. Other less common urethral anomalies include complete absence of the urethra, double urethra, urethra fistula (an opening in the urethra), urethrorectal fistula (an opening between the urethra and the rectum), and urethral diverticulum (a pouch in the wall of the urethra). Most of the above conditions are correctable by surgery.

Absence or excessive number of testes

Anorchism (absence of one or both testes) is rare; it may be associated with the absence of various other structures of the spermatic tract. Generally, if one testis is absent, the other is found to be within the abdomen rather than in the scrotum. Congenitally small testes may be a primary disorder or may occur because of underactivity of the pituitary. In both disorders, there is a lack of development of secondary sexual characteristics and some deficiency in libido and potency. Supernumerary testicles are extremely rare; when present, one or more of the supernumerary testicles usually shows some disorder such as torsion of the spermatic cord. Synorchism, the fusion of the two testicles into one mass, may occur within the scrotum or in the abdomen. Cryptorchism is the term applied to all forms of imperfectly descended testes, the commonest anomaly of the spermatic tract. Testicular descent is arrested in about 3 percent of boys at puberty; the sex gland may be found anywhere between the kidney and the scrota1 area. The majority of undescended testes are in the groin (70 percent), whereas about 25 percent are in the abdomen or in the retroperitoneal space (the space behind the membrane lining the back wall of the abdomen). The condition is often bilateral, and in the unilateral cases there is no preponderance between the left or right side. Hormonal treatment may be useful in correcting the condition, but usually surgery is necessary for correction. Cryptorchism is often associated with congenital hernia and less often with hydrocele—a collection of fluid in the membranous sac that encloses the testes—or with torsion of the spermatic cord. Tumours are more common in undescended testes.

In the female. The female external genitalia are less complex than those of the male but have anomalies that can at times severely interfere with the functioning of the female urogenital tract. The clitoris, an erectile structure that corresponds to the penis, except that it does not contain the urethra, may be absent but in other cases may be enlarged on either a congenital or a hormonal basis. Fusion of the labia minora (small folds of skin covering the clitoris, the urethral opening, and the opening of the vagina) is a midline "sealing together" of the labia minora; usually a minute unfused area is left just below the clitoris, through which the child urinates and later menstruates. The chief difficulty with this anomaly is concerned with obstruction to the flow of urine and associated urinary-tract infection. An imperforate hymen (the membrane closing off the opening of the vagina) causes distension of the uterus and vagina with fluid other than blood before puberty and with blood after puberty (the two conditions are called hydrometrocolpos and hematocolpometra, respectively). The distended vagina compresses the urethra enough to interfere with urination and commonly may even cause complete retention of urine in the bladder and distension of the entire upper urinary tract. Fusion of the urethra and the hymen is characterized by a dense hymenal ring and a stenosed urethral opening. The consequent urinary obstruction commonly results in persistent urinary infection. Most of the conditions are readily remedied by surgery.

Anomalies of vagina and uterus

Anomalies of the vagina and uterus consist of complete absence, incomplete development, and duplication. The female urethra may have a congenitally narrow opening, or meatus; may be distended; may have an abnormal pouch, or diverticulum, in its wall; or may open abnormally into the vagina. Hypospadias may occur in the female but is far less common than in the male. Epispadias is also present in the female. Reconstructive surgery is the only method of treatment. One of the rarest and most severe of the urogenital-tract anomalies, called persistent cloaca, consists in congenital intercommunication between the rectum and the bladder and vagina or between the rectum and the urethra and vagina.

Intersexuality. Intersexuality (having both male and female characteristics) may be noticeable at birth or may become apparent after puberty. Intersexuality noticeable at birth may be classified as female or male pseudohermaphroditism or true hermaphroditism. Female pseudohermaphroditism, or female intersex, may be of adrenal or nonadrenal type. The adrenal type develops because of an inborn error in the metabolism of the adrenal hormone cortisol that leads to an increased secretion of corticotropin (ACTH) and consequent excessive secretion of androgens (male sex hormones). The newborn female with this condition is a chromosomal female and resembles a normal female, but an excess male hormone has a masculinizing effect on the external genitalia; the vagina tends to be connected to the urethra; the clitoris is enlarged, as are the labia (the labia majora are prominent folds of skin, corresponding to the scrotum in the male). Effective treatment can be achieved by administration of adrenal hormones (*e.g.*, cortisone, hydrocortisone), which suppress the pituitary so that its stimulus to adrenal production of androgenic hormones is minimized. The nonadrenal type of intersex is seen in infants whose mothers have been administered synthetic androgens or progestational compounds (substances that stimulate changes in the uterus that further the implantation and growth of the fertilized ovum) during pregnancy. Rarely, the condition is associated with the presence in the mother of a masculinizing tumour of the ovary or the adrenal gland. The newborn infant is a female with varying degrees of ambiguous genitalia; no treatment is necessary, and normal female development occurs at puberty.

Male pseudohermaphrodites are males with varying deficiencies of internal and external virilization. Most commonly, the male intersex has a markedly hypospadiac penis, undescended testes, a cleft scrotum, and an enlarged prostatic utricle; a complete uterus and Fallopian tubes may be found, with the vagina opening into the posterior wall of the urethra. (Such persons are pseudohermaphrodites in that they do not have ovaries.)

True hermaphrodites have recognizable ovarian and testicular tissue. **A** uterus is always present, but the internal genitalia otherwise vary greatly, often including both male and female structures. The external genitalia are usually ambiguous, but in **75** percent of the reported cases the children have been raised as males. At puberty, over 80 percent of them develop enlarged breasts, and approximately half menstruate. Most hermaphrodites are chromatin positive—that is, they have, within and near the periphery of the nuclei of their cells, a substance, chromatin, that is normally found in the cells of females but not in those of males—and over half have a characteristically female set of chromosomes in their peripheral blood cells.

Surgical and hormonal therapy directed to producing either a male or a female configuration of the body is based on the existing physical and psychological findings.

Inter-sexuality syndromes associated with puberty

Klinefelter's, Turner's, and testicular feminizing syndromes are intersexuality syndromes that become apparent prior to or after puberty. Klinefelter's syndrome is a disorder of phenotypic males (persons who have a male body configuration) who do not produce sperm, have small testes, and have varying degrees of eunuchoidism. The essential genetic abnormality is the presence of an extra sex chromosome. Thus, instead of the X and Y and total of 46 that would be anticipated from their appearance, affected persons have two X's and a Y and a total of **47** chromosomes. Patients with this syndrome have various associated medical problems, such as chronic disease of the lungs, varicose veins, thrombophlebitis (inflammation of the blood vessels), obesity, diabetes mellitus, hyperlipemia (abnormally high blood levels of fats), and enlarged breasts at the time of puberty. Mental retardation and antisocial behaviour are also associated with this syndrome. Treatment consists of the administration of androgens to prevent eunuchoid symptoms and osteoporosis (a condition characterized by lightness, porosity, and fragility of the bones). Removal of the breasts is occasionally necessary.

Turner's syndrome is a disorder of phenotypic females —persons who are female in physical configuration. Characteristically, such persons are short, do not menstruate, and show estrogen (a female sex hormone) deficiency; there is a distinctive cluster of congenital anomalies. In the typical cases, the ovary is an undifferentiated streak of cells inside the body, with no evidence of activity related to secretion or production of ova. Internal genital development is female, with a uterus and Fallopian tubes, and external development is immaturely female, with no virilization except for occasional enlargement of the clitoris. Treatment consists of administration of estrogens.

The disorder known as the testicular feminizing syndrome is inherited. Affected persons seem to be of normally developed females but have a chromosomal sex that is that of the normal male. The gonads are well-developed testes, and evidence indicates that there is a normal production of testosterone (male hormone), but there is cellular resistance to the action of this hormone, and therefore the affected person becomes female in appearance. Because these gonads are apt to form malignant tumours, they are usually removed surgically. Female sexual characteristics are then maintained by the administration of estrogenic hormones.

FUNCTIONAL GENITAL DISORDERS

Affecting both male and female systems. Delayed puberty. The term delayed puberty may be a misnomer, because puberty delayed beyond age **19** is in fact a permanent failure of sexual development because of an abnormally low secretion by the pituitary gland of gonadotropic hormone, the hormone that stimulates growth and activity of the sex glands; this condition is called hypogonadotropic eunuchoidism. The term delayed puberty is usually applied to boys who develop more slowly than the average but who still eventually undergo full sexual development. Only in retrospect—*i.e.*, after the affected person reaches the age of 20—can one clearly differentiate these cases from the classic or incomplete forms of hypogonadotropic eunuchoidism. If there are social and psychological problems related to the sexual underdevelopment, therapy may consist of a course of chorionic gonadotropin, a hormone produced by the placenta and secured from the urine of pregnant women. If puberty is merely delayed, it will usually progress normally after this treatment. If it fails to progress, the patient does not have delayed puberty but rather has hypogonadotropic eunuchoidism.

Precocious puberty. In healthy girls living in a temperate climate, the earliest sign of puberty occurs at a mean age of 10.6 years (standard deviation of 1.2 years), whereas, in boys, testicular growth begins at a mean age of 11.8, with a standard deviation of one year. The average age of menstruation is **13.5** years (range, **9–17** years). What is called true precocious puberty is a condition in which normal pituitary–gonadal function is activated at an abnormally early age. It is always isosexual with the sex gonads (*i.e.*, it is always in keeping with the sex of the gonads) and with development of the secondary sexual characteristics and production of spermatozoa or ova. Pseudoprecocious puberty includes development of secondary sexual characteristics but not production of spermatozoa or ova and may be isosexual or may be heterosexual (*i.e.*, it may involve virilization in the female or feminization in the male).

Pseudo-precocious puberty

The causes of true precocious puberty are several—including brain lesions and hypothyroidism (abnormally low secretion by the thyroid glands); the largest proportion of cases are of unknown cause. Precocious pseudopuberty in females may be caused by ovarian tumours, a cyst of the ovary, a tumour of the adrenal cortex (outer substance of the adrenal gland), or congenital overdevelopment of the adrenal gland. In males, the causes include congenital overdevelopment of the adrenal glands, tumour of the adrenal cortex, tumour involving the Leydig cells of the testes, and teratoma (a tumour containing numerous types of tissue; in these circumstances it includes adrenal-cortical tissue).

The diagnosis of precocious puberty is warranted when sexual maturation begins before the age of eight in girls or ten in boys. The condition is twice as common in girls as in boys but in girls is far less indicative of serious disease. Most prematurely developed girls (80 percent) have no demonstrable structural abnormality of brain, pituitary gland, or ovary; more than 60 percent of sexually precocious boys have serious organic disease. In girls, ovarian tumours account for **15** percent of the cases of precocious development and intracranial lesions for only **4** percent. In boys, 20 percent of the cases of precocious development are due to intracranial disease, **25** percent to adrenal disease, and less than 10 percent to lesions of the testicle. Precocious puberty of unknown cause is four to ten times as common in girls than in boys.

Infertility. At least 10 percent of marriages are barren, and deficiencies of sperm production in the male are the causal factor in 40 percent of these. The common causes of male infertility are deficiencies in maturation of germ cells (sperm); orchitis (mumps), with destruction of the testes; obstruction of the passageways for sperm; abnormally low thyroid or high adrenal secretion; varicocele (enlargement of the veins of the spermatic cord); or formation of antibodies to sperm by the male or the female. The most important steps in the evaluation of male infertility are examination of the semen and of a specimen of the tissue of the testes. Evaluation also includes chromatin analysis and observation of thyroid, adrenal, and pituitary function. The results of treatment of infertility in the male are usually unsatisfactory, except when a varicocele or obstruction in the sperm passageways is the cause, in which case surgical correction may be beneficial.

Causes of infertility in females

Infertility in the female is related to the faulty production of ova or to interferences with their union with spermatozoa. Vaginal causes are usually uncommon, but obstruction may be due to an unruptured hymen or may be functional and arise from enlargement and contraction of the levator ani muscles (these muscles form a supporting sheet under the pelvic cavity, with openings for structures such as the anus and the vagina). Abnormalities of the cervix are among the most important causes obstructing the passage of sperm. (The sperm normally enter the uterus through the cervix and, from the uterus, move into a uterine or Fallopian tube, where fertilization of an ovum takes place.) During the few days prior to ovulation—release of an ovum from the ovary—the glands within the cervix normally secrete a thin, watery mucus that is beneficial to sperm survival and migration. Various factors, such as infection or estrogen deficiency, may decrease the quality of the mucus.

Uterine anomalies such as a bicornuate (double) uterus may play a role in infertility. Total or partial blocking of the uterine tubes can result from inflammation due to infection (*e.g.*, gonorrhea) or from endometriosis, a condition involving the presence of tissue resembling that which lines the uterus elsewhere in the pelvic cavity. Thyroid, pituitary, adrenal, or ovarian disease may interfere with ovulation, as may the presence of large numbers of cysts in the ovaries (the condition known as polycystic ovaries).

Finally, emotional factors may play a role in causing infertility. Treatment consists of the use of various hormones, surgical correction of tubal blockage, and psychotherapy. With the advent of new hormone preparations, the results in achieving pregnancy have been vastly improved.

Affecting female system. ***Menstrual disorders.*** Amenorrhea (failure to menstruate) may be caused by congenital abnormalities of the female reproductive system and disturbances of the pituitary, thyroid, and adrenal glands and the ovaries. The pituitary may be affected by cancer (adenoma or carcinoma), by cysts, by an acute infection, by the functional disorder known as Frölich's syndrome, or by anorexia nervosa (a condition in which aversion to eating, itself caused by some underlying emotional disorder, leads to emaciation and other abnormal effects). Both overactivity and underactivity of the thyroid may cause either amenorrhea or excessive uterine bleeding. Abnormally high secretion by the adrenal glands causes amenorrhea as a secondary effect of the virilization that results, but low adrenal secretions (Addison's disease), when present for some time, also causes amenorrhea as a typical symptom. Amenorrhea may be caused by masculinizing tumours of the ovaries, by polycystic ovaries (Stein-Leventhal syndrome), or by infections; radiation of the ovaries may decrease production of estrogens (female sex hormones) enough to cause amenorrhea.

There is also spontaneous premature failure of the ovaries (in women in their early 30s), which is similar to the normal menopause in its effects. Lastly, debilitating diseases such as tuberculosis or severe malnutrition may cause amenorrhea.

Abnormal uterine bleeding

Abnormal uterine bleeding may be excessive menstrual bleeding, scanty menstrual bleeding, too frequent menstruation, or bleeding other than at the time of menstruation. The causes include lesions in the uterus, pregnancy disorders, endometriosis, lesions in the ovaries, systemic diseases, administration of hormones, or underactivity or overactivity of the thyroid. Dysfunctional uterine bleeding—that is, irregular, excessive, or prolonged menstrual bleeding from nonorganic or endocrine causes—is most frequently seen in women who are near 40 years of age. Dysmenorrhea is of two types, called primary and acquired. Primary dysmenorrhea is usually not associated with any organic pelvic abnormality. Characteristically, there are severe cramps during menstruation, accompanied by pain in the pelvis, nausea, vomiting, and tension. Acquired dysmenorrhea usually accompanies some organic pelvic abnormality, such as endometriosis or cervical stenosis—narrowing of the opening in the cervix.

Affecting the male system. ***Impotence.*** Impotence is inability of the male to have satisfactory sexual intercourse and varies in form from the inability to gain an erection to weak erections, premature ejaculation, or loss of normal sensation with ejaculation. Almost all of these complaints are psychogenic in origin, but impotence may be caused by subnormal functioning of the testes, by arteriosclerosis (hardening of the arteries), diabetes mellitus (a metabolic disease in which there is inadequate secretion or utilization of insulin), or by some disease of the nervous system. Certain medications prescribed for the treatment of such diseases as peptic ulcer, hypertension, or psychiatric illnesses may adversely affect sexual ability. Therapy, usually limited in its success, includes administration of sex hormones and psychotherapy.

Priapism. Priapism is prolonged penile erection that is painful and unassociated with sexual stimulation. The blood in the spaces of the corpora cavernosa becomes sludgelike and may remain for hours or even for days. About 25 percent of the cases are associated with leukemia (a disease of the blood-forming tissues that results in extremely high numbers of white blood cells), sickle-cell anemia (an inherited disease in which red blood cells are abnormal in shape and function and the hemoglobin is of a particular type), metastatic carcinoma (cancerous development at a distance from the primary site), and diseases of the nervous system, but in the majority of cases the causation is not clear. There have been many forms of therapy, but prompt surgical treatment with evacuation of the blood from the corpora appears to be the best. Regardless of treatment, impotence is common after an episode of priapism and even more common after repeated episodes of priapism.

INFECTIONS

Venereal infections. The principal venereal infections are syphilis, gonorrhea, chancroid, lymphogranuloma venereum, and granuloma inguinale.

Syphilis. Syphilis is caused by infection with ***Treponema pallidum.*** Two to four weeks after sexual exposure, a painless lump develops on the skin or mucous membranes of the genitalia, the region around the anus, or the oral cavity and then breaks down, forming a hard, punched-out ulcer (called the primary chancre). The diagnosis is made early by identifying the organism on microscopic examination. Several weeks after exposure, syphilis can be identified by standard tests of the blood serum. The genitalia of both male and female may be affected by flat, raised nodules in the secondary state of syphilis. In the tertiary state, the testes or epididymides may be affected with nodule (gumma) formation. Treatment with massive doses of penicillin is extremely effective, and, if treatment is early, the chances of cure are excellent. A female with syphilis can transmit the disease to a fetus for an undetermined period of years. Infection of the fetus usually occurs after the fifth month of pregnancy and may result in spontaneous abortion, miscarriage, a stillborn fetus, or a fatally ill infant.

Syphilitic infection of fetus

Gonorrhea. Gonorrhea, commonest of the venereal diseases, is caused by the organism *Neisseria gonor-*

rhoeae. Pain or a burning sensation on urination is an early symptom. The infection tends to spread. In the male, it may include the posterior portion of the urethra, the prostate, and the seminal vesicles. Involvement of the epididymides may lead to sterility. In the female, a large, painful abscess may form near the opening into the vagina; spread to the uterine tubes may lead to sterility. Other possible effects include inflammation of the heart lining and arthritis. Identification of the disease depends upon finding the organism in smears from the urethral or vaginal discharge. Treatment is with antibibtics.

Chancroid. Chancroid, a common venereal disease caused by *Haemophilus ducreyi*, first becomes evident two to 14 days after sexual intercourse, with the appearance of a small red papule, or pimple, on the genitals or surrounding skin. During the next few days a series of changes occur. Pus forms in the lesion; it then goes through a necrotic stage (death of tissue), followed by ulceration. Other lesions may develop. Identification of the disease is by examination of a smear, culture (growth of the organism in the laboratory), and skin test. Treatment is with antibiotics or sulfonamide drugs.

Other venereal diseases. Lymphogranuloma venereum, an infection with an organism of the group that cause psittacosis (parrot fever, or ornithosis) and the eye disease trachoma, inflames the lymph nodes and channels of the genitals and rectum, with eventual formation of fistulas (open channels) and scar tissue. It is treated with antibiotics, usually a tetracycline or chloramphenicol. Surgical treatment may also be required.

Granuloma inguinale, a slowly developing and relatively painless disease, causes scar tissue and ulcers in the skin and subcutaneous tissues of the genitalia and adjacent regions. The micro-organism responsible, ***Donovania granulomatis,*** is transmitted during sexual intercourse. Cure is usually effected by use of tetracyclines or chloramphenicol.

Types of puerperal infection

Nonvenereal infections. ***Puerperal infection.*** Puerperal infection, one of the commonest causes of death in childbearing, is an infection of a wound in the birth canal (usually in the endometrium, the lining of the uterus). The infection may be divided into two main categories: (1) local involvement of the vagina, cervix, endometrium, and adjacent tissues and (2) extensions of the original process to the veins, to the lymph vessels, to the membranes lining the abdomen, and to the uterine tubes. The vast majority of puerperal infections are caused by the ***Streptococcus,*** but any of a number of other organisms may be iesponsible. The treatment consists chiefly of antibiotics, supportive therapy, and, occasionally, surgical drainage of an abscess.

Tuberculosis. Tuberculosis of the reproductive system usually is brought from another part of the body by way of the bloodstream. In the male, the kidney and possibly the prostate are the primary sites of tuberculosis infection in the genitourinary tract, and from these sites the infection can invade the seminal vesicles, the epididymides, and the testes. Tuberculosis of the prostate and seminal vesicles may not cause any symptoms; the first indication of involvement of these organs is the onset of epididymitis — inflammation of the epididymides — with subsequent scrotal abscess and sinus formation. The diagnosis is made by finding the tuberculosis bacilli in urine or semen or by examining a specimen of tissue from the genitourinary tract. The semen has been known to infect the vagina.

In the female, the uterine tubes are usually first infected; then the endometrium, ovaries, **cervix,** and vulva.

The treatment in both sexes is basically medical and consists of the administration of antituberculotic drugs in various combinations for periods of up to at least two years. Surgical excision of an infected organ may also be necessary. The outlook is good with early diagnosis and treatment, except that sterility may possibly result.

Other infections. Balanitis, or inflammation of the glans penis, and posthitis, or infection of the prepuce, is usually caused by retention of secretions and bacteria beneath the foreskin.

Elephantiasis of the scrotum, a chronic disease caused by interruption of the lymphatic drainage of the scrotum, leads to massive hypertrophy (enlargement) of the skin and collection of fluid subcutaneously.

Prostatitis

Acute prostatitis is an inflammation of the prostatic tissue caused by bacterial invasion or by toxins produced by these bacteria. Infection is usually associated with inflammation of the posterior part of the urethra and the seminal vesicles. Chronic prostatitis can occur at any age but is probably the most common chronic infection in the reproductive system of men over 50 and usually follows acute inflammation of the prostate. Inflammation of the epididymis (epididymitis) is the most common disease of the scrotal contents and is commonly a sequel to prostatitis, the usual pathway of invasion being the vas deferens. Herpes progenitalis is a common lesion of the penis caused by a virus; it is manifested by groups of small blisters on the surface of the glans or prepuce along the pathway of the dorsal penile nerve. In the female, similar vesicles (blisters) may appear in the area of the vulva, and in many cases its occurrence coincides with the onset of menstruation.

Condylomata acuminata are soft, painless, cauliflower-like growths caused by a virus. They appear on the prepuce, glans penis, within the adjacent part of the urethra, and on the female perineum (the area between the external genitals and the anus). Other infections that occur in the genital organs of women include Bartholinitis, or inflammation of Bartholin's glands, near the opening of the vagina; trichomonas vaginitis, infection with the parasite *Trichomonas vaginalis*; and monilia vulvovaginitis, a common inflammation caused by the fungus ***Candida*** *albicans.*

STRUCTURAL CHANGES OF UNKNOWN CAUSES

In the female: endometriosis. Endometriosis, a disease occurring only during a woman's menstrual life, is the growth of endometrial tissue in an abnormal location. This may occur in the uterus or elsewhere. The most common location of the implants of endometrial tissue are the ovaries; other areas and organs affected (in order of incidence) are uterosacral ligaments (thickened portions of the sheet of connective tissue covering the pelvic organs), the rectovaginal septum (the membrane dividing the rectum from the vagina), the sigmoid colon (that portion of the large intestine that leads into the rectum), the lower genital tract, the round ligaments of the uterus, and the peritoneum (membrane) lining the pelvis. Characteristic symptoms associated with this disease include (1) progressive, severe pain associated with menstruation or occurring just before it; (2) dyspareunia (painful intercourse); (3) painful defecation; (4) slight bleeding before menstruation and excessive flow during menstruation; (5) painful urination and blood in the urine; and (6) infertility. The treatment includes suppression of ovulation by administration of hormones, and surgical treatment. Although endometriosis is a progressive disease in most instances, pain relief following conservative surgery (surgery that preserves ability to bear children) has occurred in an estimated 80 percent of patients, and 40–50 percent were able to become pregnant.

Symptoms of endometriosis

In the male: benign hypertrophy of the prostate. Benign prostatic hypertrophy, an overgrowth of normal glandular and muscular elements of the prostate gland, arises in the immediate vicinity of the urethra and is the most frequent cause of urinary obstruction. The enlarged prostate usually causes symptoms after the age of 50. If undetected, the obstruction may cause bladder and kidney damage. The diagnosis is made by rectal examination, excretory urography (X-raying the urinary tract while an opaque substance is being excreted in the urine), and cystourethroscopy (direct viewing of the bladder and urethra). Treatment is by surgical removal of the excess tissue. The prognosis is good if detection is early and treatment is given before the kidneys are damaged.

TUMOURS

In the male. ***External genitalia.*** Tumours of the penis are almost all of epithelial (covering or lining) origin and usually involve the foreskin or glans. Cancer of the penis (epithelioma) is rarely found in men who have been circumcised during infancy. The growth arises on the glans or inner surfaces of the prepuce, and metastases (secondary growths at distant parts) occur through lymph channels that lead to the inguinal (groin) and iliac nodes (nodes along the aorta and iliac arteries). The diagnosis is made by examination of a specimen of the lesion. Treatment for small lesions consists of surgical removal of a part of the penis or by X-ray therapy, while spread to inguinal nodes may be treated by removal of the node. The outlook is good if the cancer is small and there has been no metastasis.

Tumours of the scrotal skin are rare; most are thought to arise from occupational exposure to various carcinogens (cancer-causing substances), such as the soot in chimney sweeps' clothing. Primary tumours of the epididymis are also uncommon, and most are benign.

Testicular tumours. Testicular tumours are usually malignant; the peak incidence is between the ages of *20* and 40 years. This type of cancer accounts for about 0.5 percent of all malignant growths in men and about 4 percent of all tumours affecting the genitourinary tract. The great majority of testis tumours (greater than 95 percent) are of types that do not reproduce cells resembling those of the tissue of origin. The major route of metastases is via the lymphatics. The lymph nodes in the loins and the mediastinum—the region between the lungs—are most commonly involved, but the lungs and liver are also frequent sites of tumour spread. The remaining 5 percent of the testicular tumours, which usually resemble the cells from which they arise, include the hormone-secreting tumours. In general, these tumours have been described in all age groups, have usually been benign in behaviour, and have been most frequent in poorly developed or undescended testes.

The most common symptom first observed in all groups is painless enlargement of the testis. If, after careful examination, tumour cannot be ruled out, the testicle is removed for microscopic examination. Further treatment may consist of removal of the retroperitoneal lymph nodes (the lymph nodes in the region behind the peritoneum, the membrane lining the abdomen), X-ray therapy, or chemotherapy.

Carcinoma of the prostate. Carcinoma (cancer) of the prostate is rare before the age of 60 but increases in frequency every decade thereafter. It is the second most common cause of death from cancer in the male, second to cancer of the lung. In men over *60,* it is the commonest cause of cancer deaths. Like most tumours, prostatic cancer has no known cause, but it is clear that its growth is strikingly influenced by sex hormones or their withdrawal. Viruses may also play a role. The progress of the cancer is so slow that, by the time it produces symptoms of urinary obstruction, metastasis has occurred in many cases, most frequently to the spine, the pelvic bones, or the upper portions of the thigh bones. The diagnosis is made by finding cancer cells in a specimen of tissue taken from the prostate. Elevated levels of acid phosphatase (an enzyme of the prostate) are found in the blood (in *75* percent of cases) when the cancer has extended outside the prostate capsule and metastases are present.

Slow growth of prostate cancer

If the tumour is discovered before it has extended beyond the prostate, the gland is removed. If spread has occurred, various palliative measures offer the affected person much relief.

In the female. *Carcinoma of the vulva.* Primary carcinoma of the vulva (the external female genital organs) usually occurs in women over *50* and usually arises from the labia majora or labia minora. Most patients first notice a lump on the vulva or perineum; the diagnosis is made by examination of a specimen of tissues. Treatment consists of surgical removal of the vulva and of regional lymph nodes.

Cancer of the cervix of the uterus. Cancer of the cervix is the most common malignant tumour of the female genital tract; it is second only to cancer of the breast as a cause of death from cancer in women. The average age of occurrence for cancer of the cervix is the 45th year. The initial diagnosis is made by screening with such tests as those developed by Papanicolaou and Traut. (These consist of staining smears from vaginal and other secretions and examining them for cancer cells.) The final diagnosis rests on examining specimens of tissue from the cervix. Treatment now is usually irradiation instead of surgery because of the uncertainties of total surgical excision and the illness associated with extreme surgery. The prospect of five-year survival is as good as 85 percent if the cancers have not spread beyond the cervix.

Uterine fibromyomas. Uterine fibromyomas (fibroids) are the most frequent cause of enlargement of the uterus. They are most common in Negroes and in persons who have not borne children and are most often identified in women aged 30–45 years. New tumours do not originate after the menopause, and existing ones usually regress at that time but do not disappear. The tumours, which are benign, originate from the smooth muscle cells of the uterus wall and may be single but usually are multiple, pseudoencapsulated nodules. The symptoms are quite variable and depend largely on the location and size of the tumour. Excessive menstrual bleeding is often caused by fibroids. The diagnosis is tentatively made by pelvic examination and confirmed at surgery. Small asymptomatic fibromyomas need not be treated; the larger ones are dealt with by total or partial removal of the uterus or by irradiation.

Diagnosis of fibromas

Carcinoma of the body of the uterus. Cancer of the endometrium (the lining) of the body of the uterus is the second most common malignant tumour of the uterus and the female genital tract. The peak incidence is in the mid-50s, and there is also a strikingly high incidence in women who have not borne children. The chief symptom of the cancer is postmenopausal uterine bleeding. The Papanicolaou smear is not a reliable screening test, and an examination of a specimen of endometrial tissue must be performed. The treatment is primarily surgical but is often supplemented with preoperative intrauterine radium application or preliminary deep X-ray therapy to the pelvis. The survival rate from this disease is relatively good if the tumour is confined to the uterine body.

Ovarian tumours. No other organ in the body develops such a variety of tumours as does the ovary. The symptoms and signs may be due to the hormones secreted or may be only those of an enlarging mass in the pelvis. The final diagnosis is usually made at abdominal exploration. The treatment consists of surgery, X-ray therapy, or chemotherapy. The prognosis is variable and depends on the type of tumour that is present as well as the extent of metastatic spread.

Periodic pelvic and rectal examinations for both men and women, before any symptoms have occurred, would save almost all of those who now die from genital cancers.

BIBLIOGRAPHY. P.B. BEESON and W. MCDERMOTT (eds.), *Cecil-Loeb Textbook of Medicine,* 13th ed. (1971); and M.M. WINTROBE *et al.* (eds.), *Harrison's Principles of Internal Medicine,* 6th ed. (1970), are two excellent general textbooks of medicine that cover many of the subjects in this article in a short concise manner. "The Male Reproductive System" in *Christopher's Textbook of Surgery,* 9th ed. (1968), is one of the best summaries of the diseases of the male reproductive system. F.H. NETTER, *A Compilation of Paintings on the Normal and Pathological Anatomy of the Reproductive System* (1954), is one of the best books for anatomical illustrations on the reproductive system. M.F. CAMPBELL and J. HARTWELL HARRISON (eds.), *Urology,* 3rd ed., 3 vol. (1970), represents one of the most complete textbooks in the field of urology; D.R. SMITH, *General Urology,* 6th' ed. (1969), is well written and quite complete. D.D. FEDERMAN, *Abnormal Sexual Development* (1967), presents in complete detail all aspects of abnormal sexual development. E. STEWART TAYLOR, *Essentials of Gynecology,* 4th ed. (1969); and R.W. KISTNER, *Gynecology* (1964), cover the field of gynecology in detail; W.D. and D.W. BEACHMAN, *Synopsis of Gynecology,* 7th ed. (1967), is a good outline on female reproductive diseases. N.J. EASTMAN and L.M. HELLMAN (eds.), *Williams' Obstetrics,* 13th ed. (1966), is a classic in the field of obstetrics. R.H. WILLIAMS (ed.), *Textbook of Endocrinology,* 4th ed. (1968), covers in detail the endocrine aspects of the male and female reproductive systems.

(N.A.Ro./J.K.La.)

Reproductive Systems, Animal

The role of reproduction is to provide for the continued existence of a species; it is the process by which living organisms duplicate themselves. Animals compete with other individuals in the environment to maintain themselves for a period of time sufficient to enable them to produce tissue nonessential to their own survival, but indispensable to the maintenance of the species. The additional tissue, reproductive tissue, usually becomes separated from the individual to form a new, independent organism.

This article describes the reproductive systems in metazoans (multicelled animals) from sponges to mammals, exclusive of man. It focusses on the gonads (sex organs), associated ducts and glands, and adaptations that aid in the union of gametes—*i.e.*, reproductive cells, male or female, that are capable of producing a new individual by union with a gamete of the opposite sex. Brief mention is made of how the organism provides for the development of embryos and of the regulatory role of gonads in vertebrate cycles. (Human reproduction is treated in REPRODUCTIVE SYSTEM, HUMAN.)

This article is divided into the following major sections:

I. General features

Unlike most other organ systems, the reproductive systems of higher animals have not generally become more complex than those of lower forms. Asexual reproduction (*i.e.*, reproduction not involving the union of gametes), however, occurs only in the invertebrates, in which it is common, occurring in animals as highly evolved as the sea squirts, which are closely related to the vertebrates. Temporary gonads are common among lower animals; in higher animals, however, gonads are permanent organs. Hermaphroditism, in which one individual contains functional reproductive organs of both sexes, is common among lower invertebrates; yet separate sexes occur in such primitive animals as sponges, and hermaphroditism occurs in animals more highly evolved —*e.g.*, the lower fishes. Gonads located on or near the animal surface are common in the lowest invertebrates, but in higher animals they tend to be more deeply situated and often involve intricate duct systems. In echinoderms, which are among the highest invertebrates, the gonads hang directly into the sea and spill their gametes into the water. In protochordates, gametes are released into a stream of respiratory water that passes directly into the sea. Duct systems of the invertebrate flatworms (Platyhelminthes) are relatively complex, and those of specialized arthropods (*e.g.*, insects, spiders, crabs) are more complex than those of any vertebrate. Copulatory organs occur in flatworms, but copulatory organs are not ubiquitous among vertebrates other than reptiles and mammals. The trend toward fewer eggs and increased parental care in higher animals may account for the relative lack of complexity in the reproductive systems of some advanced forms. As trends toward increasing structural complexity have often been reversed during evolution, however, reproductive behaviour patterns in many phylogenetic (*i.e.*, evolutionary) lines have become more complicated in order to enhance the opportunity for fertilization of eggs and maximum survival of offspring (see SEX AND SEXUALITY).

A direct relationship exists between behaviour and the functional state of gonads. Reproductive behaviour induced principally but not exclusively by organic substances called hormones promotes the union of sperm (spermatozoa) and eggs, as well as any parental care accorded the young. There are a number of reasons why behaviour must be synchronized with gonadal activity. Chief among these are the following:

Synchronization of behaviour and gonadal activity

Individuals of a species must congregate at the time the gonads contain mature gametes. This often entails migration, and some members of all major vertebrate groups migrate long distances to gather at spawning grounds or rookeries.

Individuals with gametes ready to be shed must recognize members of the opposite sex. Recognition is sometimes by external appearance or by chemical substances (pheromones), but sex-linked behaviour is often the only signal.

Geographical territories frequently must be established and aggressively defended.

The building of nests, however simple, is essential reproductive behaviour in many species.

When fertilization of aquatic forms is external, sperm and eggs must be discharged at approximately the same time into the water, since gametes may be quickly dispersed by currents. Courtship, often involving highly intricate behaviour patterns, serves to release the gametes of both mating individuals simultaneously.

When fertilization is internal, willingness of the female to mate is often essential. Female mammals not in a state of willingness to mate not only will not mate but may injure or even kill an aggressive male. The unwillingness of a female mammal to mate when mature eggs are not present prevents loss of sperm needed to preserve the species.

Parental care of fertilized eggs by one parent or the other has evolved in many species. Parental behaviour includes fanning the water or air around the eggs, thereby maintaining appropriate temperature and oxygen levels; secretion of oxygen from a parent's gills; transport of eggs on or in the parental body (including the mouth of some male parents); and brooding, or incubation, of eggs.

Some species extend parental care into the postnatal period, feeding and protecting the offspring. Such behaviour patterns are adaptations for survival and thus are essential; all are induced by the nervous and endocrine systems and are typically cyclical, because gonadal activity is cyclical (see also REPRODUCTIVE BEHAVIOUR).

II. Reproductive systems of invertebrates

GONADS, ASSOCIATED STRUCTURES, AND PRODUCTS

Although asexual reproduction occurs in many invertebrate species, most reproduce sexually. The basic unit of sexual reproduction is a gamete (sperm or egg), produced by specialized tissues or organs called gonads. Sexual reproduction does not necessarily imply copulation or even a union of gametes. As might be expected of such a large and diverse group as the invertebrates, many variations have evolved to ensure survival of species. In many lower invertebrates, gonads are temporary organs; in higher forms, however, they are permanent. Some invertebrates have coexistent female and male gonads; in others the same gonad produces both sperm and eggs. Animals in which both sperm and eggs are produced by the same individual (hermaphroditism) are termed monoecious. In dioecious species, the sexes are separate. Generally, the male gonads ripen first in hermaphroditic animals (protandry); this tends to ensure cross-fertilization. Self-fertilization is normal, however, in many species, and some species undergo sex reversal.

Sponges, coelenterates, flatworms, and aschelminths. Sponges are at a cellular level of organization and thus do not have organs or even well-developed tissues; nevertheless, they produce sperm and eggs and also reproduce asexually. Some species of sponge are monoecious, others are dioecious. Sperm and eggs are formed by aggregations of cells called amoebocytes in the body wall; these are not considered gonads because of their origin and transitory nature.

In hydrozoan coelenterates, temporary gonads are formed by groups of cells in either the epidermis (outer cell layer) or gastrodermis (gut lining), depending on the

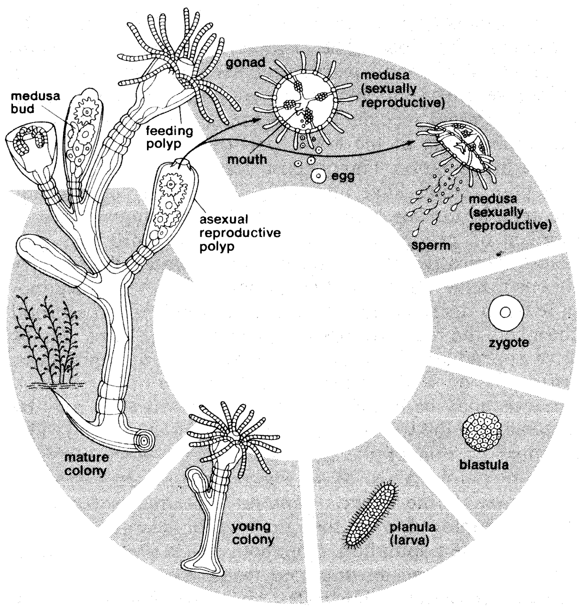


Figure 1: Life cycle of the colonial hydrozoan *Obelia*.

species; scyphozoan and anthozoan coelenterates generally have gonads in the gastrodermis. The origin and development of gonads in coelenterates, particularly freshwater species, are often associated with the seasons. Freshwater hydrozoans, for example, reproduce asexually until the onset of cold weather, which stimulates them to form testes and ovaries. Colonial hydrozoans asexually produce individuals known as polyps. Polyps, in turn, give rise to free swimming stages (medusae), in which gonads develop (Figure 1). The body organization of sponges and coelenterates is such that most of their cells are in intimate contact with the environment; consequently, gametes are shed into the water, and no ducts are necessary to convey them to the outside.

Polyps

In contrast to sponges and coelenterates, platyhelminths generally have well-developed organ systems of a permanent nature and, in addition, have evolved secondary reproductive structures to convey sex products. One exception is the acoels, a group of primitive turbellarians; they lack permanent gonads, and germinal cells develop from amoebocytes in much the same manner as in sponges. The majority of flatworms, however, are monoecious, the primary sex organs consisting of one or more ovaries and testes (Figure *2)*. The tube from the ovary to the outside is called the oviduct; it often has an outpocketing (seminal receptacle) for the storage of sperm received during copulation. In many species the oviduct receives a duct from yolk (vitelline) glands, whose cells nourish the fertilized egg. Beyond the entrance of the duct from the yolk glands the oviduct may be modified to secrete a protective capsule around the egg before it is discharged to the outside. The male organs consist of testes, from which extend numerous tubules (vasa efferentia) that unite to form a sperm duct (vas deferens); the latter becomes an ejaculatory duct through which sperm are released to the outside. The sperm duct may exhibit expanded areas that store sperm (seminal vesicles), and it may be surrounded by prostatic cells that contribute to the seminal fluid. The sperm duct eventually passes through a copulatory organ. The same basic structural pattern, somewhat modified, is found in most higher invertebrates.

Reproductive organs of nematodes

Aschelminthes (roundworms) are mostly dioecious; frequently there are external differences between males and females (sexual dimorphism). The males are generally smaller and often have copulatory spicules. Nematodes have relatively simple reproductive organs, a tubular testis or ovary being located at the end of a twisted tube. The portion of the female tract nearest the ovary forms a uterus for temporary storage of fertilized eggs. Some species lay eggs, but others retain the egg in the uterus until the larva hatches. The sperm are released into a cavity called the cloaca. A number of free-living nematodes are capable of sex reversal—if the sex ratio in a given population is not optimal or if environmental conditions are not ideal, the ratio of males to females can be altered. This sometimes results in intersexes; *i.e.*, females with some male characteristics. Hermaphroditism occurs in nematodes, and self-fertilization in such species is common. Unisexual reproduction among rotifers is described below (see Parthenogenesis).

Annelids and mollusks. Annelids have a well-developed body cavity (coelom), a part of the lining of which gives rise to gonads. In some annelids, gonads occur in several successive body segments. This is true, for example, in polychaetes, most of which are dioecious. Testes and ovaries usually develop, though not invariably, in many body segments; and the sperm and eggs, often in enormous numbers, are stored in the coelom. Fertilization is external. In oligochaetes (all of which are monoecious) on the other hand, the gonads develop in a few specific segments. Sperm are stored in a seminal vesicle and eggs in an egg sac, rather than in the coelom. A portion of the peritoneum, the membrane lining the coelom, becomes a saclike seminal receptacle that stores sperm received from the mate. The earthworm, *Lumbricus* terrestris, is an example of a specialized annelid reproductive system. Female organs consist of a pair of ovaries in segment 13; a pair of oviducts that open via a ciliated funnel (*i.e.*, with hairlike structures) into segment 13 but open to the exterior in segment 14; an egg sac near each funnel; and a pair of seminal receptacles in segment 9 and another in segment 10. Male organs consist of two pairs of minute testes in segments 10 and 11, each associated with a ciliated sperm funnel leading to a tiny duct, the vas efferens. The two ducts on each side lead to a vas deferens that opens in segment 15. Testes and funnels are contained within two of three pairs of large seminal vesicles that occupy six body segments.

Reprinted with permission of The Macmillan Company from *College Zoology* by *K.A.* Stiles. R.W. Hegner and R.A. Boolootian.
© Copyright by The Macmillan Company, 1969

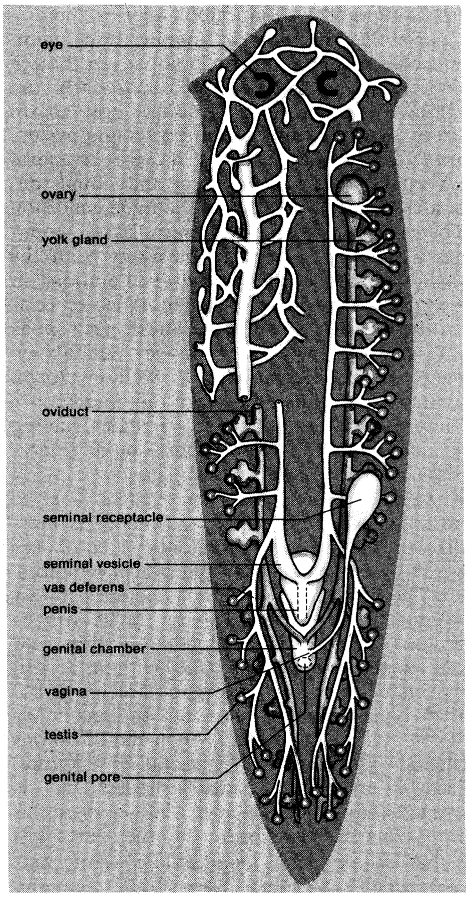


Figure 2: Reproductive system of a planarian flatworm, a monoecious animal.

Leeches (Hirudinea), also monoecious, have one pair of ovaries and a segmentally arranged series of testes with duct systems basically similar to those of earthworms.

Differences between mollusk and annelid systems

Although mollusks have a close evolutionary kinship to annelids, they have reduced or lost many structures characteristic of segmented worms. The coelom persists only as three regional cavities: gonadal, nephridial (kidney), and pericardial (heart). In ancestral forms these were interconnected so'that gametes from the gonad passed through the pericardial cavity, the nephridial cavity, then to the outside through a nephridial pore. The various groups of mollusks have tended to modify this arrangement, with the result that gonads have their own pore; among amphineurans, for example, the sexes are usually separate, and there is one gonad with an associated pore. Gastropods show considerable variability, but generally one gonad (ovary, testis, or ovotestis—a structure combining the functional gonads of both sexes) is located in the visceral hump and connected to the outside by a remnant of the right kidney. In hermaphroditic forms, one duct carries sperm as well as eggs. The gonadal ducts of gastropod females often secrete a protective capsule around the fertilized eggs; in males, the terminal portion of the duct is sometimes contained in a copulatory organ. Pelecypods may be either monoecious or dioecious, but the gonads are usually paired. In mussels and oysters, the gonads open through the nephridial pore, but in clams the reproductive system opens independently. The cephalopods are all dioecious. The single testis or ovary releases its products into the pericardiai cavity and this, in turn, leads to a gonopore, the external opening. The oviduct of the squid is terminally modified to form a shell gland. The male system is more complex—the gonoduct leads into a seminal vesicle where a complicated torpedo-shaped sperm case (spermatophore) is secreted and contains the sperm. Spermatophores are then stored in a special structure (Needham's sac) until copulation occurs.

A remarkable characteristic of some mollusks is the ability to alter their sex. Some species are clearly dioecious; however, among the monoecious species there is considerable variability in their hermaphroditic condition. In some species, male and female gonads, although in the same individual, are independent functionally and structurally. In others, an ovotestis produces both sperm and eggs. Oysters display a third condition; young oysters have a tendency toward maleness, but, if water temperature or food availability is altered, some individuals develop into females. Later, a reversal to the male condition may occur. The sexual makeup of an entire oyster population also has a seasonal aspect; in harmony with the group, an individual may undergo several alterations in the course of a year. A similar phenomenon, called consecutive sexuality, occurs in limpets. These gastropods stack themselves in piles, with the younger animals on top. The animals on top are males with well-developed testes and copulatory organs; those in the middle are hermaphroditic; those on the bottom are females, having lost the testes and copulatory organ (penis) by degeneration. A decrease in the number of females in a stack induces males to assume female characteristics, but the transition is retarded when an excess of females is present. The degree of maleness or femaleness is probably controlled in part by environmental and internal factors.

Arthropods. The phylum Arthropoda includes a vast number of organisms of great diversity. Most arthropods are dioecious, but many are hermaphroditic, and some reproduce parthenogenetically (*i.e.*, without fertilization). The primary reproductive organs are much the same as in other higher invertebrates, but the secondary structures are often greatly modified. Such modifications depend on whether fertilization is internal or external, whether the egg or zygote (*i.e.*, the fertilized egg) is retained or immediately released, and whether eggs are provided some means of protection after they have left the body of the female. The mandibulate arthropods (*e.g.*, crustaceans, insects) include more species than any other group and have invaded most habitats, a fact reflected in their reproductive processes.

Crustaceans (*e.g.*, crabs, crayfish, barnacles) are for the most part dioecious. The primary reproductive organs generally consist of paired gonads that open through paired ventral (bottom side) gonopores. Females often have a seminal receptacle (spermatheca) in the form of an outpocketing of the lower part of the female tract or as an invagination (inpocketing) of the body near the gonopore. Males have appendages modified for clasping the female during copulation or for guiding sperm.. A number of groups have members that reproduce parthenogenetically. Branchiopods (*e.g.*, water fleas, fairy shrimp) have simple paired gonads. The female gonopore often opens dorsally (on the back side) into a brood chamber; the male gonopore opens near the anus. Males have appendages for clasping females during copulation. Ostracods, or seed shrimp, have paired, tubular gonads. The eggs may be brooded by the female, or they may be released into the water via a gonoduct aad gonopore. The terminal portion of the male gonoduct is enclosed in a single or paired penis. Many species reproduce parthenogenetically. Some experts contend that this is the only method employed, even though functional males may be present in the population. Copepods (*e.g.*, ***Cyclops)*** have paired ovaries and an unpaired testis. The terminal portion of the oviduct constitutes an ovisac for storage of eggs. The male deposits sperm in a spermatophore that is transferred to the female. Sexual dimorphism is particularly evident among parasitic copepods. Frequently, parasitic females can hardly be recognized as copepods except for the distinctive ovisacs. Males, on the other hand, are free-living and are recognizable as copepods.

Reproductive variation among arthropod groups

The hermaphroditic Cirripedia (*e.g.*, barnacles) are among the exceptions to the generalization that crustaceans are dioecious. It has been suggested that hermaphroditism in barnacles is an adaptation to their sessile, or stationary, existence, but cross-fertilization is more common than self-fertilization. The ovaries lie either in the base or in the stalk of the animal, and the female gonopore is near the base of the first pair of middle appendages (cirri). The testes empty into a seminal vesicle through a series of ducts; from the vesicle extends a long sperm duct within a penis that may be extended to deposit sperm in the mantle cavity of an adjacent barnacle. The terminal portion of the oviduct secretes a substance that forms a kind of ovisac within the mantle cavity, where fertilized eggs undergo early development. Although most barnacles are hermaphrodites, some display a peculiar adaptation in that they contain parasitic dwarf or accessory males. Dwarf males are much smaller than the host barnacle in which they live and are degenerate, except for the testes. In some species they live in the mantle cavity of hermaphroditic forms and produce accessory sperm; in other species only the female organs persist in the host animal, and the accessory male is a necessity.

Amphipods and isopods (*e.g,* pill bugs, sow bugs), like most crustaceans, are dioecious and have paired gonads. Females of both groups have a ventral brood chamber (marsupium) formed by a series of medially directed (*i.e.*, toward the body midline) plates (oostegites) in the region of the thorax, the region between head and abdomen. Many isopods are parasitic and have developed unusual sex-related activities. Certain species are parasitic on other crustaceans. After a series of molts (*i.e.*, shedding of the body covering) a parasitic larval (immature) isopod attaches to the shell of a crab. If it is the only larva to do so, it increases in size and develops into an adult female. If another larva subsequently attaches, the new arrival becomes a male. It has been demonstrated that the testes of the functioning male larva will change to ovaries if the larva is removed to a new, uninfected host. Thus, the larvae of these species apparently are intersexual and can develop into either sex. This phenomenon, reminiscent of that in mollusks, demonstrates the way in which similar adaptations have evolved in diverse groups of organisms.

The gonads of crabs and lobsters are paired, as are the gonopores. The females of many species have external seminal receptacles on the ventral part of the thorax;

those of other species have internal receptacles in the same region. In some species, seminal receptacles are absent, and the male simply attaches a spermatophore to the female. Thus, males either have appendages (gonopods) by which sperm are inserted in the body of the female or produce spermatophores for sperm transfer. The sexual dimorphism of many decapods can be altered by parasitism. An example of this is the crab that is parasitized by a barnacle. A barnacle infection in male crabs induces the secondary sex characters of the crab to resemble those of a female; however, masculinization does not occur in parasitized females. At each molt a parasitized male crab increasingly resembles a female even though the testes may be completely unaffected. Feminization results from a hormonal alteration of the parasitized crab.

Parthenogenesis in insects

Insects are rarely hermaphroditic, but many species reproduce parthenogenetically (without fertilization). The insect ovary is composed of clusters of tubules (ovarioles) with no lumen, or cavity (Figure 3). The upper portion of each ovariole gives rise to oocytes (immature eggs) that mature and are nourished by yolk from the lower portion. The oviduct leads to a genital chamber (copulatory bursa, or vagina), with which are often associated accessory glands and a seminal receptacle. Some accessory glands form secretions by which eggs become attached to a hard surface; others secrete a protective envelope around the egg. The eighth and ninth body segments are often modified for egg-laying. The paired testes consist of a series of seminal tubules that form primary spermatogonia (immature spermatozoa) at their upper ends. As the spermatogonia mature a covering is secreted around them. Eventually they enter a storage area (seminal vesicle). The terminal portion of the male system is an ejaculatory duct that passes through a copulatory organ. A pair of accessory glands, often associated with the ejaculatory duct, contributes to the semen (fluid containing sperm) or participates in spermatophore formation. The ninth body segment and sometimes the tenth bear appendages for sperm transfer. Scorpions and spiders have tubular or saclike gonads; the female system is equipped to receive and store sperm, and, in some species, the female retains the eggs long after fertilization has occurred. Male spiders may have a cluster of accessory glands associated with the terminal portion of the reproductive system for the manufacture of spermatophores, or they may have expanded seminal vesicles for the retention of sperm until copulation takes place. Often specific appendages are adapted for sperm transfer.

From R. Snodgrass, *Principles of Insect Morphology*, © 1935 by McGraw-Hill Book Co., used with permission of McGraw-Hill Book Co.

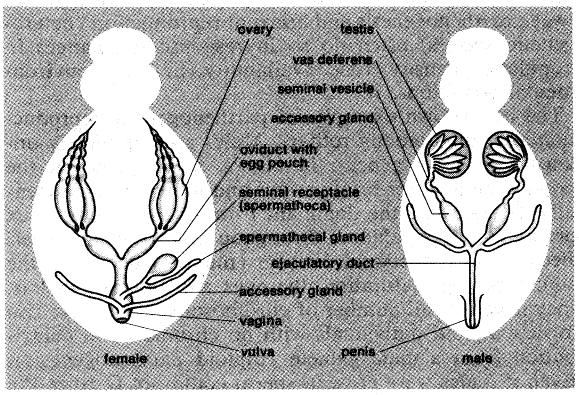


Figure 3: Reproductive organs of insects.

Echinoderms and protochordates. Echinoderms (*e.g.*, sea urchins), hemichordates (including acornworms), urochordates (*e.g.*, sea squirts), and cephalochordates (amphioxus) are restricted to a marine habitat. As with many other marine animals, their gametes are shed into the water. In echinoderms, the gonads are generally suspended from the arms directly into the sea; with few exceptions, the sexes are separate. Female starfishes have been known to release as many as 2,500,000 eggs in two hours; 200,000,000 may be shed in a season. Males produce many times that number of sperm. Acornworms reproduce only sexually, and the sexes are generally separate. The gonads lie on each side of the gut as a paired series of simple or lobed sacs. Each opens to the exterior, either directly or via a short duct. The eggs, when shed, are in coiled mucous masses, each of which contains 2,500 to 3,000 eggs.

Gonads in urochordates and cephalochordates

In urochordates and cephalochordates the gonads develop in the wall of a cavity (atrium) that receives respiratory water after it passes over the gills. Gametes are released into the cavity, then carried into the sea by the water flowing from the cavity. Most urochordates are hermaphroditic. One ovary and one testis may lie side by side, each with its own duct to the atrium; some species have many pairs of ovaries and testes. The eggs develop in so-called ovarian follicles consisting of two layers of cells, as in many vertebrates. The inner layer remains with the ovulated, or shed, egg, and the cells become filled with air spaces, which apparently help the eggs to float. Amphioxus, the highest animals lacking vertebral columns, are dioecious. They have 24 or more pairs of ovaries or testes lacking ducts. When ripe, the gonads rupture, spilling their gametes directly into the atrium.

MECHANISMS THAT AID IN THE UNION OF GAMETES

Sponges, coelenterates, flatworms, and aschelminthes. The processes of sperm transfer and fertilization have been documented for only a few species of sponges. Flagellated (*i.e.*, bearing a whiplike strand) sperm are released from the male gonad and swept out of the body and into the water by way of an elaborate system of canals. A sperm that enters another sponge, or the one from which it was released, is captured by a flagellated collar cell (choanocyte). The choanocyte completely engulfs the sperm, loses its collar and flagellum (or "whip"), and migrates to deeper tissue where the egg has matured. The choanocyte containing the sperm cell fuses with the egg, thus achieving fertilization. In freshwater coelenterates, sperm are also released into the water and carried by currents to another individual. Unlike the mechanism in sponges, however, coelenterate eggs arise in the epidermis, or surface tissue, and are exposed to sperm that may be nearby in the water; thus, no intermediate transport cell is needed. Many species of marine coelenterates expel both sperm and eggs into the water, and fertilization takes place there. Some medusoid coelenterates (jellyfish), however, offer some protection to the egg. After leaving the gonad, the egg becomes temporarily lodged in the epidermis on the underside of the organism, where fertilization and early development occur.

In all flatworms, fertilization is internal. Among species with no female duct, sperm are injected, and fertilization occurs in the inner layer of tissue. Most flatworms, however, have an elaborate system of male and female ducts. Generally, the male gonoduct passes through a penis-like organ, and sperm are transferred during copulation. In parasitic species, which often cannot find a mate, self-fertilization serves as the means for reproduction. Sperm and ova unite in the oviduct, which then secretes yolk around the zygote.

Male nematodes (roundworms) are usually equipped with a pair of copulatory structures (spicules) that guide the sperm during copulation. The posterior end of some males also exhibits a lateral (sideward) expansion (copulatory bursa) that clasps the female during copulation. Other males loop their tail around the female in the region of her gonopore. Unlike many other aschelminthes, nematodes have sperm cells that are amoeboid; *i.e.*, their cell contents seem to flow. Some male rotifers have a copulatory organ.

Annelids and mollusks. In some species of annelid polychaetes (marine worms) reproductive activity is synchronized with lunar cycles. At breeding time the body of both sexes differentiates into two regions, an anterior atoke and a posterior epitoke, in which gonads develop. When the moon is in a specific phase, the epitoke separates from the rest of the body and swims to the surface. The female epitoke apparently stimulates the male epitoke to release sperm, and sperm release, in turn, evokes expulsion of eggs. Fertilization is external. So well coordinated is this phenomenon that tremendous numbers of epitokes appear on the surface at about the same time.

Sexually mature oligochaetes have a clitellum, which is a modification of a section of the body wall consisting of a glandular, saddlelike thickening near the gonopores. During copulation, the clitellum secretes a mucus that keeps the worms paired while sperm are being exchanged. Following copulation, the clitellum secretes substance for a cocoon, which encircles the worm and into which eggs and sperm are deposited. The worm then manipulates the cocoon until it slips off over the head. Thereupon, the ends of the cocoon become sealed, and fertilization and development take place inside. Many leeches also form a cocoon; but the males of some species have a penis that can be inserted into the female gonopore. In other leeches, a spermatophore is thrust into the body of the mate during copulation.

Union of gametes among mollusks is effected in a number of ways. Marine pelecypods synchronously discharge sperm and eggs into the sea; some freshwater clams are apparently self-fertilizing. One of the more unusual types of reproductive diversity occurs in marine gastropods of the family Scalidae that produce two kinds of sperm cells. A large sperm with a degenerate nucleus acts as a transport cell for carrying numerous small fertilizing sperm through the water and into the oviduct of another individual. Cephalopod males have modified arms for the transfer of spermatophores. The right or left fourth arm of the squid, for example, is so modified. Following an often elaborate courtship, the male squid uses the modified appendage to remove spermatophores from their storage place in his body and place them in the mantle cavity of the female. A cementing substance, which is released from the spermatophore, firmly attaches the spermatophore to the female's body near the oviduct. In some species, the male loses the arm. Manipulation of the eggs by the female's arms may also occur.

Sexual behaviour among mollusks

Some unusual behaviour patterns have evolved in conjunction with sperm transfer in mollusks. Prior to copulation of certain land snails, a dart composed of calcium carbonate is propelled forcefully from the gonopore of each of the mating individuals and lodges in the viscera of the mate. Even though the snails have assumed a mating posture, sperm transfer cannot occur until each snail has been stimulated by a dart.

Arthropods. Arthropods are as varied as mollusks in their methods of effecting union of sperm and eggs. They have relatively few devices for sperm transfer, but many display a high degree of behavioral complexity.

The male and female scorpion participate in a courtship ritual involving complicated manoeuvres. In some species the male produces spermatophores that are anchored to the ground. In the course of the ritual dance the female is positioned over the spermatophore. The male then presses her down until the sperm packet is forced into her genital chamber, where it becomes attached by means of small hooks. Thus, ultimately, fertilization takes place internally.

Among some spiders the male's pedipalp, a grasping or crushing appendage, contains a bulb and an extensible, coiled structure (embolus). As mating begins, the male dips the pedipalp into semen from his gonopore. The embolus is then placed in the female gonopore, and the sperm are transferred to her seminal receptacle. The female deposits the sperm along with her eggs into a silken cocoon, which she attaches to her body or to an object such as a rock or a leaf.

Sperm transfer in copepods, isopods, and many decapods, often preceded by courtship, is effected by modified appendages, gonopods, or spermatophores. Copepods clasp the female with their antennae while placing a spermatophore at the opening of the seminal receptacle. In some decapods fertilization occurs as eggs are being released into the water.

Fertilization among insects is always internal; there is much variation in the manner in which sperm are transferred to the female. Males of some species form spermatophores that are deposited in a copulatory bursa (vagina) of the female; the wall of the spermatophore breaks down, and the sperm swim to the seminal receptacle. In other species, sperm are introduced directly into the seminal receptacle by an intromittent organ. In still others, sperm, but no spermatophores, are deposited in the copulatory bursa and migrate to the seminal receptacle. In all instances, sperm are retained in the seminal receptacle until after fertilization. An exception to the usual route of sperm transfer occurs in insects that inject sperm into the female's hemocoel (*i.e.*, the space between the body organs). The sperm then migrate to the ovary and oviduct and unite with eggs before the eggshell is formed.

PARTHENOGENESIS

Most frequently, parthenogenesis is the development of a new individual from an unfertilized gamete. Often referred to as unisexual reproduction, it has been observed in almost every major invertebrate group, with the exception of protochordates (including hemichordates), and frequently occurs alternately with bisexual reproduction (reproduction by union of gametes). Some species, in which males are completely unknown, apparently reproduce only by parthenogenesis. Species that alternate between parthenogenesis and bisexual reproduction (heterogenetic species) often do so in response to changes in population density, food availability, or other environmental conditions.

Examples of parthenogenetic reproduction

The best known examples of parthenogenetic reproduction are found among rotifers. Males are completely unknown in some genera; in others, they appear in the population only for brief periods and more or less seasonally. Females are the dominant form or are the only sex present in a population throughout most of the year. Because no reductional division (meiosis) occurs in the course of egg maturation, the eggs are diploid—that is, they have the full number of chromosomes; they give rise to new diploid individuals with no chromosomal contribution from a male gamete (diploid parthenogenesis). Even if males were present, sperm could not fertilize the eggs because the latter are already diploid. Under conditions of environmental stress such as seasonal changes, some females form eggs that undergo reductional division, resulting in eggs with the haploid number of chromosomes; such eggs must be fertilized by a male gamete to produce a new female. When the new individual matures, it will probably reproduce parthenogenetically. If, however, there are no males in the population, the haploid eggs can develop into haploid males (haploid parthenogenesis), which then participate in bisexual reproduction. Bisexually produced eggs are often referred to as winter eggs since they have a thick covering that protects the embryo during adverse environmental conditions. Summer eggs, produced parthenogenetically, are thin shelled. Bisexual reproduction occurs, therefore, only often enough to ensure survival of the species.

Nematodes, especially free-living species such as some dioecious soil nematodes, exhibit a type of parthenogenesis known as gynogenesis. In this type of reproduction, the sperm produced by males do not unite with the haploid female egg but merely activate it to begin development. The result is haploid females.

Parthenogenesis, which apparently occurs only rarely in the annelids and mollusks, is found more frequently among the arthropods. The cladocerans (*e.g.*, water fleas), for example, have a reproductive cycle much like that of rotifers—so long as environmental conditions are optimal and food is plentiful, females produce other females by diploid parthenogenesis. When conditions become adverse, males begin to appear in the population, and bisexual reproduction follows. The precise trigger for the appearance of males is not yet known. Fertilized eggs, covered with a highly resistant case, enter a resting stage (ephippium) and can withstand severe temperatures and drying out. The return of favourable conditions leads to the emergence of females that reproduce parthenogenetically. The ability to form a resting stage regulates population density. Whenever the food supply becomes short because of overpopulation by parthenogenetic females, bisexual reproduction is induced, and a dormant stage ensues. During periods of food shortage, the excess females die from lack of food, but the ephippia remain to restore the population.

Insects provide numerous examples of parthenogenesis of varying degrees of complexity. One of the most notable is that of the honeybee. Unfertilized eggs develop into drones, which are males. Fertilized eggs become worker females, which are kept in a nonreproductive state by secretions from the reproductive female, the queen bee.

Life cycles involving alternation between parthenogenesis and bisexual reproduction can be found in many species of Homoptera and Diptera (flies). Aphids (Homoptera) have a seasonal cycle consisting of a bisexual winter phase and a parthenogenetic summer phase; some species spend each phase on a different host plant. Temperature change, length of day, and food availability play major roles in initiating the phases. In the midge, a type of fly, the bisexual phase occurs in adults, and parthenogenesis takes place among the larvae (paedogenesis). Adult female midges deposit fertilized eggs, from which hatch larvae whose ovaries develop while the rest of the body retains a larval form. The ovaries of the larvae release eggs that enter the larval hemocoel (the space between body organs), where they undergo development while feeding on larval tissue. When sufficiently developed, the parthenogenetically produced young emerge either as larvae that continue parthenogenetic reproduction, forming larvae like themselves, or as male or female larvae that mature to become bisexually reproducing adults.

PROVISIONS FOR THE DEVELOPING EMBRYO

Differences in amount of yolk

Invertebrates have developed a great many methods for protecting the fertilized egg and young embryo and for providing nutrients for the developing young. This is especially true of freshwater and terrestrial forms. Sponges and freshwater coelenterates, exposed to seasonal drying out, provide a tough covering for the eggs that prevents water loss. Many turbellarians envelop the eggs with a capsule and attach it to a hard surface, where it remains until the young emerge. Other turbellarians retain encapsulated eggs in the body until development is complete and the young emerge. All parasitic flatworms enclose their eggs in a protective capsule within which development occurs after it has left the parent's body. Most nematodes and rotifers do likewise, but a few species are ovoviviparous; *i.e.*, the egg hatches in the mother's body. In many forms the amount of yolk provided in the egg and the nature of the egg capsule are correlated with annual seasons—summer eggs generally have less yolk and thinner capsules than do winter eggs. This is true also in a number of crustaceans. Freshwater and terrestrial annelids provide a cocoon for their young and often deposit it in a moist place. One group of leeches, however, does not form a cocoon; instead, the egg, surrounded by a protective membrane, is attached to the underside of the parent. As the young develop, the adult leech undulates its body so that water currents flow over the young. Presumably this serves as a means of aeration. Mollusks that live in freshwater may provide a protective covering for the eggs, or the eggs may be brooded by the female. Some pelecypods (bivalves) release mature eggs into their gill chambers; here the eggs are fertilized, and embryonic development is completed in a protected location. Cephalopods (*e.g.*, squid, octopus) attach the eggs to a surface, then continuously force jets of water over the egg masses, thereby keeping them free of debris and perhaps aerating them. Some echinoderms also brood the eggs until the young emerge.

Arthropods have a particularly wide range of methods for ensuring offspring survival. Brood pouches, common in branchiopods, isopods, and amphipods, are sometimes part of the carapace, or back plate. In other instances, expanded plates on the lower side (sternum) form the pouches. Crayfish cement the fertilized eggs to their swimmerets (modified appendages) and carry them about as they are brooded by the female. The most elaborate provisions for the embryo are found among terrestrial arthropods, especially insects. Although some species simply deposit their eggs and abandon them, many retain the encapsulated egg within the body during early development. Some are viviparous; that is, they Gear living young. The eggs of certain species of scorpions have little or no yolk; the embryo is nourished by the parent in a manner similar to that in mammals—part of the scorpion oviduct becomes modified as a uterus for the embryo; another part lies close to the female's gut and absorbs nutritive substances that are conveyed to the developing young. A similar arrangement has evolved in some insects. Other viviparous insects nourish the larvae ky glandular secretions from the uterine lining.

III. Reproductive systems of vertebrates

GONADS, ASSOCIATED STRUCTURES, AND PRODUCTS

The reproductive organs of vertebrates consist of gonads and associated ducts and glands. In addition, some vertebrates, including some of the more primitive fishes, have organs for sperm transfer or ovipository (egg-laying) organs. Gonads produce the gametes and hormones essential for reproduction. Associated ducts and glands store and transport the gametes and secrete necessary substances. In addition to these structures, most male and female vertebrates have a cloaca, a cavity that serves as a common terminal chamber for the digestive, urinary, and reproductive tracts and empties to the outside. In lampreys and most ray-finned fishes in which the cloaca is small or absent, the alimentary canal has a separate external opening, the anus. In some teleosts the alimentary, genital, and urinary tracts open independently. Hagfishes, which are closely related to the lampreys, have a short cloaca. In many vertebrates other than mammals, especially reptiles and birds, the cephalic, or head, end of the cloaca is partitioned by folds into a urinogenital chamber (urodeum) and an alimentary chamber (coprodeum) that open into a common terminal chamber (proctodeum). Above monotremes (*e.g.*, platypus, echidna) the embryonic cloaca becomes completely partitioned into a urinogenital sinus conveying urine and the products of the gonads, and an alimentary pathway; the two open independently to the exterior.

Embryonic development of gonads

Gonads arise as a pair of longitudinal thickenings of the coelomic epithelium and underlying mesenchyme (unspecialized tissue) on either side of the attachment of a supporting membrane, the dorsal mesentery, to the body wall. At first, gonadal ridges bulge into the coelom and are continuous with the embryonic kidney. The germinal epithelium covering the gonadal ridges gives rise to primary sex cords (medullary cords) that invade the underlying mesenchyme. These cords establish within the gonadal blastema (a tissue mass that gives rise to an organ) a potentially male component, the medulla. Secondary sex cords grow inward, spreading just beneath the germinal epithelium to form a cortex. If the gonad is to become a testis, only the medullary component differentiates. If the gonad is to become an ovary, only the cortex differentiates (see DEVELOPMENT, ANIMAL).

The length of an adult gonad depends, in part, upon the extent of gonadal-ridge differentiation. In cyclostomes (lampreys and hagfish), elasmobranchs (sharks, skates, and rays), and teleosts most of it differentiates, and the gonads extend nearly the length of the body trunk. In tetrapods (amphibians, reptiles, birds, and mammals), the cranial portion, at the anterior end, generally does not differentiate; in toads only the more caudal, or posterior, portion does so. The middle segment in toads of both sexes gives rise to a Bidder's organ containing immature eggs. In anurans (frogs and toads) and some lizards of both sexes, one segment of the gonadal ridge gives rise to yellow fat bodies that, especially in anurans, diminish in size just prior to the breeding season. In mammals, only the middle portion of the gonadal ridge differentiates.

Some vertebrate species have only one gonad, which may lie in the midline or on one side; the condition is more common among females. Adult cyclostomes of both sexes have one gonad. In lampreys it is in the middle of the body; in hagfishes it is on the right side. Birds are the only other major group of vertebrates in which most females have one gonad, the right ovary being typically absent. Male birds have a pair of testes, however. Exceptions to the condition of single ovaries among birds include members of the falcon family, in which more than 50 percent of mature hawks have two well-developed ovaries. In all bird species a small percentage of females probably have two ovaries; reported instances include owls, parrots, sparrows, and doves, with estimates for doves ranging from 5 percent to 25 percent. A few teleosts and viviparous elasmobranchs have only one ovary; in sharks the right one is usually present, in rays, the left. In amniotes (*i.e.*, reptiles, birds, and mammals) unpaired gonads are unusual. Some lizards have one testis, and some female crocodiles have one ovary. Among mammals, the platypus usually has only a left ovary, and some bat species (family Vespertilionidae) have only the right.

Theories about unpaired gonads

One of two explanations may account for unpaired gonads: the paired embryonic gonadal ridges may fuse to form a median gonad—as in lampreys and the perch--or only one gonadal ridge may receive immigrating primordial germ cells (immature sperm or eggs), with the result that the opposite gonad does not develop—as in chickens and ducks. Both gonadal ridges have been reported to exhibit an equal number of primordial germ cells in embryonic hawks, and these typically have two ovaries.

Among lower vertebrates, mature gonads sometimes produce both sperm and eggs. Hermaphroditism is more general in cyclostomes and teleosts than in other fishes. A teleost may function as a male during the early part of its sexual life and as a female later. In some teleost families sperm and eggs mature simultaneously but in different regions of the gonad. These fish normally function as males during one season and as females the next. Cyclostomes generally are ambisexual during juvenile life—*i.e.*, immature male and female sex cells exist side by side, or, as in *Myxine*, the anterior part of the immature gonad may be ovary and the caudal part, the testis. It is thought that cyclostomes normally become unisexual at maturity. Hermaphroditism is uncommon among amphibians, although it frequently occurs as an anomaly. In vertebrates above amphibians, true hermaphroditism probably does not exist.

Both male and female duct systems are occasionally absent. In cyclostomes, a few elasmobranchs, and some teleosts, such as salmon, trout, and eels, the gametes are propelled toward the posterior within the coelom, often by cilia (minute hairlike structures), and exit via a pair of funnel-like genital pores near the base of the tail. In cyclostomes, the pores lead to a sinus, or cavity, within a median papilla (*i.e.*, a fingerlike structure) and are open only during breeding seasons.

Seminiferous tubules

Male systems. Testes. In anurans, amniotes (reptiles, birds, and mammals), and even some teleosts, testes are composed largely of seminiferous tubules--coiled tubes, the walls of which contain cells that produce sperm—and are surrounded by a capsule, the tunica albuginea. Seminiferous tubules may constitute up to 90 percent of the testis. The tubule walls consist of a multilayered germinal epithelium containing spermatogenic cells and Sertoli cells, nutritive cells that have the heads of maturing sperm embedded in them. Seminiferous tubules may begin blindly at the tunic, or outermost tissue layer, and pass toward the centre, becoming tortuous before emptying into a system of collecting tubules, the rete testis. Such an arrangement is characteristic of frogs. In certain amniotes—the rat, for example—the tubules may be open ended, running a zigzag course from the rete to the periphery and back again. The average length of such tubules is 30 centimetres (12 inches), and they seldom communicate with each other. In many mammals the tubules are grouped into lobules separated by connective-tissue septa, or walls. The arrangement permits the packing of an extensive amount of germinal epithelium into a small space. In immature males and in adult males between breeding seasons, the tubules are inconspicuous and the epithelium is inactive; in some species, however, spermatogenesis, or production of sperm, proceeds at a variable pace throughout the year. An active epithelium may exhibit all stages of developing sperm. The lumen, or tubule cavity, contains the tails of many sperm (the heads of which are embedded in Sertoli cells), free sperm, and fluid that is probably resorbed. In mammals, in any single zone along a tubule, all sperm are at the same stage of maturation; adjacent zones contain different generations of sperm, and a period of sperm formation and discharge is followed by an interval of inactivity.

In cyclostomes, most fishes, and tailed amphibians the germinal epithelium is arranged differently. Instead of seminiferous tubules there are large numbers of spermatogonial cysts (also called spermatocysts, sperm follicles, ampullae, crypts, sacs, acini, and capsules) in which sperm develop, but in which the epithelium is not germinal. Spermatogenic cells migrate into the cysts from a permanent germinal layer, which, depending on the species, may lie among cysts at the periphery of the testes or in a ridge along one margin of the testis. After invading the thin nongerminal epithelium of a cyst, spermatogenic cells multiply, producing enormous numbers of sperm. The cysts become greatly swollen and whitish in colour; the entire testis also swells and has a granular appearance. As sperm mature, they separate from the epithelium and move freely in the cystic fluid. Finally, the cysts burst, and the sperm are shed into ducts. In the case of cyclostomes and a few teleosts the sperm are shed into the coelom. The cysts, totally emptied, collapse. Then either they are replaced by new ones, or they become repopulated by additional spermatogenic cells. It is not yet known which of these processes occurs.

Testicular stroma, which fills the spaces between seminiferous tubules or spermatogenic cysts, consists chiefly of connective tissue, blood and lymphatic vessels, and nerves; it is more abundant in some vertebrates than in others. Leydig (interstitial) cells, which are undifferentiated connective-tissue cells, are also present in most, if not all, vertebrates. Thought to be a primary source of androgens, or male hormones, Leydig cells are not always readily distinguishable, and, in some bird species, they may be seen only with the electron microscope. The capillary system of the rat testis, and probably that of many other vertebrates, is such that blood that has bathed the Leydig cells flows to the tubules; it is thus probable that Leydig cell hormones have an immediate effect on the germinal epithelium.

Location of testes in mammalian orders

Testes in vertebrates below mammals lie within the body. This is also true of many, sometimes all, members of the mammalian orders Monotremata, Insectivora, Hyracoidea, Edentata, Sirenia, Cetacea, and Proboscidea. Some male mammals—most marsupials, ungulates, carnivores, and primates after infancy—have a special pouch (scrotum) that the testes occupy permanently. A few mammals have a pouch into which the testes descend and from which they can be retracted by muscular action. These include a few rodents such as ground squirrels; most, if not all, bats; and some primitive primates (loris, potto). The scrotum consists of two scrota1 sacs, each connected to the abdominal cavity by an inguinal canal lined with the peritoneal membrane. The canals are the path of descent (and retraction) of the testes to the sacs. In descending, the testes carry along a spermatic duct, blood and lymphatic vessels, and a nerve supply wrapped in peritoneum and constituting, collectively, the spermatic cord. Rabbits, most rodents, and some insectivores, which lack scrotal sacs, have instead a wide inguinal canal into which the testes may be drawn and from which they are retracted when in danger of injury. In these mammals, descended testes cause a temporary bulge in the perineal region (*i.e.*, between the anus and the urinogenital opening). In a small number of mammals, the testes permanently occupy the perineal location.

The scrotum is a temperature-regulating device. Warm blood approaching the testis comes close to the vessels carrying cool blood leaving the testis, so that the blood approaching the testis is cooled; the vessels form an intricate vascular network (pampiniform plexus) within the spermatic cord. Failure of both testes to enter the scrotal sacs (cryptorchidism) results in permanent sterility. In cold weather two sets of muscles, the dartos and cremasteric, pull the testes close to the body. The dartos lies between the two scrotal sacs and is attached to the scrotal skin. The cremaster, wrapped around the spermatic cord, is an extension of the abdominal wall musculature. It retracts the testis. Birds, like mammals, are homoiothermic (warm-blooded), and their testes are near air sacs (extensions of incurrent respiratory tubes). Air in the sacs may help regulate the temperature of the testes.

Ducts. The male duct system begins as the rete testis, a network within the testis of thin-walled ductules, or minute ducts, that collects sperm from the seminiferous tubules. The rete is drained by a number of small ducts—usually fewer than ten—called the vasa efferentia, which are modified kidney tubules. In some fishes and amphibians the vasa efferentia connect the testes with the cranial (anterior) end of the kidneys (Figure 4). In anamniotes (*e.g.*, fish and amphibians), therefore, except teleosts, the ducts that drain the kidneys usually drain the testes also. In most amphibians these ducts pass caudad, or posteriorly, to empty independently into the cloaca; in some fishes they pass through a median urinogenital papilla.

(Left and right) From George C. Kent Jr., Comparative Anatomy of the Vertebrates. 2nd ed. (1969); The C.V. Mosby Co., St. Louis. (left) redrawn from C.J. Baker and W.W. Taylor, Journal of the Tennessee Academy of Science, vol. 39, no. 1 (1964)

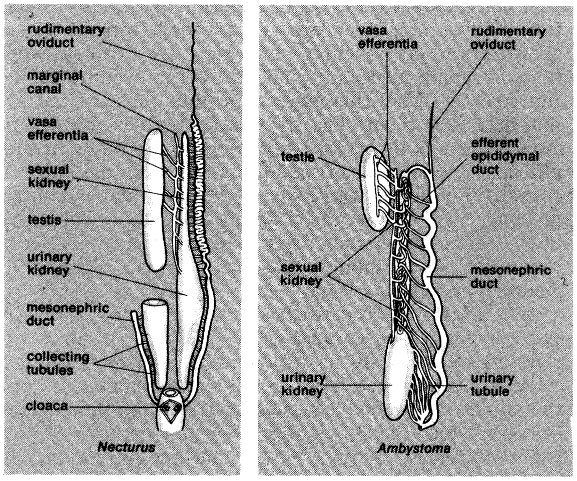


Figure 4: Reproductive systems of two male amphibians. (Left) The sexual kidney is composed of one row of modified tubules that are ciliated for sperm transport and drain into the adult (mesonephric) kidney duct. This duct also collects urine from the kidney. (Right) The mesonephric duct collects only sperm.

Although drainage of the testis and the kidney by the same duct is a basic pattern, there has been a tendency in many vertebrates toward separate spermatic and urinary ducts. This tendency is manifested in one of two ways among anamniotes. In many sharks and in some amphibians (Plethodontidae, Salamandridae, Ambystomatidae), the embryonic kidney duct ultimately drains the testis, and one or more new ducts (ureters) drain the adult kidney. On the other hand, in the primitive fish Polypterus and in most teleosts, the embryonic kidney duct drains the adult kidney, and a new duct arises to drain the testis. Many degrees of separation of the two ducts occur in anamniotes, from the condition of the sturgeon, in which the spermatic duct unites with the urinary duct far toward the head, to the condition in Esox (a pike), in which spermatic and urinary ducts empty independently to the exterior.

In amniotes, the mesonephric kidney is a temporary structure confined to the embryo, but the mesonephric duct persists in the adult male as a sperm duct. A separate ureter drains the adult kidney. The spermatic and urinary ducts empty independently into the cloaca except in mammals above monotremes, in which they are confluent with the urethra. The epididymis of amniotes, a highly tortuous duct draining the vasa efferentia, usually serves as a temporary storage place for sperm; it is small in birds and large in turtles. In mammals, the first part of the epididymis consists of a head, body, and tail that wrap around the testis; it gradually straightens to become the spermatic duct. The epididymis secretes substances that prolong the life of stored sperm and increase their capacity for motility.

In all vertebrates certain regions of the spermatic duct are lined by cilia and a variety of secretory epithelial cells. One end may enlarge to form a sperm reservoir or secrete seminal fluid. In the catfish Tmchycorystes *mirabilis* secretions of the spermatic duct form a gelatinous plug in the female similar to the vaginal plug of mammals. A seminal glomulus in birds functions as a sperm reservoir. In some mammals an enlargement of the spermatic duct called the ampulla contributes to the seminal fluid and stores sperm. Small mucous glands (of Littré) and other glandular structures open into the urethra along its length. Cloacal glands in basking sharks and many salamanders form a jelly that encloses sperm in a spermatophore. Cloacal glands of some lizards produce secretions called pheromones. The siphon sac of elasmobranchs is one of the few accessory sex glands that is a separate organ in animals below mammals. It extends as an elongated pocket into the pelvic fin and secretes a nutritive mucus that enters the female reproductive tract with sperm.

Accessory glands. Accessory sex glands that are conspicuous outgrowths of the genital tract are almost uniquely mammalian. The major mammalian sex glands include the prostate, the bulbourethral, and the ampullary glands, and the seminal vesicles. All are outgrowths of the spermatic duct or of the urethra (Figure 5) and all four occur in elephants and horses and in most moles, bats, rodents, rabbits, cattle, and primates. A few members of these groups lack ampullary glands, or ampullary glands and seminal vesicles. Cetaceans (whales, porpoises) have only the prostate, as do some carnivores, including dogs, weasels, ferrets, and bears.

The prostate gland

The prostate, the most widely distributed mammalian accessory sex gland, is absent only in Echidna (a marsupial) and a few carnivores. It empties into the urethra by multiple ducts. Many rodents, insectivores, and lagomorphs have three separate prostatic lobes; in a few mammals (some primates and carnivores) the prostate is a single mass with lobules and encircles the urethra at the base of the bladder. In a few mammals (*e.g.*, opossum), the prostate is not a compact mass but a partly diffuse gland. In many rodents (*e.g.*, rat, guinea pig, mouse, hamster) and some other mammals, the semen coagulates quickly after ejaculation as a result of a secretion from a male coagulating gland, which is usually considered part of the prostatic mass. Coagulated semen forms a vaginal plug that temporarily prevents copulation.

From George C. Kent Jr., Comparative Anatomy of the Vertebrates, 2nd ed. (1969); The C.V. Mosby Co., St. Louis

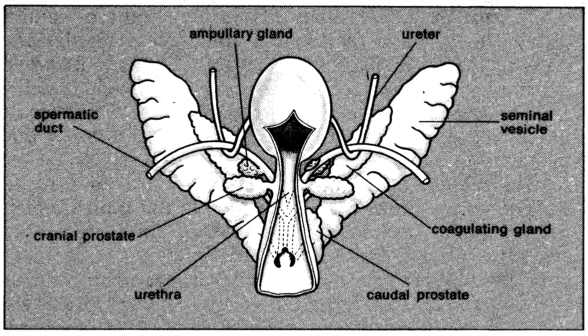


Figure 5: Accessory sex organs of a male golden hamster. Bulbourethral glands arise more toward the tail end and are not shown. The bladder and urethra have been opened to show entrances of ducts.

Bulbourethral (Cowper's) glands arise from the urethra near the penis and are surrounded by the muscle of the urethra or penis. Typically, there is one pair, but as many as three (marsupials) may be found. The glands, small in man, large in rodents, elephants, and some ungulates including pigs, camels, and horses, are absent in cetaceans, mustelids (*e.g.*, mink, weasel), sirenians (manatees, dugongs), pholidotans (pangolins), some edentates, and carnivores such as walrus, sea lion, bear, and dog.

Although many mammals have an ampullary swelling on the spermatic duct near the urethra, only a small number form a separate ampullary gland as an outgrowth of the duct. It is very large in some bats, absent in many mammalian orders, and variable in the rest. Although common in rodents, it is absent in guinea pigs and some strains of mice.

Seminal vesicles are paired, typically elongated and coiled fibromuscular sacs that empty into either the spermatic duct or the urethra. Absent in monotremes, marsupials, carnivores, cetaceans, and in some insectivores, chiropterans, and primates, seminal vesicles are exceptionally large in rhesus monkeys and small in man. They are absent in domesticated rabbits, small or rudimentary in cottontails, large in armadillos, and variable in sloths. They contribute the sugar fructose and citric acid to the semen but do not serve as sperm reservoirs.

Female systems. Ovaries. Ovaries lie within the body cavity and are suspended by a dorsal mesentery (mesovarium), through which pass blood and lymph vessels and nerves. Primitive vertebrate ovaries occur in the hagfish, in which a mesentery-like fold of gonadal tissue stretches nearly the length of the body cavity. Unique in the hagfish is the fact that functional ovarian tissue occupies only the forward half of the gonadal mass, the rear part containing rudimentary testicular tissue. In most fishes except very primitive forms, the ovaries are similarly elongated. In tetrapods other than mammals, the ovaries are usually confined to the middle third or half of the body cavity, particularly during nonbreeding seasons. The ovaries of mammals undergo moderate caudal displacement, finally coming to lie between the kidney and the pelvis.

Characteristicsof the ovary

The appearance of an ovary depends on many factors —*e.g.*, whether one egg or thousands are discharged (ovulated); whether the eggs are immature or ripe; whether mature eggs are small or large; or whether pigments occur in the egg cytoplasm, such as those responsible for yellow yolk. Other factors also affect the appearance of the ovary: the season of the year in seasonal breeders (the ovary enlarges during breeding seasons, diminishes in size between seasons); the age of the animal (whether juvenile, reproductively active, or senile, particularly in birds and mammals); and the fate of ovulated, or discharged, egg follicles, or sacs.

The ovaries are covered with a germinal epithelium that is continuous with the peritoneum lining the body cavity. The term germinal epithelium is inappropriate because in most adults it contains no germ cells, these having moved deeper into the ovary. In hagfishes and amphibians, cells that give rise to eggs are known to occur in the germinal epithelium, and it may be that the germinal epithelium in a few other vertebrates contains similar cells. The germinal epithelium undergoes cell division, however. This is particularly true of species in which enormous expansion of the ovary occurs each breeding season. Beneath the epithelium is a layer of connective tissue, the tunica albuginea, which is much thinner than that surrounding the testes.

A typical vertebrate ovary consists of cortex and medulla. The cortex, immediately internal to the tunica albuginea, contains future eggs and, at one time or another, eggs in ovarian follicles (*i.e.*, developing eggs); it undergoes fluctuations in size and appearance that correlate with stages of the reproductive cycle. The cortex also contains remnants of ovulated follicles and, in mammals, clusters of interstitial cells that, in some species, are glandular. The cortical components are embedded in a supportive framework of connective, vascular, and neural tissue constituting the stroma. Internal to the cortex is the medulla, consisting of blood and lymph vessels, nerves, and connective tissue. The medulla, which contains no germinal elements, exhibits no significant cyclical activity, is usually inconspicuous, is continuous with the dorsal mesentery, and, in cyclostomes, is hardly distinguishable from the latter. The mammalian medulla, on the contrary, is almost completely surrounded by cortex and converges on the mesovarium (*i.e.*, the part of the peritoneum that supports the ovary) at a narrow hilus, at which nerves and vessels enter the ovary. In the medulla of the mammalian ovary near the hilus are small masses of blind tubules or solid cords—the rete ovarii—which are homologous (*i.e.*, of the same embryonic origin) with the rete testis in the male. The microscopic right ovary of birds usually consists only of medullary tissue.

Ovaries are characterized as saccular, hollow, lacunate (*i.e.*, compartmented), or compact. The ovary of many teleosts, especially viviparous ones, contains a permanent cavity, which is formed during ovarian development when an invagination of the ovarian surface traps a portion of the coelom. The cavity is therefore unique in that it is lined by germinal epithelium. The lining develops numerous ovigerous folds that project into the lumen and greatly increase the surface area for proliferation of eggs. In most other teleosts, a temporary ovarian cavity develops after each ovulation, when the shrinking cortex withdraws from the outside ovarian wall along one side of the ovary. The resulting cavity is obliterated as eggs of the next generation enlarge. The permanent and temporary cavities of teleost ovaries and a similar cavity in garfish ovaries are continuous with the lumen of the oviduct, and eggs are shed into them. The ovaries of other fishes lack cavities and are characterized as compact. The amphibian ovary, which contains six or more central, hollow sacs that give it a lobed appearance, is characterized as saccular. The sacs are formed when the embryonic medullary and rete cords become hollow and coalesce. Maturing eggs bulge into the sacs but are not shed into them. The ovaries of reptiles, birds, and monotremes have cavities homologous to those in amphibians; the number of medullary spaces in the adults is considerably larger, however, so that the ovaries contain an extensive network of fluid-filled cavities (lacunae). Such ovaries are characterized as lacunate. The ovaries of mammals above monotremes are compact, having no medullary cavities.

Formation and fate of ovarian follicles

An ovarian follicle consists of an oocyte, or immature egg, surrounded by an epithelium, the cells of which are referred to variously as follicular, nurse, or granulosa cells. In cyclostomes, teleosts, and amphibians, the epithelium is one layer thick. In the hagfish and those vertebrates in which the oocyte receives heavy deposits of yolk (elasmobranchs, reptiles, birds, and monotremes), the epithelium appears to be two cells thick, apparently the result of layering of nuclei in a simple columnar epithelium (*i.e.*, epithelium consisting of relatively "tall" cells). Above monotremes the follicular epithelium appears to be many cells thick; in at least one species, however, this is considered an artifact, and all granulosa cells are said to extend between the outer boundary of the epithelium and the oocyte.

The follicular epithelium originates as a few flattened cells derived from the germinal epithelium. Primary follicles are usually situated just under the tunica albuginea; secondary follicles lie deeper in the cortex. The primitive role of the follicular cells appears to be the secretion of the yolk-forming material onto or into the oocyte. Evidence from mammals indicates that the follicular cells may also have a role in converting substances produced elsewhere into female hormones, or estrogens. In some hibernating bats the granulosa cells are filled with glycogen, or animal starch, which may be a source of energy. Mammalian follicles above monotremes are unique in that they develop a fluid-filled cavity (antrum) within the granulosa layer. During antmm formation cell division of the granulosa cells increases, and fluid-filled spaces develop among the cells. The spaces coalesce to form the antrum. Under the influence of pituitary gonadotropic hormones, many antral follicles thereafter continue to grow, forming large so-called Graafian follicles—less than 400 microns, or 0.4 millimetre (0.02 inch), in diameter in large mammals, 150–200 microns, or 0.15–0.2 millimetre (0.006–0.008 inch), in small ones. Graafian follicles contain mature eggs and appear as large blisters on the ovary. At this stage the ovum, suspended within the fluid of the antrum (liquor folliculi) by a slender stalk of granulosa cells, is surrounded by a cluster of these cells, the cumulus oophorus, or discus proligerus. The remaining follicular cells form a thin wall surrounding the antrum. Antra are lacking in a few insectivores (Hemicentetes, *Euriculus*) because the granulosa cells swell and multiply to form corpora lutea, masses of yellow tissue. In the bat Myotis the antrum is likewise compressed and disappears just before discharge of the egg, or ovulation.

In all vertebrates, oocytes that have begun to grow and mature may, at any time until just before ovulation, cease development and undergo atresia, or degeneration. This is a normal process that reduces the number of eggs ovulated. In small laboratory rodents, atresia takes place in 50 percent of the Graafian follicles in each ovary one or two days before ovulation, thus reducing the number of ovulatable eggs by 50 percent. A similar reduction takes place in hagfish prior to ovulation. Atretic follicles eventually become lost in the stroma of the cortex of the ovary. In mammals especially, follicles lacking oocytes and antra, called anovular follicles, as well as polyovular follicles (*i.e.*, containing more than one oocyte), occasionally occur.

The theca

The ovarian follicle of vertebrates, commencing with hagfish, is surrounded by a theca, or sheath, composed of two concentric layers of stromal cells. The outer layer (theca externa) is chiefly connective tissue but may contain smooth muscle fibres. The inner layer (theca interna) has more blood vessels and, in vertebrates that produce heavily yolked eggs, the largest vessels carry venous blood. In these species the cell membranes of the oocyte and granulosa cells have many microvilli (*i.e.*, fingerlike projections), which probably facilitate transport of substances important in yolk formation from the blood vessels to the egg. Mature follicles in the marsupial Dasyatus are said to lack theca, and in some bats only one thecal layer has been described.

During the growth phase, eggs in species with massive amounts of yolk may increase in size 106 (1,000,000) or more times as a result of vitellogenesis (deposit of yolk). In goldfish, on the other hand, when vitellogenesis commences, the egg has a diameter of 150 microns (0.15 millimetre [0.006 inch]); that of the mature egg is only 500 microns (0.5 millimetre [0.02 inch]). Mammalian eggs contain little yolk and vary little in size. Oogonia (*i.e.*, cells that form oocytes) of the golden hamster average 15 microns (0.015 millimetre [0.0006 inch]) in diameter, and eggs in Graafian follicles average 70 microns (0.07 millimetre [0.003 inch]). The mature eggs of horses and humans are approximately the same size—somewhat less than 150 microns. In seasonally breeding oviparous fishes and amphibians, all eggs are usually in the same stage of development, and the ovary grows to a mature state quite rapidly as a result of growth of the eggs, which frequently number more than 1,000,000. Such ovaries distend the body wall when mature; following spawning, the ovaries shrink rapidly to inconspicuous bodies consisting mainly of oogonia, immature oocytes, and a few stromal cells. In reptiles and birds, ovarian weight also is high in proportion to body weight during egg-laying seasons. The weight of the ovary of the starling, for example, may increase from eight milligrams in early winter to 1,400 milligrams immediately before ovulation. The mature eggs of reptiles and birds are unique in that they are suspended from the ovary by a short stalk (pedicle). The stalk contains a cortex with additional oocytes in various stages of development and extensions of vessels and nerves. Full growth of the follicle in reptiles and birds requires only a few days or weeks (nine days in the domestic hen). In mammals, the ratio of ovarian weight to body weight varies insignificantly throughout the reproductive life of the female, and follicles in many stages of development are constantly present.

Vertebrate eggs are almost universally shed into the coelom or into a subdivision thereof, from which they enter the female reproductive tract. Even in those teleosts in which the eggs are shed into an ovarian cavity, the latter is often of coelomic origin. In many mammals a membranous sac of peritoneum, the ovarian bursa, traps part of the coelom in a chamber along with the ovary. The bursal cavity (periovarian space) may be broadly open to the main coelom, completely closed off from the coelom, or in communication with the coelom by a narrow, slitlike passage. The bursa, moderately developed in lower primates and catarrhines (Old World monkeys), is poorly developed in man. In horses, one edge of the ovary contains a long groove (ovulation fossa) into which all eggs are shed; the groove is found in a cleftlike ovarian bursa. The ovarian bursa increases the probability that all ovulated eggs will enter the oviduct.

Ovulation

The process of ovulation has been described for all vertebrate classes. Elasmobranchs, reptiles, and birds have massively yolked eggs. As ovulation approaches, the fimbria (*i.e.*, frills, or fringes) of the membranous and muscular funnel surrounding the entrance to the oviduct wave in a gentle, undulating motion. An egg that is nearly free of the ovary is grasped and partially encompassed by the fimbria; when the egg is freed, the fimbria draw the egg into the funnel. At this time, the egg has little shape and is partly squirted and partly flows into the oviduct; never completely free in the coelom, its chances of not entering the oviduct are small. In the case of moderately or poorly yolked eggs cilia help to sweep the eggs into the ostium, or opening, of the oviduct. During ovulation in Japanese rice fish, Oryzias *latipes*, a tiny papilla, or fingerlike process, develops on the surface of a bulging mature follicle in the centre (stigma) of the follicle. The follicle becomes thin at the stigma, an aperture appears, and the egg rolls out. In rabbits this process differs only in detail. During the final 20 minutes before ovulation in rabbits, some of the tiny blood vessels surrounding the stigma rupture, and a small pool of blood forms under the apex of the cone-shaped papilla. The follicular wall shortly gives way at the apex, and follicular fluid oozes from the opening, followed soon after by the egg. The ovulated mammalian egg typically is surrounded by a layer of columnar follicular cells, the corona radiata; but it is naked in some insectivores and some marsupials. Following ovulation in all vertebrates, the ovary may become smaller, become modified for maintenance of pregnancy, or proceed to form additional eggs.

The process of ovulation in vertebrates has been documented, but the immediate causes remain to be clarified. It is almost certain that an ovulatory hormone is secreted by the pituitary gland (*i.e.*, the so-called master endocrine gland) of all vertebrates. It is highly probable that breakdown of very small fibres that bind the follicular cells together may occur at the stigma, weakening the follicular wall at that location. Hormones from the ovary and other sources may play a role, as may neurohormones, which are hormones released at nerve endings. Rhythmic contractions of the entire ovary occur at ovulation in many vertebrates and have been described in rabbits. The role of mechanical pressure within the follicle, however, is not understood. Ovulation in most mammals (spontaneous ovulators) occurs cyclically as a result of the spontaneous release of the ovulatory hormone. In a few mammals (reflex ovulators) the stimulus of copulation is essential for release of the ovulatory hormone.

Striking postovulatory changes take place in the follicles of mammals and, to lesser degrees, of lower vertebrates. Blood vessels from the theca interna invade the ovulated follicles; the granulosa cells divide, enlarge, accumulate fats, and obliterate any remnants of the collapsed antra. Thereafter, they are known as lutein cells. Theca interna cells undergo changes identical to those of the granulosa cells. The result in mammals is the formation of solid masses called corpora lutea, recognizable as prominent reddish-yellow bulges on the ovary. Corpora lutea produce the hormone progesterone, which is essential for the maintenance of pregnancy. The conversion of postovulatory follicles into structures more or less resembling mammalian corpora lutea has been demonstrated in numerous viviparous reptiles, amphibians, and elasmobranchs; in certain other fishes, including cyclostomes; and in some oviparous amphibians and reptiles. In birds, the postovulatory follicle shrinks, and identifiable corpora lutea do not develop, although some granulosa cells accumulate lipids of unknown significance.

Tracts. The female reproductive tract consists of a pair of tubes (gonoducts) extending from anterior, funnel-like openings (ostia) to the cloaca (Figure 6), except as noted below. The gonoducts are specialized along their length for secretion of substances added to the eggs; for transport, storage, nutrition, and expulsion of eggs or the products of conception; and, in species with internal fertilization, for receipt, transport, storage, and nutrition of inseminated sperm. The predominately muscular tracts are lined by a secretory epithelium and ciliated over at least part of their length. Fusion of the caudal (tail) ends of the paired ducts may occur. Gonoducts are absent in cyclostomes and a few gnathostome fishes that have abdominal pores. A few vertebrates have only one functional gonoduct.

From George C. Kent Jr.. *Comparative Anatomy of the Verfebrates*, 2nd *ed.* (1969); *The C.V. Mosby Co., St. Louis*

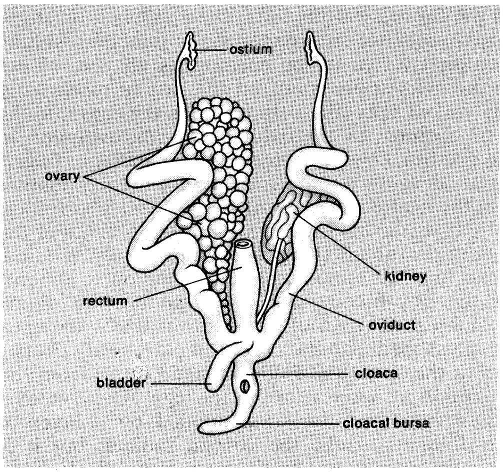


Figure 6: Reproductive tract of female turtle, *Trionyx euphraticus*. One ovary has been removed.

Gonoducts of fishes and amphibians

Gonoducts in lungfishes and amphibians are coiled muscular tubes that are ciliated over most of their length. Only occasionally do they unite caudally in a genital papilla before opening into the cloaca. During breeding seasons their diameter increases severalfold because of the highly active secretory epithelium. Between breeding seasons they are small. In some anurans (frogs, toads), such as *Rana*, the lower end of each gonoduct is expanded to form an ovisac, in which ovulated eggs are stored until spawning; the tube between the ostium (funnel-like opening) and ovisac is the oviduct. In viviparous amphibians the young develop in the ovisac. In amphibians, numerous multicellular glands extend deep into the lining of the female tract. Six successive glandular zones have been described in some urodeles, and these secrete six different gelatinous substances upon the egg. Female urodeles often have convoluted tubular outpocketings of the cloaca called spermatheca; they temporarily store sperm liberated from the male spermatophore.

The two gonoducts of elasmobranchs share a single ostium, a trait found only in Chondrichthyes. The ostium is a wide caudally directed funnel supported in the falciform ligament, which is attached to the liver. The role of the fimbria of the ostium at ovulation has been described (see above Ovaries). Two oviducts pass forward from the ostium to the septum transversum (*i.e.*, between the heart and abdominal cavities), curve around one end of the liver, then pass posteriorly on each side. Approximately midway between ostium and uterus each oviduct has a shell (nidamental) gland. Fertilization takes place above the shell gland, which may be immense or almost undifferentiated. Half of the shell gland secretes a substance high in protein content (albumen), and the other half secretes the shell—delicate in viviparous forms, thick and horny in most oviparous species. Horny shells may have spiral ridges and many long tendrils, which entwine about an appropriate surface after the egg is deposited. In the viviparous shark *Squalus* acanthias several eggs pass one after the other through the shell gland, where they are enclosed in one long delicate membranous shell that soon disintegrates. Beyond the shell gland the oviducts terminate in an enlargement, which, in viviparous species, serves as a uterus. An oviducal valve may be found at the junction of oviduct and uterus. Although the two uteri usually open independently into the cloaca, they occasionally unite to form a bicornuate (two-horned) structure. In immature females, the uterus may be separated from the cloaca by a hymen, or membrane. The tract enlarges enormously during the first pregnancy and does not thereafter fully regress to its original size.

The gonoducts of most lower ray-finned fishes resemble those of lungfish, but those of gars and teleosts are exceptional in that the oviducts are usually continuous with the ovarian cavities. A median genital papilla receives the oviducts in teleosts, and the papilla is sometimes elongated to form an ovipositor. European bitterlings deposit their eggs in a mussel by means of the ovipositor. and female pipefish and sea horses deposit them in the brood pouch of a male.

Gonoducts of reptiles, birds, and mammals

With certain modifications, the gonoducts of reptiles and birds are comparable to those of lower vertebrates. Crocodilians, some lizards, and nearly all birds have one gonoduct; the other is not well developed. Even in birds of prey having two functional ovaries, the right oviduct is sometimes undeveloped. The tracts of reptiles generally show less regional differentiation than do those of birds. The oviduct funnel (ostium) in birds forms the chalazae—two coiled, springlike cords extending from the yolk to the ends of the egg. In both reptiles and birds, much of the length of the female tract is oviduct. This region, called the magnum in birds, secretes albumen; lizards and snakes do not form albumen. Behind the albumen-secreting region is a shell gland. In lizards, the gland is midway along the tract. In birds, the shell gland is at the posterior end, has thick muscular walls, and is often inappropriately called a uterus. It is preceded by a narrow region, or isthmus, which secretes the noncalcareous, or soft, membranes of the shell. The shell gland leads to a narrow muscular vagina that empties into the cloaca. The vagina secretes mucus that seals the pores of the shell before the egg is expelled. Special vaginal tubules (spermatheca) store sperm over winter in some snakes and lizards; seminal receptacles have been described in the oviduct funnel in some snakes. In birds, sperm storage glands (sperm nests) often occur in the funnel and at the uterovaginal junction. In lizards and birds, ovulation does not usually occur into a tract already containing an egg. Some lizards shed very few eggs per season; the gecko, for example, sheds only two.

The female reproductive tracts of monotremes, the egg-laying mammals, consist of two oviducts, the lower ends of which are shell glands. These open into a urinogenital sinus, which, in turn, empties into a cloaca. Marsupials have two oviducts, two uteri (duplex uterus), and two vaginas. The upper parts of the vaginas unite to form a median vagina that may or may not be paired internally. Beyond the median vagina, the vaginas are again paired (lateral vaginas) and lead to a urinogenital sinus. The posterior end of the pouchlike median vagina is separated from the forward end of the urinogenital sinus by a partition. When the female is delivering young, the fetuses are usually forced through the partition and into the urinogenital sinus, bypassing the lateral vaginas. The ruptured partition may remain open thereafter, resulting in a pseudovagina. It closes in opossums, and in kangaroos both the median and lateral routes may serve as birth canals. The lateral vaginas in marsupials receive the forked tips of the male penis. Fertilization in all mammals takes place in the oviducts (Fallopian tubes).

In eutherian mammals (*i.e.*, all mammals except monotremes and marsupials), with exceptions noted below, female reproductive tracts beyond the ostia (oviduct funnels) consist of two narrow and somewhat tortuous Fallopian tubes, two large uterine horns (each of which receives a Fallopian tube), a uterine body, and one vagina. Fallopian tubes often have a short dilated ampulla, or saclike swelling, just beyond the ostium. Implantation of the egg occurs only in the uterine horns; the embryos become spaced equidistant from one another in both horns even if only one ovary has ovulated. In some species one horn is rudimentary—the left in the impala (an African antelope)—and the embryos become implanted in the other horn, even though both ovaries ovulate. The body of the uterus in some mammals (*e.g.*, rabbits, elephants, aardvarks; some rodents, bats, insectivores) contains two separate canals (bipartite uterus). In other mammals (ungulates, many cetaceans, most carnivores and bats) the body of the uterus has one chamber into which the two horns empty (bicornuate uterus). There are numerous intermediate conditions between the bipartite and bicornuate condition. Apes, monkeys, and man have no horns, and the Fallopian tubes empty directly into the body of the uterus (simplex uterus). In all mammals, the uterine body tapers to a narrow neck (cervix). The opening (os uteri) into the vagina is guarded by fleshy folds (lips of the cervix). The vagina in eutherian mammals other than rodents and primates terminates in a urinogenital sinus that opens to the exterior by a urinogenital aperture. In some rodents and in higher primates the vagina opens directly to the exterior. In the young of many species a membrane, the hymen, closes the vaginal opening. In guinea pigs the hymen reseals the opening after each reproductive period. Sperm are stored over winter in the uterus of some bats and in vaginal pouches in others.

Accessory glands. Female mammals have fewer accessory sex glands than males, the most prominent being Bartholin's glands and prostates. Bartholin's (bulbovestibular) glands are homologues of the bulbourethral glands of males. One pair usually opens into the urinogenital sinus or, in primates, into a shallow vestibule at the opening of the vagina. Prostates develop as buds from the urethra in many female embryos but often remain partially developed. They become well developed, however, in some insectivores, chiropterans, rodents, and lagomorphs, although their function is obscure. A variety of glands (labial, preputial, urethral) are found in the mucosa, or mucous membrane. Glands in the uterine mucosa provide nourishment for embryos before implantation. Cervical uterine glands secrete mucus that lubricates the vagina, which has no glands (see also REPRODUCTIVE SYSTEM, HUMAN).

ADAPTATIONS FOR INTERNAL FERTILIZATION

Fertilization among vertebrates may be external or internal, but internal fertilization is not always correlated with viviparity or the presence of intromittent (copulatory) organs. The latter, uncommon among fishes, amphibians, and birds, are present in all reptiles (except Sphenodon) and mammals.

Copulatory organs in fishes and amphibians

A considerable number of fishes are viviparous; in them, fertilization is internal, and the males have intromittent organs. The claspers of most male elasmobranchs are usually paired extensions of pelvic fins that are inserted into the female's uterus for transfer of sperm. The clasper, supported by modified fin cartilages, contains a groove along which sperm are conveyed into the uterus and is raised, or erected, by muscles at its base. Gonopodia of male teleosts are fleshy, often elongated modifications of pelvic or anal fins that are directed posteriorly, have a genital pore at the end, and often serve as intromittent organs. In some teleosts, a large penis-like papilla located under the throat is supported by bones. The spermatic duct opens on one side of the papilla. In a few teleosts, hemal spines (ventral projections of vertebrae) form the skeleton of an intromittent organ. Occasionally, the intromittent organ is an asymmetrical tube that matches the asymmetrical genital opening of the female. Still other teleosts have uncomplicated fleshy genital papillae that can be erected. In at least one teleost species, the female has a copulatory organ that she inserts into the genital pore of the male for receiving sperm.

Certain amphibians have internal fertilization but no intromittent organs. The muscular cloaca of the male coecilian, however, can be everted (turned outward) to protrude into that of the female. The male urodele deposits a spermatophore that the female picks up with the lips of her cloaca. Among anurans, Nectophrynoides (a viviparous frog) and Ascaphus (a toad) have internal fertilization, but only Ascaphus has an intromittent organ. It is a permanent tubular extension of the cloaca and resembles a tail. Other anurans have external fertilization and no intromittent organs.

Copulatory organs in reptiles, birds, and mammals

The provision of an eggshell in reptiles requires that fertilization be internal, and all reptiles have intromittent organs except Sphenodon. Reptilian intromittent organs are of two types. Crocodilians and chelonians (turtles) have a penis (phallus), a median thickening in the floor of the cloaca consisting of two parallel cylinders of spongy vascular erectile tissue, the corpora spongiosa (Figure 6). The caudal tip of the penis protrudes into the cloaca as a genital tubercle, or glans penis. The penis is held in the cloacal floor by retractor muscles. When the blood vessels within the spongy bodies are filled with blood, the penis swells, the retractor muscle relaxes, and the genital tubercle protrudes from the vent to serve as an intromittent organ. A longitudinal groove on the surface of the penis directs the flow of sperm. When the spongy bodies are no longer filled with blood, the retractor muscle returns the penis to the cloacal floor. Snakes and lizards have hemipenes, paired elongated outpocketings of the caudal wall of the cloaca that extend under the skin at the base of the tail. Each hemipenis is held in place by a retractor muscle. During copulation the muscle relaxes, the pocket turns inside out and protrudes through the vent in an erect condition. Semen passes along grooves on its surface. Except in pythons, erectile tissue is lacking in hemipenes. Hemipenes protrude independently of each other and are often covered with spines. Very small hemipenes of unknown function are usually present in females.

All birds have internal fertilization, although they are not viviparous; most lack intromittent organs. Male swans, ducks, geese, tinamous, ostriches, and some other ratites (flightless birds), however, have an erectile median penis like that of crocodiles and turtles. Chickens have an organ consisting of a small amount of erectile tissue, but lymph vessels, rather than blood vessels, become engorged. Some birds have a vestigial penis.

All mammals have internal fertilization and an erectile penis. That of monotremes is of the reptilian type, nonprotrusible and in the cloacal floor. In higher mammals the penis has been modified. The groove on the surface of the embryonic penis becomes enclosed in a tube along with the corpus spongiosum and two additional erectile masses, the corpora cavemosa. The proximal ends (crura) of the corpora cavernosa are anchored laterally to the pubic and ischial bones by various muscles and constitute the root of the penis. The crura converge in the midline to enter the body of the penis, which also contains the urethra, surrounded by the corpus spongiosum. The latter begins on the pelvic floor as the bulb of the penis and contains a dilation of the urethra (urethral bulb). The body of the penis extends a variable distance beyond the body of the mammal, in contrast to the short genital tubercle of reptiles. Except in ruminants (*i.e.*, cud-chewing animals, such as cattle and deer), cetaceans, and some rodents, the penis terminates in a glans penis, a swelling of the corpus spongiosum that caps the ends of the corpora cavernosa and contains the urinogenital aperture. The glans is supplied with nerve endings and is partly or wholly sheathed, except during erection, by a circular fold of skin, the prepuce. The inner surface of the prepuce is moistened by preputial glands, and the external surface is sometimes covered with spines or hard scales, as in the cat, guinea pig, and wombat. The glans penis of the male Virginia opossum (Didelphis *virginiana*), the bandicoot, and some other species is bifid (*i.e.*, with two equal tips), correlated with the paired vaginas of females. In boars, the glans penis is corkscrew-shaped, and in goats, rams, and many antelopes a urethral (vermiform) process of much smaller diameter extends three or four centimetres (about an inch to an inch and a half) beyond the glans. In some cattle, a sigmoid, or S-shaped, flexure bends the penis, which otherwise would be too long to fit into the preputial sac. The penis of marsupials is directed backward, and that of cats and rodents is directed backward, except during copulation. In some mammals (*e.g.*, bats) it is pendulous; and in armadillos it may extend one third the length of the body during copulation.

Erection of the mammalian penis is initiated typically by an increase in the volume of blood reaching the cavernous and spongy bodies, engorgement of the vessels, and consequent compression of the veins leaving the organ. When a retractor muscle is present (wolf, fox, dog), it relaxes as erection occurs. The amount of erectile tissue in bovines (cattle) is small, and the penis has much fibroelastic tissue. Erection in such species results primarily from relaxation of the retractor muscle, and vascular engorgement provides only rigidity. Among mechanisms that reverse the erectile state are disgorgement of blood from the cavernous spaces, elasticity of the walls of the spaces, and action of a retractor muscle. A penis bone (baculum, os priapi) is present in various degrees of development in many mammals.

Female mammals have an erectile penile organ known as the clitoris in the floor of the urinogenital sinus or vagina. In the young spider monkey ***Ateles,*** the clitoris is six or seven centimetres long. In a few mammals (some rodents, insectivores, lemurs, and hyenas) the urethral canal becomes enclosed within the clitoris, as in males. In hyenas, the clitoris is large and often mistaken for a penis, and female scrotal pouches, lacking gonads, are present. So much do the male and female external genitalia resemble each other that the ancients regarded the hyena as a hermaphrodite. The clitoris of female mammals often contains cartilage or bone. A specialized clitoris is present in female turtles, crocodiles, alligators, and a few species of birds in which the male has a penis.

The spermatic duct of male mammals between the seminal vesicle and the prostatic urethra has a heavy muscular coat and serves as an ejaculatory duct. In mammals in which the seminal vesicles empty directly into the urethra, the latter is ejaculatory. In birds, the terminal segments of the spermatic ducts are erectile and ejaculatory, and in fish the posterior end of whatever duct transports sperm may be ejaculatory.

ROLE OF GONADS IN HORMONE CYCLES

Neurosecretions formed in the brain in response to environmental stimuli regulate the synthesis and release of hormones known as gonadotropins, which, in turn, stimulate the gonads. Cyclical intervals of illumination (photoperiods) may be the principal environmental factor regulating gonadal activity. Although cyclical temperature changes are experienced by many species, as are fluctuations in food supply, rainfall, and salinity, their precise effects and those of many other stimuli, independently or in combination, have not yet been defined for any species. Photoperiodicity, temperature, and perhaps all other cycles are attributable to the seasons, and to the 24-hour day.

As a result of rhythmic stimulation by gonadotropins secreted by the pituitary gland, the gonads grow, mature, and produce gametes and hormones. Certain of these hormones, known as androgens, are thought to be produced chiefly by interstitial cells and are more abundant in males. Hormones known as estrogens are probably produced chiefly by ovarian follicles and their thecas. Circulating progestins are produced in greatest quantities by corpora lutea. Although the gonadal hormones of different species vary somewhat in structure, their effects are essentially the same. As the quantity of pituitary gonadotropins decreases, the activity of the gonads slows and may temporarily cease.

Effects of gonadal hormones

The effects of gonadal hormones may be summarized as follows:

Gonadal hormones induce growth of and maintain the cyclical function of the reproductive tracts, accessory sex glands, and copulatory or ovipository organs. They thereby provide for the storage, nutrition, and transport of gametes; the secretion of necessary substances onto the surface of gametes; and the ultimate extrusion of sperm, eggs, or the products of conception. In mammals, therefore, they prepare the vagina for copulation and the uterus for implantation of eggs; in addition, gonadal hormones maintain pregnancy until birth or until placental hormones can take over their function. The hormonal basis for the maintenance of viviparity in vertebrates below mammals is almost unknown.

Gonadal hormones participate in the maturation of gametes still in the gonads by augmenting the metabolic effects of other hormones.

Gonadal hormones are essential for the differentiation of manv secondarv sex characters—the physical differences between the sexes—facilitate amplexus (copulatory embrace) and provide for the protection or nutrition of young. Secondary sex characters include scent glands; sexually linked pigmentation of the skin or its appendages; the nature of any vocal apparatus; hardened areas on the appendages that facilitate amplexus; distribution of hair; body size; mammary gland development; and other features.

Gonadal hormones participate in the induction of behaviour necessary for the union of sperm and eggs; this includes migratory phenomena, heat (estrus) in mammals, courtship, territorial defense, mating, and care of eggs or young.

Gonadal hormones participate in a mechanism that affects the pituitary, thereby imposing certain restraints on the secretion of gonadotropins.

The role of light in establishing periodic reproductive activity

The effects of a cyclical environment on gonads is illustrated in mammals that ovulate spontaneously. Ovulation is induced by ovulatory hormones released rhythmically from the pituitary gland. Newborn mice maintained during the first week of life in regular, natural photoperiods will, on reaching maturity, ovulate regularly. Newborn mice kept in continuous light during this interval will not ovulate regularly. The photoperiods in which these animals live as neonates, or newborn, establish an intrinsic brain rhythm that subsequently results in cyclical reproductive activity. If mature female mice that have been ovulating regularly are subjected to continuous light, ovulation ultimately becomes arrhythmical. This suggests that the rhythmical environment is the ultimate regulator of the gonads. Because of the effects of cyclical photoperiods, spontaneous ovulation occurs about the same time of day or night in all members of species intensively studied thus far. Golden hamsters ovulate shortly after midnight; chickens and Japanese rice fish ovulate in the morning. Not all mammals ovulate spontaneously, however. In those that do not (*e.g.*, reflex ovulators), including some cats, rodents, weasels, shrews, rabbits, the act of mating substitutes for the environmental effects on the pituitary gland in releasing ovulatory hormones (see HORMONE).

PROVISIONS FOR THE DEVELOPING EMBRYO

Among the requirements of developing embryos are nutrients, oxygen, a site in which to discharge metabolic wastes, and protection from the environment. These needs exist whether the embryo is developing outside the body of the female parent (oviparity), or within, so that she delivers living young (viviparity). Combinations of yolk, albumen, jellies, and shells contributed by the female parent, as well as membranes constructed from the tissues of the embryo meet the embryo's needs.

Oviparous eggs are usually supplied with enough nutrients to last until the new individual is able to obtain food from the environment. The alternative, postnatal parental feeding, is uncommon. Oviparous animals that develop from yolk-laden eggs are not hatched until they resemble adults. Those that develop from eggs with moderate amounts of yolk hatch sooner, usually into free-living larvae; in this case the larvae transform, or undergo metamorphosis, into adults. The eggs of amphioxus, an oviparous protochordate, contain almost no nutrients; the embryos hatch in an extremely undeveloped but self-sustaining state as few as eight hours after fertilization. The yolk mass is large in some animals and becomes surrounded by a membrane called the yolk sac, the vessels of which convey yolk to the embryo. In some species, yolk also passes from the yolk sac directly into the fetal intestine.

Oviparous fishes and amphibians develop in an aquatic environment, and exchange of oxygen and carbon dioxide and elimination of metabolic wastes occur through the egg membranes. Oviparous reptiles, birds, and monotremes develop on land, and gaseous exchange is accomplished by two membranes (allantois, chorion) applied closely to the shell. The allantois also receives some wastes. Drying out or mechanical injury of embryos of reptiles, birds, and mammals is prevented by still another membrane, the amnion, which is a fluid-filled sac immediately surrounding the embryo.

Types of viviparity

Viviparity has evolved in some members of all vertebrate classes except birds. When eggs heavily laden with yolk and surrounded by a well-formed shell develop within the female, the parent may provide the developing young only with shelter and oxygen (ovoviviparity). At the opposite extreme, if eggs contain only enough nutrients to supply energy for a few cell divisions after fertrlization, the female provides shelter, oxygen, and nourishment, and, in addition, excretes all metabolic wastes produced by the developing organism (euviviparity). Between these extremes are numerous intermediate degrees of dependence on the parent.

Teleosts have evolved many unusual adaptations for viviparity. In some viviparous teleosts the eggs are fertilized in the ovarian follicle, where development occurs. The granulosa cells either form a membrane that secretes nutrients and assists in respiratory and excretory functions or they may be ingested along with follicular fluid, nearby eggs, and other ovarian tissue. A common site for development is the ovarian cavity, which may become distended with as many as nine series of embryos of different ages. Embryos in this location are bathed with nutritive fluids secreted by the epithelium of the cavity. In some species, mortality rates of intraovarian young are high, and surviving individuals ingest those that die. In still other species, extensions of villi in the ovarian lining invade the mouth and opercular (gill) openings of the embryo, filling the opercular chamber, mouth, and pharynx with surfaces that secrete nutrients. The embryos also develop specialized surfaces for nutrition, respiration, and excretion. An enlarged pericardial (heart) sac or an expansion of the'hindgut of the embryo may occur next to the blood-vessel containing (vascular) follicular wall. Vascular extensions may grow out of the anus, urinogenital pore, or gills of the embryo. Other embryonic surfaces —including ventral body wall, fins, and tail—may participate in the support of viviparity. These embryonic surfaces may lie in contact with the follicular or ovarian epithelium, or they may simply be bathed by ovarian fluids. One or more combinations of the maternal and embryonic specializations described above, as well as many others, make viviparity possible among teleost fishes. In a number of teleosts the eggs are incubated, or brooded, in the mouth of the male for periods as long as 80 days. The oral epithelium becomes vascular and highly glandular. In sea horses and pipefish the female deposits her eggs in a ventral brood pouch of the male, and the embryos develop there.

In viviparous elasmobranchs development takes place in the uterus, the lining of which develops parallel ridges or folds covered with villi or papillae (trophonemata) that constitute a simple placenta (site of fetal–maternal contact). In contact with this region is the yolk sac of the embryo, which serves as a respiratory and nutritive membrane. Trophonemata secrete uterine fluids that supplement the yolk as a source of energy. In one shark (Pteroplatea micrura), trophonemata extend into the spiracular chamber (an opening for the passage of respiratory water) of the young and secrete nutrients into the fetal gut. In another (Mustelus antarcticus), the uterine folds form fluid-filled compartments for each embryo. The yolk sac may lie in contact with the uterine lining, or projections of the sac may extend into uterine pits. When the stored yolk is used up before birth, the yolk sac may serve far the absorption of nutrients; *i.e.*, as a placenta. In a few species, immature eggs that enter the oviduct are eaten by the developing young.

Very few amphibians bear living young. In the viviparous frog Nectophrynoides, all development, including larval stages, occurs in the uteri and the young are born fully metamorphosed; *i.e.*, except for size they resemble adults. N. occidentalis, an African species, has a nine-month gestation period. There is almost no yolk in the egg and no placenta, so it is probable that uterine fluids provide nourishment and oxygen. In N. vivipara there are as many as 100 larvae in the uteri, each with long vascular tails that may function as respiratory membranes. *Gastrotheca* marsupiata is an ovoviviparous anuran with a gestation period of three to four months. In certain viviparous salamanders the extent of the nutritional dependence on the mother varies. After depleting their own yolk supply, the larvae of some forms eat other embryos and blood that escapes from the uterine lining. Conventional viviparity is rare among amphibians; however, they have evolved unusual alternatives. In some anurans the young develop in such places as around the legs of the male (Alytes), or in pouches in the skin of the back (some females of the genera Nototrema, Protopipa, and *Pipa*). In *Pipa*, vascular partitions in the skin pouch separate developing young, and the larvae have vascular tails that absorb substances. In Nototrema larval gills have vascular extensions with a similar function. The male Chilean toad (Rhinoderma darwinii) carries developing eggs in the vocal sac until the young frogs emerge.

Some snakes and lizards and all mammals except monotremes exhibit viviparity to some degree. The same extra-embryonic membranes found in oviparous reptiles and mammals (yolk sac, chorioallantoic membrane, amnion) function in viviparous ones. Here, the extra-embryonic membranes lie against the uterine lining instead of against an egg shell. At special sites of fetal–maternal contact (placentas), viviparous young receive oxygen and give up carbon dioxide; metabolic wastes are transferred to maternal fluids and tissues; and, in euviviparous species, the young receive all their nutrients. Yolk-sac placentas are common in marsupials with short gestation periods (opossum, kangaroo) and in lizards. Chorioallantoic placentas (*i.e.*, a large chorion fused with a large allantois) occur in certain lizards, in marsupials with long gestation periods, and in mammals above marsupials. The yolk-sac placenta does not invade maternal tissues, but intimate interlocking folds may occur between the two. The chorioallantoic membranes of reptiles and mammals exhibit many degrees of intimacy with maternal tissues, from simple contact to a deeply rooted condition (deciduate placentas). Chorioallantoic or chorionic placentas represent specializations in a chorionic sac surrounding the embryo. The entire surface of the sac may serve as a placenta (diffuse placenta, as in pigs); numerous separate patches of placental thickenings may develop (cotyledonary placenta, as in sheep); a thickened placental band may develop at the equator of the chorionic sac (zonary placenta, as in cats); or there may he a single oval patch of placental tissue (discoidal placenta, as in higher primates).

BIBLIOGRAPHY. E.J.W. BARRINGTON, The Biology of *Hemi*chordata and Protochordata (1965), contains much information on reproduction and life histories that can be read profitably by specialist and nonspecialists alike; V.N. BEKLEMISHEV, Principles of Comparative Anatomy of Invertebrates, 2 vol. (1969; orig. pub. in Russian, 1964), a concise account of basic structural patterns in major groups; R.P. DALES, Annelids (1963), a college-level account; K.G. DAVEY, Reproduction in the Insects (1965), a somewhat advanced but excellent review; R.W. HEGNER and J.G. ENGEMANN, Invertebrate Zoology (1968), a college-level text covering major groups but in somewhat less detail than the work by Meglitsch cited below; L.H. HYMAN, The Invertebrates, 6 vol. (1940–67), the most authoritative and detailed work in existence on Protozoa through Mollusca, with extensive bibliographies; P.A. MEGLITSCH, Invertebrate Zoology (1967), a college-level text covering all major groups, highly readable and well illustrated; J.E. MORTON, *Molluscs,* 4th rev. ed. (1967), a readable and interesting account for the nonspecialist; W.D. RUSSELL-HUNTER, A Biology of Lower Invertebrates (1968), an introductory work, not extensive but very readable; W.L. SCHMITT, Crustaceans (1965), an elementary account; J.H. WILMOTH, Biology of Invertebrata (1967), readable, introductory accounts of reproduction; s . ~ASDELL, *Pat-*

terns of Mammalian Reproduction, 2nd ed. (1964), species-by-species data in capsule form on selected reproductive processes of the world's mammals; J.F. DANIEL, *The Elasmobranch Fishes,* 3rd rev. ed. (1934), a definitive work on the anatomy of representative cartilaginous fishes, including reproductive systems; P. ECKSTEIN and S. ZUCKERMAN, "Morphology of the Reproduction Tract," in F.H.A. MARSHALL, *Physiology of Reproduction,* 3rd ed. by A.S. PARKES, vol. 1 (1959); E.S. GOODRICH, *Studies on the Structure and Development of Vertebrates,* 2 vol. (1930, reprinted 1958), out of date with respect to development and certain theoretical discussions but an excellent reference for morphological details of vertebrates, with extensive bibliography; A.E. HARROP, *Reprodztction in the Dog* (1960), a thorough, readable text including basic reproductive anatomy and physiology and veterinary information; A.D. JOHNSON, W.R. GOMES, and N.L. VANDEMARK (eds.), *The Testis,* vol. 1, *Development, Anatomy, and Physiology* (1970), one of the few complete works on the subject, with extensive bibliography; S. SISSON, *Anatotny of the Domestic Animal,* 4th ed. rev. by J.D. GROSSMAN (1953), a basic anatomy for veterinary medical students, but quite readable by the nonspecialist; A. VAN TIENHOVEN, *Reproductive Physiology of Vertebrates* (1968), primarily for the reproductive physiologist, but contains valuable anatomical data relating to all vertebrate classes; W.C. YOUNG (ed.), *Sex and Internal Secretions,* 2 vol. (1961), primarily for endocrinologists, but with a wealth of data on reproductive activities and structures of vertebrates, with extensive bibliographies; S. ZUCKERMAN (ed.), *The Ovary* (1962), a detailed account of the development, structure, and function of vertebrate ovaries, commencing with protochordates; *The Cambridge Natural History,* 10 vol. (1895–1936), phylogenetically arranged detailed discussions of anatomy and natural history of invertebrates and vertebrates, with a section on reproductive organs included for most major taxonomic groups; PETER GRAY (ed.), *The Encyclopedia of the Biological Sciences* (1961), includes very brief but valuable information on reproduction in most animal groups; T.J. PARKER and W.A. HASWELL, *A Text-Book of Zoology,* 7th ed., 2 vol. (1962–63), a thorough college-level study of the basic morphology of the major invertebrate and vertebrate phyla, extensively illustrated.

(G.C.K.)

Reproductive Systems, Plant

In plants, as in animals, the end result of reproduction is the continuation of a given species, and the ability to reproduce is, therefore, rather conservative, or given to only moderate change, during evolution. Changes have occurred, however, and the pattern is demonstrable through a survey of plant groups. Reproduction is basically either asexual or sexual. Asexual reproduction in plants involves a variety of widely disparate methods for creating new plants identical in every respect to the parent plant. Sexual reproduction, on the other hand, depends on a complex series of basic cellular events, involving chromosomes and their genes, that take place within an elaborate sexual apparatus evolved precisely for the creation of new plants in some respects different from the two parents that played a role in their production. (See REPRODUCTION for an account of the common details of asexual and sexual reproduction and the evolutionary significance of the two methods.)

In order to describe the modification of reproductive systems, plant groups must be identified. One convenient classification of organisms sets plants apart from lower forms such as bacteria, algae, fungi, and protozoans. Under such an arrangement, the plants, as separated, comprise two great divisions (or phyla)—the Bryophyta (mosses and liverworts) and the Tracheophyta (vascular plants). The vascular plants include four subdivisions: the three entirely seedless groups are the Psilopsida, Lycopsida, and Sphenopsida; the fourth group, the Pteropsida, consists of the ferns (seedless) and the seed plants (gymnosperms and angiosperms).

A comparative treatment of the two patterns of reproductive systems will introduce the terms required for an understanding of the survey of those systems as they appear in selected plant groups.

GENERAL FEATURES

Asexual systems. Asexual reproduction involves no union of cells or nuclei of cells and, therefore, no mingling of genetic traits, since the nucleus contains the genetic material (chromosomes) of the cell. Only those systems of asexual reproduction that are not really modifications of sexual reproduction are considered below. They fall into two basic types: systems that utilize almost any fragment or part of a plant body; and systems that depend upon specialized structures that have evolved as reproductive agents.

Two basic asexual processes

Reproduction by fragments. In many plant groups, fragmentation of the plant body, followed by regeneration and development of the fragments into whole new organisms, serves as a reproductive system. Fragments of the plant bodies of liverworts and mosses regenerate to form new plants. In nature and in laboratory and greenhouse culture, liverworts fragment as a result of growth; the growing fragments separate by decay at the region of attachment to the parent. During prolonged drought, the mature portions of liverworts often die, but their tips resume growth and produce a series of new plants from the original parent plant.

In mosses, small fragments of the stems and leaves (even single cells of the latter) can, with sufficient moisture and under proper conditions, regenerate and ultimately develop into new plants.

It is common horticultural practice to propagate desirable varieties of garden plants by means of plant fragments, or cuttings. These may be severed leaves or portions of roots or stems, which are stimulated to develop roots and produce leafy shoots. Naturally fallen branches of willows *(Salix)* and poplars *(Populus)* root under suitable conditions in nature and eventually develop into trees. Other horticultural practices that exemplify asexual reproduction include budding (the removal of buds of one plant and their implantation on another) and grafting (the implantation of small branches of one individual on another).

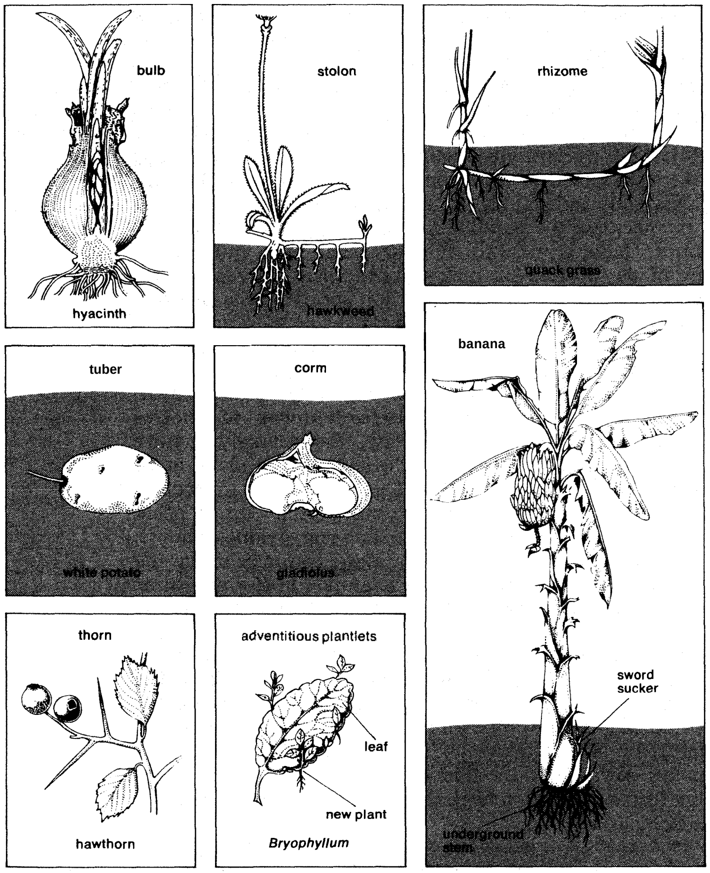
Reproduction by special asexual structures. Throughout the plant kingdom, specially differentiated or modified cells, groups of cells, or organs have, during the course of evolution, come to function as organs of asexual reproduction. These structures are asexual in that the individual reproductive agent develops into a new individual without the union of sex cells (gametes). A number of examples of special asexual agents of reproduction from several plant groups are cited in this section.

Spores and vegetative organs

Airborne spores characterize most nonflowering land plants, such as mosses, liverworts, and ferns. Although the spores arise as products of meiosis, a cellular event in which the number of chromosomes in the nucleus is halved, such spores are asexual in the sense that they may grow directly into new individuals, without prior sexual union.

Among liverworts, mosses, lycopods, ferns, and seed plants, few to many celled, specially organized buds, or gemmae, also serve as agents of asexual reproduction.

The vegetative, or somatic, organs of plants may, in their entirety, be modified to serve as organs of reproduction. In this category belong such flowering-plant structures as stolons, rhizomes, tubers, corms, and bulbs, as well as the tubers of liverworts, ferns, and horsetails, the dormant buds of certain moss stages, and the leaves of many succulents (Figure 1). Stolons are elongated runners, or horizontal stems, such as those of the strawberry, which root and form new plantlets when they make proper contact with a moist soil surface. Rhizomes, as seen in iris, are fleshy, elongated, horizontal stems that grow within or upon the soil. The branching of rhizomes results in multiplication of the plant. The enlarged, fleshy tips of subterranean rhizomes or stolons are known as tubers, examples of which are potatoes. Tubers are fleshy storage stems, the buds ("eyes") of which, under proper conditions, can develop into new individuals. Erect, vertical, fleshy, subterranean stems, which are known as corms, are exemplified by crocuses and gladiolas. These organs tide the plants over periods of dormancy and may develop secondary cormlets, which give rise to new plantlets. Unlike the corm, only a small portion of the bulb, as in lilies and the onion, represents stem tissue. The latter is surrounded by the fleshy, food-storage bases of earlier



Fiaure 1: Structures servina asexual reproduction.

From (rhizome, tuber, corm, thorn, adventitious plantlets, banana) *Biological* Science: An *Inquiry* into Life 2nd ed. (1968) Harcourt Brace Jovanovich, Inc., New York, by permission of ihe Biological Sciences Curriculum Study; (bulb) Gilbert M. Smith, et *al.*, A Textbook of General Botany (© 1953), The Macmillan Company. (stolon) *Botany*, 3rd ed. by Carl L. Wilson and Walter E. Loams, Copyright 1952. © 1957, 1962 by Holt. Rinehart and Winston. Inc., reproduced by permission of Holt, Rinehait and Winston, Inc

formed leaves. After a period of dormancy, bulbs develop into new individuals. Large bulbs produce secondary bulbs through development of buds, resulting in an increase in number of individuals.

Sexual systems. In most plant groups both sexual and asexual methods of reproduction occur. Some species, however, seem secondarily to have lost the capacity for sexual reproduction. Such cases are described below (see Variations in reproductive cycles).

The cellular basis. Sexual reproduction at the cellular level generally involves the following phenomena: the union of sex cells and their nuclei, with concomitant association of their chromosomes, which contain the genes, and the nuclear division called meiosis. The sex cells are called gametes, and the product of their union is a zygote. All gametes are normally haploid (having a single set of chromosomes) and all zygotes, diploid (having a double set of chromosomes, one set from each parent). Gametes may be motile, by means of whiplike hairs (flagella) or of flowing cytoplasm (amoeboid motion). In their union, gametes may be morphologically indistinguishable (*i.e.*, isogamous) or they may be distinguishable only on the criterion of size (*i.e.*, heterogamous). The larger gamete, or egg, is nonmotile; the smaller gamete, or sperm, is motile. The last type of gametic difference, egg and sperm, is often designated as oogamy. In oogamous reproduction, the union of sperm and egg is called fertilization. Isogamy, heterogamy, and oogamy are often considered to represent an increasingly specialized evolutionary series (Figure *2).*

In the plants included in this article—bryophytes (mosses and liverworts) and tracheophytes (vascular plants)—sexual reproduction is of the oogamous type, or a modification thereof, in which the sex cells, or gametes, are of two types, a larger nonmotile egg and a smaller motile sperm. These gametes are often produced in special containers called gametangia, which are multicellular. In cases in which special gametangia are lacking, every cell produces a gamete. In oogamy, the male gametangia are called antheridia and the female oogonia or archegonia. A female gametangium with a sterile cellular jacket is called an archegonium although, like an oogonium, it produces eggs. In most of the plants dealt with in this article the eggs are produced in archegonia and the sperms in antheridia with surface layers of sterile cells.

The plant basis. Individual plants may be either bisexual (hermaphroditic), with male and female gametes produced by the same organism, or unisexual, producing either male or female gametes but not both. A bisexual individual, however, may not be capable of fertilizing its own eggs. In certain ferns, for example, male gametes of one individual are not compatible with the female gametes of the same individual, so that cross-fertilization (with another individual) is obligatory. This, of course, is similar in adaptive significance to cross-pollination (which leads to cross-fertilization) among seed plants.

The "alternation of generations"

Among the liverworts, mosses, and vascular plants, the life cycle involves two different phases, often called generations, although only one plant generation is, in fact, involved in one complete cycle. This type of life cycle is often said to illustrate the "alternation of generations" in which a haploid individual (*i.e.*, with one set of chromosomes), or tissue, called a gametophyte, at maturity produces gametes that unite in pairs to form diploid (*i.e.*, containing two sets of chromosomes) zygotes. The latter develop directly into individuals, or tissues, called sporophytes, in which the nuclei of certain fertile cells, called spore mother cells, or sporocytes, give rise to haploid spores (sometimes called meiospores). These spores are lightweight and are borne by air currents; they germinate to form the haploid, sexual, gamete-producing phase, usually designated the gametophyte.

There are several variations in the above described life cycle. The haploid gametophyte and sporophyte may be free-living, independent plants (*e.g.*, certain algae and yeasts), in which case the life cycle is said to be diplobiontic; or the sporophyte may be physically attached to the gametophyte, as it is in liverworts and mosses. By contrast, the gametophytic phases develop as parasites on the sporophytes of the seed plants, as in certain algae. In further variation, the alternating phases may be similar morphologically except for the type of reproductive cells, gametes or spores they produce (isomorphic life cycle); or they may be strikingly dissimilar, as in some algae, mosses, ferns, and seed plants (heteromorphic life cycle). Only heteromorphic life cycles occur in liverworts, mosses, vascular plants, and certain fungi.

The differences between the gametophyte and sporophyte are often great, especially those of the diplobiontic types, so that the alternates seem to be two different, unrelated individuals rather than different manifestations of the same organism.

From P.B. Weisz, The Science *of Biology*, 3rd ed. (1967), used with permission of McGraw-Hill Book Company

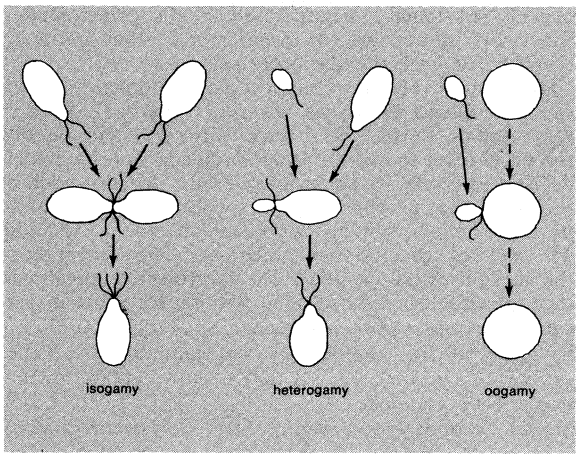


Figure 2: Patterns of fertilization based on gamete types. In isogamy and heterogamy all gametes are motile; in oogamy oniy the smaller is motile, and moves to the larger.

BRYOPHYTE REPRODUCTIVE SYSTEMS

Liverworts and hornworts. The plant bodies of liverworts and hornworts represent the gametophytic (sexual) phase of the life cycle, which is dominant in these plants. In the liverworts, the sporophyte is borne upon or within the gametophyte but is transitory. Liverwort and hornwort plants, depending on the species, may be bisexual or unisexual, and the sex organs may be distributed on the surface (Riccia, Ricciocarpus, *Sphaerocarpos, PelIia*) or localized in groups and borne on special branches (antheriodiophores and archegoniophores) as in *Marchantia*; the sperms are biflagellate (Figure 3).

From (top) Botany 3rd ed by Carl L Wilson and Walter E Loomis, Copyright 1952 © 1957 1962 by Holt Rinehart and Winston. Inc , reproduced by permission of Hoit, Rinehart and Winston Inc , (bottom) T E Weir *et al* , Botany (1967) John Wiley & Sons Inc

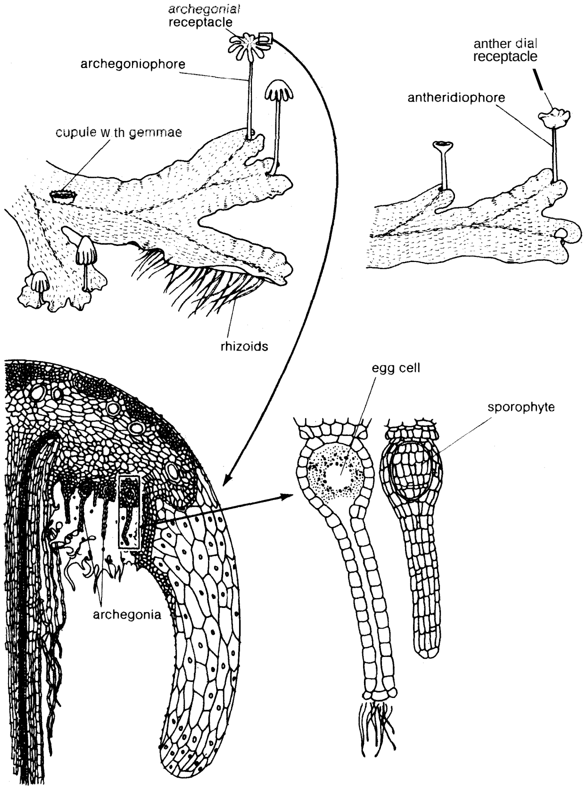


Figure 3: Archegonia and antheridia in the liverwort Marchantia.

Release of the mature sperm and the process of fertilization require moisture in the form of heavy dew or raindrops. In all but a few genera (Riccia, Ricciocarpus), the developing sporophytes are actively photosynthetic—*i.e.,* capable of utilizing light energy to form organic substances. They are, however, dependent on gametophytic tissues for water (and the inorganic salts dissolved in it) and probably derive and utilize in their nutrition some organic substances manufactured by the gametophytes. Liverwort spores are meiospores; *i.e.,* they arise by meiosis from cells called sporocytes.

The sporophytes may consist almost completely of fertile (sporogenous) tissues (Riccia, Oxymitra), or they may contain sterile cells (nurse cells or elaters) among the developing spores. In Marchantia and Porella, a sterile foot and seta, or stalk, are present; the foot anchors the spore-bearing capsule (sporangium) to the gametophyte and also probably serves an absorptive function. The seta connects the foot and capsule. The elongation of the seta raises the capsule from its protective envelopes, thus, placing it in a favourable position for spore dispersal. The capsules of liverworts may shed their spores only by decay of the capsule wall and gametophytic tissues (Riccia, Oxymitra), or they may open irregularly or into two or four segments.

Spore germination in some species may occur immediately after deposition if the spores are in a favourable environment; or, as in other species, the spores may require a period of dormancy before germination.

Mosses. In mosses, as in liverworts and hornworts, the leafy shoots belong to the gametophytic phase and produce sex organs when they mature (Figure 4). The leafy shoots (often called gametophores, because they bear the sex organs) arise from a preliminary phase called the protonema, the direct product of spore germination. Filamentous, straplike, or membranous, it grows along the soil surface. A protonema of a moss may proliferate, apparently indefinitely, under favourable conditions and thus increase the population of leafy shoots that arise as buds. Under adverse conditions, certain buds and branches of the protonema may thicken their walls and thus serve to tide the species over an unfavourable growing period.

The gametophores of mosses

The antheridia and archegonia may be borne at the tips (apices) of the main shoots or on special, lateral branchlets. Both bisexual and unisexual leafy shoots occur, depending on the species. In a number of mosses (*Mnium, Polytrichum,* Funaria), the sexually mature shoots become recognizable through the production of special, prominent leaves that form an apical cup around the sex organs. If brightly coloured, the cup is often flowerlike. In species with bisexual leafy gametophores, the archegonia and antheridia may be present on the same apex (as can be seen, for example, in *Bryum*) or at the apices of separate branches as is evidenced in the moss *Funaria.*

The archegonia and antheridia of mosses are large enough in many species to be just barely visible to the unaided eye. The jacket cells of the antheridia are often coloured bright orange or rust; their sperm are biflagellate. As in liverworts and hornworts, rains and even heavy dews evoke the liberation of sperm and the opening of the mature archegonia so that fertilization may be accomplished.

The moss sporophyte, which is attached to the gametophyte, photosynthesizes during much of its development and is more or less self-supporting. It is, to a certain degree, dependent upon the gametophyte for nutrients

From *Biological* Science: An inquiry into Life, 2nd ed. (1968); Harcourt Brace Jovanovich. Inc., New York: by permission of the *Biological* Sciences Curriculum Study

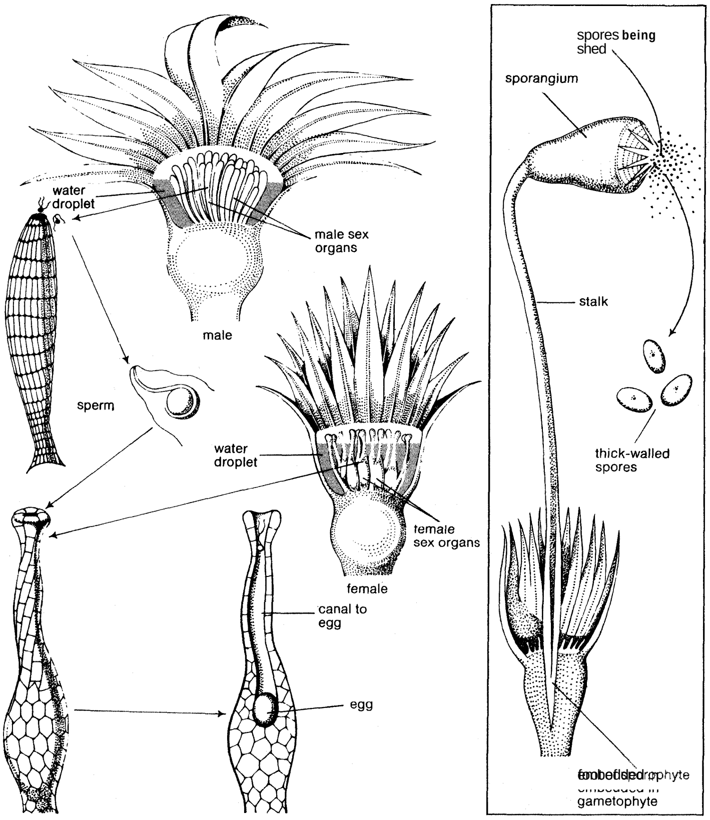


Figure 4: The two phases of a moss plant.
(Left) Sectioned tips of male and female moss gametophyte plants. (Right) Mature sporophyte releasing spores that will develop into garnetophyte.

such as water and mineral salts and, in some cases, even for elaborated foods.

After elongation of the moss sporophyte has ceased, the distal portion (farthest away) enlarges to form the capsule (sporangium), or spore-bearing region. The spores (meiospores), which arise by meiosis, are shed from the capsules gradually through a variety of mechanisms. After the operculum (cover) of the capsule has been shed, its mouth is usually partially closed by the peristome (teeth) and sometimes by associated structures. These teeth absorb moisture, and their resultant swelling and contraction open spaces through which the spores are shed.

TRACHEOPHYTE REPRODUCTIVE SYSTEMS

Dominance of the gametophyte

Spore plants. In liverworts, hornworts, and mosses, the dominant phase in the life cycle is the sexual gametophyte. In the tracheophytes (vascular cryptogams and seed plants), on the other hand, the sporophyte is the dominant phase in the life cycle. The gametophytes of the vascular cryptogams mature after the spores that initiated them have been shed from the parent plant, so that the gametophytes are free-living. In the seed plants the garnetophytes mature as parasites on the sporophytes.

Psilopsids. The trilobed sporangia of the whisk fern *Psilotum*—not a true fern but a psilopsid—are borne terminally on short lateral branches. During development, some of the potentially spore-bearing tissue is used as nutrient by the sporocytes as they complete the meiotic divisions that result in colourless kidney-shaped spores. The latter, which are shed as the sporangia open along three lines, germinate and slowly develop into cylindrical, sparingly branched gametophytes, about 0.5 to two millimetres (0.02 to 0.08 inch) in diameter and several millimetres long. They presumably derive nourishment from decaying matter and occur in humus-rich soil, in rock fissures, or among roots on the trunks of tree ferns. The cells of the gametophyte contain fungal structures (hyphae) that probably are involved in some type of nutritional relation with the gametophyte.

The gametophytes are bisexual and the sperms multiflagellate. The embryonic sporophyte is not easily distinguished from the gametophyte that bears it. At first anchored to the gametophyte by an absorptive foot, the sporophyte ultimately becomes separated from both the foot and the gametophyte.

Lycopsids. In the genus Lycopodium, the sporangia are closely associated with the leaves. In some species (L. *lucidulum*), the sporangium-bearing leaves (sporophylls) occur in zones among the vegetative portions of the stems. In most, however, the sporophylls occur in specialized compressed stems, called cones, or strobili (Figure 5). Each sporophyll is associated with one yellow to orange, kidney-shaped sporangium.

From Biolooical Science: An Inquiry into Life, 2nd ed. (1968): Harcourt Brace Jovanovich, Inc., New York; by permission of the Biological Sciences Curriculum Study

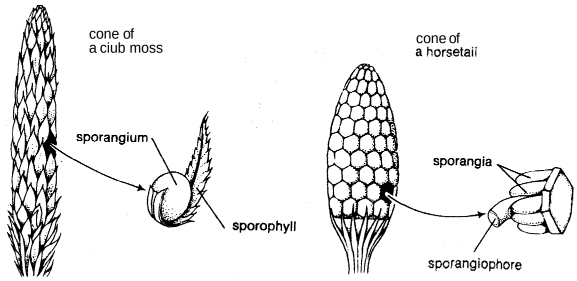


Figure 5: Sporangial protection of a club moss and a horsetail.

In several species the spores develop rapidly on the soil surface into ovoid-cylindrical gametophytes about two to three millimetres (0.08 to 0.12 inch) long, with green lobes and colourless bases; they usually contain a fungus. In other species, development of a colourless gametophyte is slow, so that at maturation, which may require up to eight years, the fleshy gametophyte will have become buried in successive layers of humus. These subterranean gametophytes, which contain fungi, are long-lived and are larger (up to two centimetres) than the surface types.

The gametophytes of Lycopodium are bisexual, although the antheridia and archegonia may develop into separate groups. The sperms are biflagellate and apparently more than one egg of the same gametophyte may be fertilized.

The zygote divides at a right angle to the long axis of the archegonium. The inner cell gives rise to the embryo, which thus is oriented as if it will develop within the gametophyte; it turns 180" during later development, however, and the axis grows vertically outward from the gametophyte.

The heterosporous condition

In contrast to *Lycopodium*, all *Selaginella* sporophytes have sporophylls localized in strobili, and all species of *Selaginella* are heterosporous: that is, they produce spores of two sizes, the larger designated as megaspores and the smaller as microspores. The megaspores develop into female gametophytes and the microspores into male gametophytes. Accordingly, strobili bear megasporophylls that contain megasporangia, which will produce megaspores, and microsporophylls that contain microsporangia, which will yield microspores. Although the evolutionary origin of two kinds of spores (dimorphism) is unknown, the development of megaspores in living plants suggests that differences in nutrition in the two kinds of sporangia are significant. In a microsporangium, most of the microsporocytes undergo meiosis, forming four spores each; by contrast, all but one or, occasionally, several of the sporocytes in the megasporangium do not complete development. As a result, only four megaspores usually mature in such a sporangium, enlarging as they become gorged with the nutrients made available by disintegration of the other cells. The megaspores, accordingly, are much larger than microspores, although both contain stored food. Both types of spores are thick walled, and both have prominent three-part (triradiate) ridges.

Unlike the homosporous spores of most liverworts, hornworts, mosses, ferns, and Lycopodium, the spores of *Selaginella* begin to develop into gametophytes before they have been shed from their sporangia and attain maturity on a suitable, moist substrate.

The microscopic male gametophyte is composed essentially of a single antheridium, which produces biflagellate sperm. The female gametophyte, which protrudes after the megaspore wall cracks open in the region of the triradiate ridge, consists of vegetative cells, has several archegonia at maturity, and usually has three groups of rhizoids. Both male and female gametophytes lack the chlorophyll (green pigment) necessary for photosynthesis; they utilize nutrients stored in the spores.

After fertilization, one zygote of each female gametophyte develops into an embryonic sporophyte. There is considerable variation in details of development among the species of *Selaginella.* In some, the spores may develop mature gametophytes before they are shed from their sporangia, and fertilization may occur, so that female gametophytes with embryos may be found in the strobili (compressed stems, or cones). The megaspores of *Selaginella,* containing female gametophytes with still-attached juvenile sporophytes, have the superficial appearance of germinating seeds, from which, however, they differ in many significant respects.

Zsoetes, like *Selaginella,* is monoecious and heterosporous. Most of the leaves are fertile; some bear one large megasporangium each, and others support a single microsporangium on the inner surface of a spoonlike leaf base. The microsporangia can produce enormous numbers of microspores—as many as 1,000,000—and the megasporangia give rise to 50 to 300 megaspores. The spores are liberated as the older sporophylls decay. Unlike those of *Selaginella,* the spores of Zsoetes do not germinate until they have been shed from their sporangia. The unisexual garnetophytes are much like those of *Selaginella,* but the sperm are multifiagellate. The embryonic sporophyte is nourished by food stored in the megaspore and transported through a massive foot.

Sphenopsids. The perennial sporophytes of horsetails (Equisetum species) produce strobili once during every growing season. They ma, be borne at the tips of green shoots (E. *hyemale*, E. kansanum); at the tips of nongreen shoots that become green after the spores have been shed (E. *fluviatile*, E. sylvaticum); or on special nongreen branches that wither and die after the spores have been shed (E. arvense, E. talmateia). The appendages of the strobilus are often called sporangiophores and have been considered to be both stem branches and of leafy origin; in the latter case, they are called sporophylls (Figure 5). Each sporangiophore bears a number of fingerlike sporangia, which produce large numbers of thin-walled, green spores. The outermost wall layer of the spore breaks down into four appendages, which, by their sensitivity to moisture, coil and uncoil, thereby disseminating the spores.

The spores of *Equisetum* germinate rapidly and grow into green, pincushion-like gametophytes anchored to the surface by rhizoids. Apparently, two types of gametophytes are produced from the homosporous spores; some mature slowly, are smaller than others, and always produce antheridia, never archegonia. Others are larger and hermaphroditic, producing archegonia at first and, later, antheridia. The ratios of male to hermaphroditic gametophytes vary among species but are relatively uniform within a species. The ratios are altered by changes in environmental conditions; for example, at certain temperatures (*e.g.*, 32° C, or 90° F) only male gametophytes develop from the spores of five species; whereas at 15" C (59° F) approximately 50 percent are male and 50 percent hermaphroditic gametophytes.

Environmental influences on sex ratios

Self-fertilization of hermaphroditic gametophytes can occur, and several sporophytes may be produced on one gametophyte. The embryo consists of an absorptive foot, a primary root or radicle, and a shoot with whorled appendages.

Ferns. As they mature, many fern sporophytes begin to produce spores in clusters of sporangia on the undersurfaces of their vegetative "leaves." Others produce their sporangia on highly modified leaves or portions thereof.

The site of origin of the sporangia is the receptacle; the latter, with its groups of sporangia, is called a sorus. In many ferns each sorus is covered with a special outgrowth, the indusium; in others, the sporangia are covered during development by the margin of the leaf. In a few ferns (*e.g.*, Polypodium), the sori remain uncovered.

In primitive ferns, such as Ophioglossum and *Botrychium*, the spores are borne upon a specialized axis, the fertile spike. The sporangia of such primitive ferns are massive, with several layeis of cellular walls, and produce an Indefinite but large number of spores. In most other ferns, the sporangia are smaller, long stalked, with single-layered walls and a definite number of spores. The spores of the latter are shed explosively by breakage and shrinking as the sporangia open and then slam shut.

Homospory and heterospory in ferns

Most ferns produce one kind of spore (homospory), but a few genera of aquatic and amphibious ferns (Marsilea, Salvinia, and Azolla) produce two kinds (heterospory), small microspores and much larger megaspores. In either case, after being shed from the parent sporophyte, the spores that have suitable environmental conditions germinate and develop into the gametophytic phase. The ribbonlike, filamentous or heart-shaped gametophytes of most ferns contain chlorophyll, are anchored to some surface—moist soil, moist rocks, or tree bark—by unicellular root-like rhizoids, and rarely exceed one-half inch (13 millimetres) in diameter. In a few ferns (*Ophioglossum*, Botrychium, and certain species of Schizaea), the gametophytes are subterranean, lack chlorophyll, are cylindrical or tuberous, and contain the filamentous structures (hyphae) of an associated fungus.

Fern gametophytes, often called prothalli (singular, prothallus) are one cell layer thick except in the centre. Most fern prothalli are bisexual—*i.e.*, have both male (antheridia) and female (archegonia) sex organs, which develop usually on the undersurface of the prothallus.

Although the eggs of several archegonia may be fertilized, only one zygote usually develops into a juvenile sporophyte. The latter consists of an absorbing foot; a primary root, or radicle, which promptly penetrates the surface; a prominent first leaf; and a rudimentary, slow-growing stem. As the juvenile sporophyte becomes established, the parental gametophyte dies. The series of leaves formed from the stem of the juvenile sporophyte gradually attain the form and vein pattern that characterize the mature sporophyte.

In most ferns, the antheridia appear before the archegonia and continue to develop as the latter mature; furthermore, the archegonial necks curve toward the mature antheridia so that fertilization can readily occur. Both gametes may be derived from one individual, or from different individuals. In the bracken fern (*Pteridium* aquilinum), although the gametophytes are bisexual, self-incornpatability factors reduce self-fertilization. In Onoclea *sensiblis*, the gametophytes are unisexual in early development, thus favouring cross-fertilization; but, later, the gametophytes become bisexual so that, if cross-fertilization fails, the species can still be maintained.

Seed plants. In the two great groups of seed plants, gymnosperms and angiosperms, the sporophyte is the dominant phase in the life cycle, as it is also in the vascular cryptogams; the gametophytes are microscopic parasites on the sporophytes (Figure *6)*.

Reprtnted from The Spectrum of *Life* by Harold A. Moore and John R. Carlock; Copyright © 1970. Harper & Row, Publtshers. Inc.

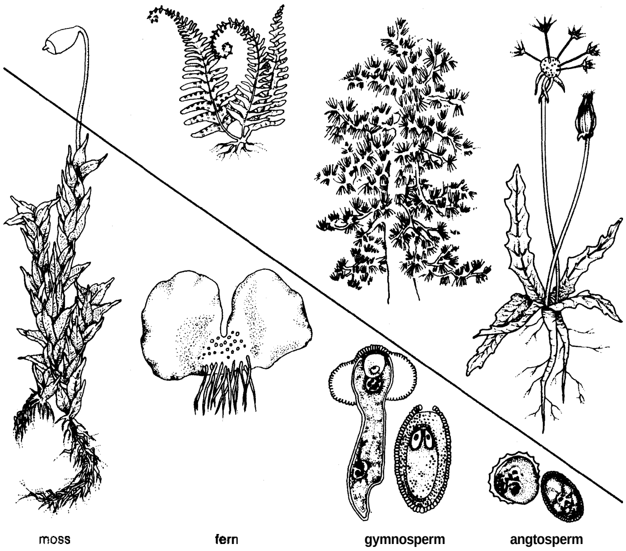


Figure 6: Change in relative size of the gametophyte (below the line) and sporophyte (above the line) in the course of plant evolution. In the moss, the sporophyte is entirely dependent upon the gametophyte.

In the gymnosperms, the seeds occur individually, exposed at the ends of stalks, sometimes in whorls on an axis, or on the scales of a cone, or megastrobilus. In angiosperms, or flowering plants, by contrast, the seeds are enclosed during development in a structure variously termed a pistil or carpel, which is sometimes considered to represent an enfolded megasporophyll.

Events in the reproductive process

A number of parts of the reproductive process are common to both angiosperms and gymnosperms: (1) they produce seeds at maturity; (2) the megasporangium, unlike that of heterosporous seedless plants, is covered by one or two cellular layers called integuments and is termed an ovule; (3) there is a minute passageway, or micropyle, through the integuments; (4) the ovule matures as a seed; (5) only one megasporocyte is present and undergoes meiosis in the megasporangium to produce four megaspores, only one of which usually is functional; *(6)* the megaspore is never discharged from its megasporangium and ovule; (7) one female gametophyte is produced within each megasporangium and ovule; (8) the microspores begin their development into male gametophytes while still enclosed in the microsporangia; (9) as they mature, the male gametophytes, which are contained within the microspore wall and are termed pollen grains, develop a tube that conveys sperm to the egg cell;

(10) union of sperm and egg and development of an embryonic sporophyte from the zygote occur within the female gametophyte (sometimes called the "embryo sac"), which is covered by the remains of the megasporangium and by integuments; (11) as the embryo develops, the ovule matures as a seed.

In contrast to this impressive list of similarities are important differences, which, in addition to seed position, serve to distinguish angiosperms from gymnosperms. The reproductive cycle in most angiosperms is completed quicker than that in gymnosperms, and the gametophytes are smaller and simpler and, unlike those of most gymnosperms, lack archegonia. The pollen in angiosperms is transferred to the surface of the megasporophyll, whereas in gymnosperms it is brought to the micropyle of the ovule itself. Two sperm are involved in the sexual union in angiosperms: one unites with the egg to form a zygote; the other unites with two nuclei of the female gametophyte to form the primary endosperm nucleus. The latter divides to form a postfertilization storage tissue, which serves as a food source for the embryo; the embryo of gymnosperms is nourished by the somatic (nonreproductive) tissues of the female gametophyte. The angiosperm ovule increases to mature seed size after fertilization, whereas in gymnosperms, this enlargement occurs prior to fertilization.

The general features of the reproduction of seed plants having now been summarized, certain special aspects of the reproduction in representative seed plants are described below.

Gymnosperms. The cycads are slow-growing dioecious gymnosperms, the microsporangia (potential pollen) and megasporangia (potential ovules) occurring on different individual sporophytes. In all cycads except the genus Cycas, the ovules are borne on megasporophylls in megastrobili; in Cycas the ovules develop on individual, leaflike megasporophylls in what is regarded as a primitive arrangement. The microspores of all cycads develop in microstrobili.

The microspores reach the three-celled stage of development of the male gametophyte before they are shed as pollen grains from the microsporangia. At this time, elongation of the megastrobilus separates the megasporophylls, and the wind-borne pollen grains have access to the micropyles of the ovules. At the time of pollination each ovule exudes a mucilaginous droplet, the pollination droplet, through the micropyle; some of the pollen grains become engulfed in this droplet and are drawn into the ovule.

The interval between pollination and fertilization is several months in cycads. The sperm are multiflagellate and are among the largest (about 300 microns, or 0.01 inch) in the plant kingdom. Each pollen tube may contain from two to 22 sperm, depending on the genus. The pollen tubes, which develop from the pollen grains, work their way through the megasporangium of the ovule to the archegonia of the female gametophyte. Fertilization of the eggs of the several archegonia is followed by the early development of several embryos (polyembryony), only one of which survives in the mature seeds. Cycad embryos produce two seed leaves, or cotyledons. The seeds are brightly coloured (yellow or scarlet) and covered by an outer fleshy layer and a stony layer of the integument. The seeds of some cycads (*e.g.*, Cycas) may germinate in the megastrobilus without a period of dormancy.

Systems of the ginkgo and the conifers

The maidenhair tree, or ginkgo (Ginkgo biloba), is sometimes classified with the conifers (see below) or separately in a group of which it is the sole living representative. The mature ginkgo (sporophyte) produces microstrobili and ovules each spring as the buds unfold. They occur on the spur shoots among the bases of the young leaves. The ginkgo, like the cycads, is strictly dioecious, so that some trees produce ovules and others produce pollen. The ovules occur in pairs at the tips of stalks that emerge among the leaf bases.

Ginkgo pollen, like that of pines, is four-celled at the time of pollination (spring season), which is accomplished by wind. Development of male and female gametophytes is similar to that in cycads, and the sperm are also multiflagellate. The female gametophyte, within the ovule of Ginkgo, is unique among seed plants in containing chlorophyll. The ovules enlarge tremendously after pollination, and as the seeds mature the integument differentiates into several coats, of which a stony layer and outer fleshy layer are most prominent. The latter becomes mottled, purplish green and foul smelling. Its tissues may cause nausea or skin eruptions in man. The inner tissues of the seed (the embryo and the female gametophyte) are palatable and prized among some peoples. Fertilization often occurs after the ovules have fallen from the trees, three or four months after pollination. The ginkgo embryo has two cotyledons.

The sporophytes of most of the species of living conifers, like those of the ginkgo, are woody trees at maturity. They usually grow for a number of years beyond the seedling stage before they mature and produce seeds.

The sporophyte of a typical conifer, such as a pine, may become a large tree. Unlike the cycads and ginkgo, a pine is monoecious, both microstrobili and megastrobili occurring on the same tree. At the beginning of each growing season, the microstrobili enlarge and emerge from their bud scales; they are borne at the base of the terminal bud, which is destined to develop into the current season's growth. The megastrobili, by contrast, arise singly or in a whorl near the apex of the current season's growth.

The microstrobili are called simple strobili because the microsporangia are borne in pairs on the appendages (microsporophylls) that emerge from the axis of the strobilus. The megastrobili, however, are compound, for the ovules are borne in pairs upon the upper (adaxial) surface of scales, which, in turn, are borne on bracts attached to the megastrobilus.

The pollen of pine, four-celled when shed, is characterized by two lateral, air-filled "wings," enlarged cavities between two layers of the pollen-grain wall. The pollen is produced in large amounts and may be transported great distances by air currents. During the time of pollination, the ovuliferous scales on the megastrobili separate slightly, and pollen can be trapped in the pollination droplet of the micropyles of the ovules. Pollen grains that make contact with a droplet are transferred by its subsequent contraction through the micropyle and to the surface of a small depression (pollen chamber) at the tip of the megasporangium.

As the pollen grain germinates, forming a tube that works its way through the megasporangium, it arrives at the female gametophyte as the latter matures its several archegonia. The pollen tubes discharge their sperm nuclei into the archegonia, and fertilization is accomplished. As in the cycads and ginkgo, the zygotes of several archegonia may initiate embryogeny. Furthermore, in pine and certain other conifers, the young embryos may form several embryos. At maturity of the seed, however, only one embryo is normally present, embedded in the remains of the female gametophyte and megasporangium, all surrounded by the seed coat (the former integument).

The reproductive process in pine occupies two full growing seasons: ovules pollinated in the spring of a given year do not mature as seeds until the late summer of the next year. The interval between pollination and fertilization is about 14 months.

Reproductive variations in gymnosperms

Among the numerous other gymnosperm species are many different reproductive processes. Some gymnosperms, for example, are dioecious, with microstrobili and megastrobili being borne on separate plants, as in junipers (Juniperus), plum yews (Cepkalotaxus), yews (*Taxus*), and podocarps (Podocarpus). Furthermore, in larches (Larix) and other groups the pollen grains lack wings. The pollen grains in larches become attached at pollination to a special receptive enlargement of the integument. In podocarps, the megasporangium bulges through the micropyle at pollination and receives the pollen directly. The interval between pollination and fertilization may be as short as four to five weeks in firs (Abies). The number of ovules formed on the ovuliferous scale varies, as does the number of microsporangia on the microsporophyll. There may be only one ovule in a megastrobilus, as in some junipers, and the megastrobili may become fleshy, also in junipers. In yews the solitary ovules are terminal on dwarf shoots; each ovule is surrounded by a cuplike structure called an **aril,** which becomes fleshy and brightly coloured as the seed matures. The number of sperm produced in each male gametophyte varies also—from two in pine to 20 in some cypresses (*Cupressus*).

The genera Ephedra, Gnetum, and Welwitschia, which are often grouped together in one category (Gnetales, or Gnetophyta), differ among themselves and from other gymnosperms with respect to several details of reproduction. The microsporangia and ovules of both Ephedra and Welwitschia are produced in compound strobili; those of Gnetum are borne in a series of whorls on elongated axes sometimes called misleadingly "inflorescences." The ovules of these genera, unlike those of other gymnosperms, have two integuments instead of one, as in angiospermous ovules. Archegonia are present in the female gametophytes of Ephedra, but only eggs occur in those of Gnetum and Welwitschia. The sperm, like those of the conifers, lack flagella.

Angiosperms. Although the angiosperms are known as flowering plants, they are difficult to distinguish from gymnosperms solely on the basis of bearing flowers, for, like the strobilus, a flower is a compressed stem, with crowded, spore-bearing appendages. The occurrence of coloured petals and attractive scents is not essential and is by no means characteristic of all flowers. The most important distinguishing feature separating flowering plants from gymnosperms is that the ovules of flowering plants are produced within enclosed contaniers called carpels (Figure 7).

Flowers may occur singly at the ends of stems (*e.g.,* tulip, poppy, rose), or they may be grouped in various clusters or inflorescences (gladiolus, sunflower, delphinium, and yucca).

An individual flower may be complete, in that a given floral receptacle produces sepals (often greenish and leaflike), petals (often white or coloured other than green), stamens, and a pistil (or pistils). The sepals are collectively known as the calyx, and the petals as the corolla; the calyx and corolla comprise the perianth. If sepals or petals are lacking, the flower is said to be incomplete. Although incomplete, a flower that has both stamens and a pistil is said to be perfect; lacking either of these parts, it is imperfect.

In practice, groups of solitary flowers are not easily distinguished from inflorescences; the latter seemingly evolved from a system of branches, each with a terminal, solitary flower. The inflorescence may be few flowered or have up to **6,000,000** flowers, as in certain palms. Inflorescences vary also in their position, being terminal, axillary, or intercalary. Terminal inflorescences are at the tips of the major or dominant branches; axillary ones are at the tips of axillary, or side, branches. In intercalary inflorescences, the stem continues beyond the inflorescence, which may result in alternating fertile and sterile areas of the axis.

Reprinted from The Spectrum of Life by Harold A. Moore and John R. Carlock; Copyright © 1970, Harper & Row. Publishers. Inc.

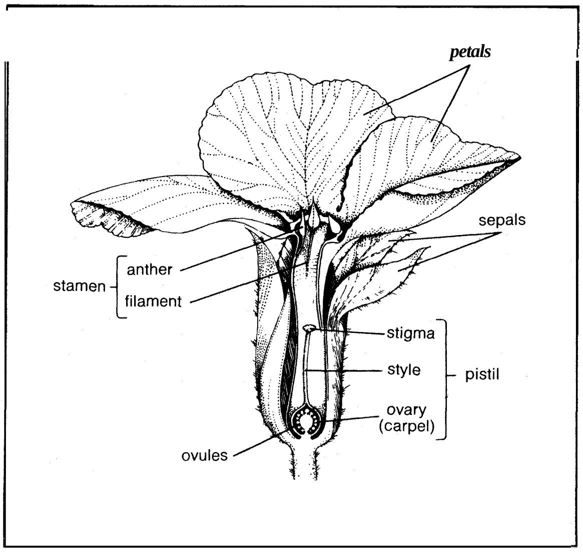


Figure 7: Gross morphology of the flower.

Types of floral clusters

Inflorescences can be distinguished by their growth patterns as determinate or indeterminate. In determinate inflorescences the first formed flower at the tip of the dominant stem matures first, and younger flowers develop on lower lateral branches; the cyme of the forget-me-not (Myosotis) is a typical example. In indeterminate inflorescences the growing region of the axis functions for extended periods so that as the older flowers mature and set fruit near the base of the inflorescence axis, younger buds develop and continue to expand into flowers at the apex. This is exemplified in the spikes of yucca and the racemes of delphinium, in which the youngest flowers are farthest away from the root. Other types of indeterminate inflorescences include umbels and capitula, or heads. The youngest flower is terminal or central in umbels and in heads.

The head is the type of inflorescence that characterizes the composite, or aster family. It may be few to many flowered and usually has at its base one or more series of leaflike bracts. The small individual flowers arise in spiral order on the receptacle, the youngest being at the centre. The basal calyx of each flower, known as a **pappus,** is bristle-like, scaly, or feathery and borne at the top of the ovary. The corolla, formed of the petals, may be: (1) tubular, with five petal lobes, sometimes split open; (2) ligulate, or tonguelike, with a very short basal tube; or (3) bilabiate, with the tube split into two tips. In some genera, all of the flowers are ligulate, whereas in others, the marginal flowers are ligulate (ray flowers) and the others tubular or all are tubular. The marginal ray flowers are either female (pistillate) or sterile. The tubular flowers are characterized by male and female parts: five united pollen-bearing stamens and a pistil, which matures as a one-seeded fruit (achene).

The position of the floral organs with reference to each other and to the tip of the floral receptacle varies in different flowers; in some, the perianth (sepals and petals) and stamens are attached to the receptacle below the pistil; such flowers are hypogynous (*e.g.,* buttercup and magnolia). In others (rose, cherry, peach), the perianth and stamens are borne on the rim of a concave structure in the depression of which the pistil is borne; such flowers are perigynous. Finally, there are flowers in which the ovary is enclosed by a tissue composed of the fused bases of the perianth and stamens (apple, pear, aster); the blossom seems to arise upon or above the ovary and is called epigynous.

The stamen, seemingly the equivalent of the gymnospermous microsporophyll, consists of an anther (a group of two to four microsporangia) borne at the tip of a blade stalk, or filament. The pistil, most often composed of an enlarged basal ovary, a columnar style, and distal stigma, is the ovule-producing organ of the flower. It is often considered to have evolved from enfolded megasporophyll or some other ovuliferous structure with enclosed ovules (angiospermy); alternatively, it is thought to have arisen from the cuplike bracts of extinct seed-bearing plants on which the leafy bracts grew together and thus enclosed the ovules.

There may be one or more pistils on the floral receptacle, depending on the species. Furthermore, pistils may be simple (composed of one ovule-bearing unit, megasporophyll, or carpel) or compound (composed of more than one carpel). Compound pistils are thought to have arisen as a result of crowding of simple pistils on the floral axis; for example, variation in the degree of fusion may be observed in members of the saxifrage family. The ovary —which matures as the fruit—usually reveals by the number of ovule-containing chambers (locules) the number of carpels it contains. The stigma is a specially adapted portion of the pistil modified for the reception of pollen. It may be feathery and branched or elongated, as in such wind-pollinated flowers as those of the grasses, or it may be compact and with a sticky surface. The ovary may contain one ovule (*e.g.*, buckwheat, avocado), a few ovules (*e.g.*, grape, bean) or a large number of ovules (tobacco, begonia, snapdragon).

In some angiosperms (*e.g.*, corn, hickory, walnut, pecan, oak) both types of imperfect flower are borne on the same plant, which is, therefore, called monoecious. By contrast, staminate flowers may occur on one plant and pistillate flowers on another, as in willows, poplars, and mulberries, which are said to be dioecious. In common parlance (and unfortunately in some botanical textbooks) staminate flowers and plants that bear them are often designated "male," and pistillate flowers and the plants that bear them are called "female." This may be traced back at least as far as to the time of Linnaeus (1753), who interpreted stamens and pistils as sex organs. Comparative morphology indicates clearly, however, that stamens and pistils are the spore-bearing structures of the sporophyte and not actually the gamete-bearing organs of the gametophyte. The terms "male" and "female," applied to angiosperm plants and their flowers, is often condoned because the gametophytic phase is so condensed in angiosperms. The designations suggest to the uninitiated, however, that pollen grains and sperm, on the one hand, and eggs and ovules, on the other, are identical, which is not the case.

Floral organization

Among the vast number of species of angiosperms there is considerable variation in floral organization. The perianth may be absent or present; clearly differentiated as calyx and corolla (*e.g.*, pea); or the perianth segments may be similar (magnolia, tulip tree). The number of stamens and pistils may be large and separately attached to the receptacle in a spiral pattern (buttercup), or the numbers may be reduced and the attachment cyclic or whorled (lily). The stamens may be fused together by their anthers (daisy) or their filaments (peas, beans). The filaments may be petallike (water lilies) or stalklike. Opening of the anther may be by longitudinal or transverse fissures or by terminal pores.

The reproductive cycle in angiosperms can be traced from before the shedding of pollen (Figure 8). The microspores begin their development of male gametophytes, which involves formation of a small generative cell and a tube cell. The generative cell may divide to form two sperm before the pollen grain (developing male gametophyte) is shed, or while the pollen tube is growing during germination. The pollen grains of angiosperms have variously, and often elaborately, ornamented walls characteristic of the species.

Pollination in angiosperms is the transfer of the pollen grains from the anther of a stamen to the stigma of a pistil. It may be accomplished by the force of gravity, wind, or water or by agency of insects, birds, or bats. A pollen grain that is trapped on the stigma may germinate, the pollen tube growing rapidly through stigma and style into the ovary, where it enters an ovule.

From (top) P.B.Weisz, The Science of *Biology*, 4th ed. (1971), used with permission of McGraw-Hill Book Company; (bottom) P.B.Weisz, The Science of *Biology*, 3rd ed. (1967), used with permission of McGraw-Hill Book Company

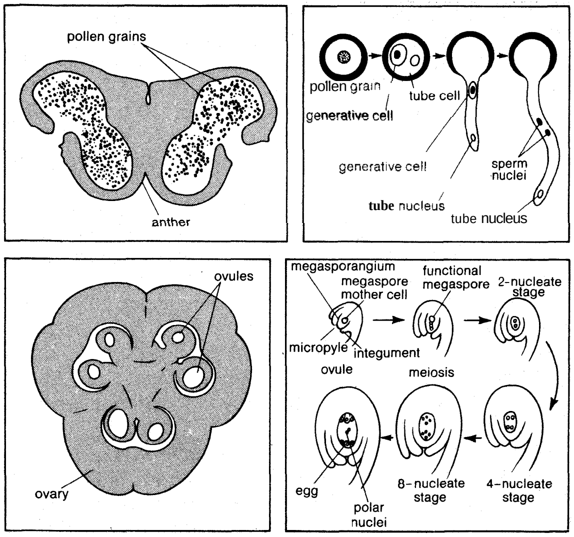


Figure 8: Gametophyte development in a flowering plant (lily). (Top left) Cross section through the anther of a lily. (Top right) Microgametophyte from the microspore. (Bottom left) Cross section through the ovary of a lily. (Bottom right) megagametophyte from the megaspore.

The pistil of a flower may receive pollen from the stamens of the same flower, in self-pollination (*e.g.*, peas and tomatoes). In many other flowers, however, pollen from one or more flowers is transferred to the stigmas of other flowers. A number of specialized relationships have evolved between floral organization and animal pollinators such as insects (see POLLINATION).

In the majority of angiosperms one megasporocyte develops in the megasporangium (often called the nucellus) of the ovule, and a tetrad of megaspores is formed as a result of meiosis. Three megaspores (nearest the micropyle) degenerate; only one enlarges, and then undergoes divisions to form the eight-nucleate, seven-celled female gametophyte ("embryo sac"). Of the three cells of this gametophyte near the micropyle, one functions as an egg. As the pollen tube discharges its contents into the female gametophyte, the egg nucleus is fertilized by one of the sperm, the other unites with the two nuclei (polar nuclei) within the large central cell of the female gametophyte. The resultant nucleus, which has three sets of chromosomes, is the primary endosperm nucleus. This process, double fertilization, occurs only in angiosperms.

Pattern of development

Both pollination and fertilization stimulate cell division in the ovary, ovules, and zygotes, all of which enter upon a period of rapid enlargement. In most angiosperms, the primary endosperm nucleus divides to form endosperm tissue, the cells of which become filled with stored food, such as starches, oils, and proteins. As the rate of embryonic development decreases, the seeds of most angiosperms enter a period of dormancy, accompanied by dehydration and hardening of the integuments, which form seed coats. At this period, the enlarged ovary (and sometimes adjacent structures) matures as fruit.

Angiosperm seeds may germinate as soon as they reach maturity, or they may undergo various kinds of dormancy. In some cases (*e.g.*, coconut) the embryo is rudimentary and undifferentiated when the seed is shed, so that a period of preparation, or after-ripening, is required. In other types of dormancy, germination is retarded by the hardness and impermeability of the seed coat or by special requirements of light, temperature, and moisture.

Some representative variations occur in the reproductive process of angiosperms. In violets (Viola), in addition to the ordinary flowers produced first during the usual flowering season, less conspicuous flowers later develop; called cleistogamous flowers, they do not open but are self-pollinated, thus insuring augmentation of the population during a period less favourable for the usual blossoms.

The pollen grains of most angiosperms separate from each other, but in some cases (*e.g.*, Rhododendron) they remain attached in original groups of four, called tetrads. The very tiny pollen grains of orchids, certain mimosas, and milkweeds are clustered in waxy masses called pollinia (singular pollinium).

A number of variations in pattern of development of the female gametophyte occur in various angiosperms; for example, in certain species of evening primrose (*Oenothera*), the female garnetophyte contains only four nuclei, whereas, in Peperomia, as many as 16 may be present. In lily, all four megaspore nuclei are involved in the formation of the female gametophyte.

Pollen may germinate immediately after contact with a stigma (sugar cane), within five minutes (corn), in two hours (beet), or after one or two days. The pollen grains of most plants produce only one pollen tube, but ten or more pollen tubes have been observed to develop from one pollen grain in plants of the mallow family. The pollen tubes usually enter through the micropyle (porogamy), but they may also enter through the base of the ovule (chalazogarny).

The interval between pollination and fertilization varies. It may be as long as 12 to 14 months in certain species of oak, five to seven months in witch hazel; two to 20 weeka among the orchids; three to four hours in lettuce and as little as 15 to 45 minutes in dandelions.

The postfertilization endosperm fails to develop in orchid seeds but is present at least during early embryogeny in most others. The endosperm may arise by nuclear divisions and become cellular as nuclear divisions terminate, or its development may involve both nuclear and cell divisions from the beginning. In a number of cases (*e.g.*, legumes) the embryo consumes the endosperm during its development, resulting in mature seeds with massive embryos and no endosperm. Most angiosperm embryos have two seed leaves (are dicotyledonous), some have one lateral cotyledon (are monocotyledonous), and a few (*e.g.*, Degeneria) have three to four cotyledons.

In seed germination, the cotyledons may remain below the soil surface within the seed (hypogean germination) and may function in digesting and absorbing endosperm (corn), may serve as sources of stored food themselves (pea!, or may rise above the soil surface (epigean germination) by elongation of the hypocotyl, the embryonic axis between the root, the growing stem, or epicotyl. Cotyledons that emerge above the soil may wither and drop off as their food is used (*e.g.*, bean), or they may persist and function as photosynthetic leaves (*e.g.*, castor bean).

An even greater range of variation occurs in angiospermous fruits. The fruit may arise from one pistil (simple or compound) of one flower (*i.e.*, the simple fruits of pea and peach), from several pistils of one flower (*i.e.*, the aggregate fruits of strawberry and raspberry), or from the pistils of several flowers (*i.e.*, the multiple fruits of pineapple, mulberry, and corn). Simple fruits may be dry (legumes) or fleshy (peach, apple, tomato) at maturity. Dry fruits may open (dehisce; many legumes) or remain closed about the seed (be indehiscent; grasses and sunflower).

The manner of ovular attachment is known as placentation. The ovary may contain one to many ovules, which may be attached to the ovary wall (parietal placentation) or to the central axis (axial, or free-central placentation). Despite these and other variations in the morphology of flower parts, the reproductive process is, with minor diversities, remarkably uniform.

VARIATIONS IN REPRODUCTIVE CYCLES

The life cycles and reproductive processes described above characterize the vast majority of their respective plant groups.

Among the liverworts it has been demonstrated that small fragments of the stalk of the sporophyte are capable of regenerating diploid gametophytes. In the mosses, both haploid and diploid apospory have been experimentally evoked. As in the liverworts, injury and regeneration of fragments of the sporophytic seta result in diploid gametophytes. By contrast, fragments of moss leaves, stems, and rhizoids (and even the sterile tissues of the sex organs) can regenerate haploid gametophytes.

In certain strains of mosses, the gametophyte can give rise to clusters of presumably haploid sporophytes without the functioning of gametes; such apogamous formation of sporophytes may also be chemically induced (by application of a solution containing a specific amount of chloral hydrate to both the protonema and leafy shoots).

Apogamy, apospory, and apomixis

Among the vascular plants both natural and induced apogamy and apospory are known. In certain ferns, gametophytes may develop at the leaf margins or in sori from transformed sporangia. Certain other ferns reproduce apogamously in nature; thus, for example, in the holly fern (*Crytomium falcatum*), the gametophytes give rise directly to sporophytes by nuclear and cell division on vegetative cells of the gametophyte. In almost every group, however, variations of the usual reproductive process occur. These may involve substitution of asexual reproduction for sexual or the direct production of plants by cells other than the usual ones (apomixis). Apomictic phenomena—which are in the strictest sense asexual—include apospory, in which the gametophyte phase is produced without the need of spores, and apogamy, in which the sporophyte phase is produced without the need of gametes, or sex cells.

Apogamy may be induced in normally sexual ferns by withholding water from the gametophytes, which prevents the liberation and functioning of sperm. Similarly, when gametophytes are grown in inorganic culture media supplemented by a variety of sugars, they produce sporophytes apogamously. Colourless roots removed from the bracken fern (Pteridium aquilinum) have been induced to develop diploid gametophytes aposporously, as have the injured juvenile leaves of a number of ferns.

Apomictic phenomena occur also among many angiosperms. In some species, haploid sporophytes may develop either from the unfertilized egg or from some other cell of the gametophyte. Such apogamy occurs, for example, after stimulation of one species with the pollen of a related one (*e.g.*, *Solanum nigrum* by the pollen of S. *lu*teum). Apogamy involving an unfertilized egg (a phenomenon termed parthenogenesis) occurs in certain orchids. Male parthenogenesis, or the production of a sporophyte from a sperm, has been detected in tobacco hybrids. Finally, a form of haploid apogamy is known in which a cell of the female gametophyte other than an egg may develop into an embryo.

In certain species of hawkweed, the embryo develops from a certain cell of the ovule or the megasporangium. In others, the female gametophyte is diploid through an impairment of the meiotic process; in this case, the egg (diploid parthenogenesis) or one of the related cells may form an embryo. In citrus trees a number of embryos (polyembryony) arise from diploid cells of the megasporangium or integuments.

PHYSIOLOGY OF REPRODUCTION

The maturation of sporophytes and gametophytes, as manifested by their ability to produce spores and gametes respectively, involves both internal and environmental factors. With respect to the former, the organism must have completed a certain minimum period of vegetative development before environmental factors are able to stimulate formation of spores and gametes.

Effects of light and temperature

Among environmental factors affecting reproduction, the duration, intensity, and quality of light, as well as temperature, have primary roles; for example, the liverwort Marchantia polymorpha continues in the vegetative state indefinitely under daily fluorescent illumination of 16 hours. Control plants exposed to daily incandescent lighting of 16 hours become sexually mature after 30 days. Addition of the sugar sucrose hastens the development of the sexual phase, an indication that chemical factors also play a role. In hornworts (*e.g.*, Anthoceros and Phaeoceros), antheridia develop under daily light periods (photoperiods) of four to 12 hours; none develop when the plants are illuminated for longer periods.

Temperature also affects sexual maturation. In mosses, for example, initiation of sex organs in bacteria-free laboratory cultures of *Funaria* occurs at 10° C (50" F) when cultures are illuminated six, 12, or 20 hours daily. On the other hand, the moss Polytrichum is seemingly not affected by duration of light but forms sex organs best at 21° C (70° F).

Among vascular cryptogams little is known about the various environmental factors that affect development of sex organs in nature, and, except for certain ferns and horsetails, experimental studies of rigidly controlled laboratory cultures are lacking. It has been shown that laboratory cultures of gametophytes of certain horsetails derived from single spores produce sex organs 40 to 60 days after the spores germinate. In horsetails generally, conditions favouring vegetative growth evoke a preponderance of initially female gametophytes; less favourable conditions induce the production of male gametophytes.

Considerable information is available about the physiology of reproduction in ferns, especially with respect to the gametophyte generation. Spore germination occurs if adequate moisture is present and in temperatures between 15° and 35° C (59° and 95" F). The spores germinate and the young gametophytes do best under neutral or slightly acid conditions. Most fern spores require light for germination. Some fern spores remain viable for as long as 20 years.

Spores germinating in darkness or in the absence of blue wavelengths of light remain in the protonemal, or filamentous, condition. Growth to the mature gametophyte requires light of blue wavelengths in most species.

An interesting aspect of fern reproductive physiology was the discovery that antheridial production is under hormonal control. Several types of antheridiurn-evoking substances (hormones) have been recognized, and there is some evidence that they are specific. The sperm of ferns are attracted chemically to the vicinity of the maturing archegonia, where they become trapped in extruded mucilage. Although only one sporophyte usually develops on a gametophyte, as many as five may be induced to form if the gametophyte is repeatedly exposed to sperm.

The preceding paragraphs, dealing with vascular cryptogams, have emphasized the influence of environmental factors on the expression of sexual reproduction; the remaining account deals with the angiosperms. Because the gametophytic phase of angiosperms is abbreviated and because it is parasitic on the sporophyte, the maturation of the sporophyte, as manifested by flowering, has received more intensive study in angiosperms than has that of the gametophyte.

Durations of life cycles

Angiosperms that complete their life cycles within one growing season (in the temperate zone) are known as annuals. Perhaps the shortest angiospermous life cycle known is that of *Plantago insularis,* a southern California species, that can grow from seed to the production of its own seed in four to six weeks under domestication; the cycle of most annuals, however, is longer. A number of angiosperms are biennial in the temperate zone: they grow vegetatively for one season, but their flowering and seed production are delayed until a second growing season, after which the plants die (*e.g.*, beets, carrots). Still other angiosperms are perennial: they continue growing, flowering, and producing seeds for a number of growing seasons (*e.g.*, irises, roses, oaks).

It has been demonstrated that duration and wavelength of light are of paramount importance in controlling the flowering of angiosperms. Based on extensive experimental studies, flowering plants have been classified as "long-day," "short-day," or "day-neutral" with respect to their requirements of light for flowering (see PHOTOPERIODISM).

Temperature also plays an important role in flowering. Thus, members of the cabbage family can be grown without flowering when high enough temperatures are maintained. A number of perennial and biennial plants (bulbous plants, beets) require a prior period of low temperature before they flower. In addition to internal (*i.e.*, genetic factors), therefore, temperature and light are of paramount importance for many plants in evoking maturation.

Several other physiological aspects of reproduction in angiosperms are noteworthy. Once transported to the stigma, the germination of the pollen grains is markedly affected by the chemical composition of the stigmatic exudate; *e.g.*, pollen of an unrelated species will not germinate on the stigma. There is evidence also that the directional growth of the pollen tubes through the style toward the ovular micropyle is stimulated by a chemical substance produced by the style, ovules, and probably other tissues of the pistil.

It is also clear that both pollination and fertilization have physiological effects on fruit production. In a number of cases pollination alone (not followed by fertilization) is sufficient stimulus to evoke enlargement of the pistil to form a seedless (parthenocarpic) fruit. This phenomenon occurs in the banana and in certain varieties of citrus fruits, grapes, and cucumbers. Parthenocarpy can also be induced by exposure of the stigma to indoleacetic acid, naphthoxyacetic acid, indole butyric acid, naphthaleneacetic acid, or other synthetic hormones. The hormone gibberellin is effective in producing seedless grapes and is the active component in preparations used to prevent premature dropping of fruit.

Also significant in the reproduction of flowering plants are the phenomena of self-sterility and self-incompatibility, which prevail in certain plants with perfect flowers (having both stamens and pistils). Pollen from one flower or from another flower of the same plant or from a plant with identical genetic constitution, when applied to the stigma, fails to germinate or grows so slowly that fertilization does not occur. For example, sweet cherries are self-incompatible: a number of groups of genetically intrasterile types are known, but cross-pollination of trees of different genetic groups results in fruit production. Self-incompatibility occurs also in certain strains of garden plants and in plum, apple, pear, and tobacco.

Self-sterility, usually associated with differences in chromosome number, often occurs in hybrids from widely divergent parents. This is true of cultivated varieties of blackberry and raspberry. Such chromosomal imbalance creates irregularities in meiosis during formation of the pollen, which results in infertility.

BIBLIOGRAPHY. D.W. BIERHORST, *Morphology of Vascular Plants* (1971), a treatment of the vascular plants at an advanced level, with many photographs and an extensive bibliography; H.C. BOLD, *The Plant Kingdom,* 3rd ed. (1970), a brief account of the structure and reproduction of representative plant types, and *Morphology of Plants* (1967), *a* profusely illustrated textbook designed for the advanced undergraduate and beginning graduate students; A.S. FOSTER and E. M. GIFFORD, *Comparative Morphology of Vascular Plants* (1959), a textbook account of the structure of vascular plants; R.F. SCAGEL *et al., An Evolutionary Survey of the Plant Kingdom* (1965), a multi-authored treatment of the plant kingdom, with abundant illustrations; K.R. SPORNE, *The Morphology of Gymnosperms* (1965), a brief but recent summary of the structure and reproduction of gymnospermous seed plants, living and fossil, and *The Morphology of Pteridophytes,* 2nd ed. (1966), a short but broadly conceived treatment of vascular cryptogams, living and fossil.

(H.C.B.)

Reptilia

Reptilia is a class of vertebrate animals that includes the snakes, lizards, crocodiles, alligators, turtles, and the tuatara, among the living forms, and a great many extinct types such as dinosaurs, pterosaurs, and ichthyosaurs.

Intermediate between amphibians and the warm-blooded vertebrates, reptiles may be described as those air-breathing vertebrates with internal fertilization and a scaly body covering instead of hair or feathers.

GENERAL FEATURES

Importance to man. The economic and ecological importance of reptiles to man is not great, particularly in comparison to the other major vertebrate groups—birds, fishes, and mammals. Locally, some species are eaten on occasion if not regularly. Turtles, in particular, are popular. The green turtle (*Chelonia rnydas)* is the most widely eaten species of reptile. The giant Galápagos tortoise was especially popular as food among 19th-century seafarers and, for this reason, nearly became extinct. Among the lizards, iguanas are perhaps the most popular as a local food. Many species of snakes are eaten locally; canned rattlesnake meat has limited popularity as a delicacy.

Leather goods, including luggage, gloves, belts, handbags, and shoes, are made from the skins of lizards, crocodilians, and snakes. This has led to the virtual extinction of several species of crocodilians and to severe reduction of populations of large lizards, snakes, and turtles. As living subjects for biological research, lizards in particular have been useful to the scientist. Venomous species constitute little hazard to man except in limited rural areas.

Size range. The maximum size of the largest snake or largest crocodile is considerably less than estimated by excitable and gullible travellers. Persistent but unsubstantiated reports have been made of 12-metre (40-foot) anacondas. This gigantic South American snake, although it probably is the largest living species, does not usually exceed nine metres (30 feet) in length. The reticulated python of Southeast Asia and the East Indies has been recorded at 8.4 metres. The rock python *(Python sebae)* of Africa reaches 7.5 metres. No other group of living

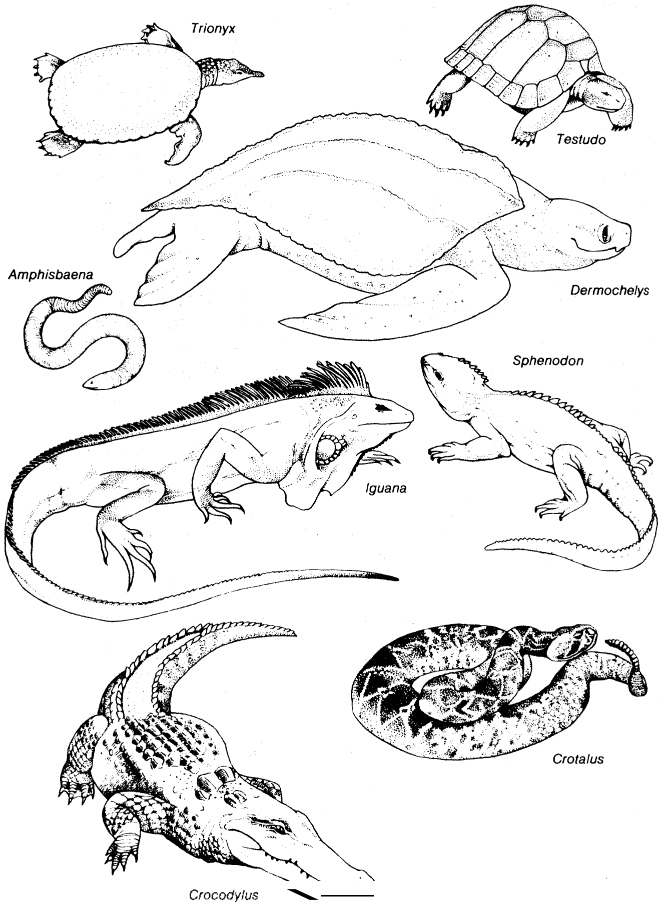


Figure 1: Body plans of living reptiles.

Drawing by J. Helmer based on (Crotalus) photograph courtesy of the National Marine Fisheries Service; (Crocodylus) Reader's Digest Living World of Animals; (others) J.Z. Young, The Life of Vertebrates, The Clarendon Press, Oxford

snakes approaches the pythons and boas in weight, although the king cobra of Asia and the East Indies comes close in length (5.4 metres) and is the longest venomous snake. The heaviest venomous snake is probably the eastern diamondback rattlesnake (Crotalus *adamanteus*), which, though not exceeding 2.4 metres, may weigh as much as 15.5 kilograms (34 pounds). The largest of the common nonvenomous snakes of the family Colubridae is probably the Oriental rat snake, Ptyas carinatus (3.73 metres).

Four living species of crocodilians grow larger than six metres: the American crocodile (Crocody*lus* acutus), the Orinoco crocodile (C. *intermedius*), the saltwater crocodile (C. porosus), and the gavial (*Gavialis* gangeticus). The last two may approach nine metres.

The giant among living turtles is the marine leatherback (*Dermochelys*), which reaches a total length of about 2.7 metres and a weight of about 680 kilograms (1,500 pounds). The largest of the land turtles is a Galápagos tortoise weighing 255 kilograms (560 pounds).

The largest modern lizard is a monitor, the Komodo dragon of the East Indies; it attains a length of three metres (ten feet). Two or three other species of monitors reach 1.8 metres. The common iguana comes close to that size, but no other lizard does.

None of the living reptiles, with the possible exception of snakes, is as large as the largest extinct representative of its particular group. The 2.7-metre leatherback turtle is smaller than the extinct 3.6-metre marine turtle *Arche*lon, and no modern crocodile approaches the estimated 15-metre length of Phobosuchus. The Komodo dragon does not compare with the six-metre or more mosasaur Tylosaurus. Extinct dinosaurs such as Brontosaurus (Apatosaurus), Diplodocus, and Brachiosaurus grew to about 24 metres (80 feet) and probably weighed close to 45,000 kilograms (100,000 pounds).

The smallest reptiles

The smallest reptiles are the geckos, some of which grow no longer than three centimetres (slightly more than one inch). Certain blind snakes (Typhlopidae) are less than ten centimetres (four inches) in length when fully grown. The smallest turtles weigh less than 450 grams (one pound) and reach a maximum length of 12.5 centimetres (about five inches). The smallest crocodilians are the dwarf crocodile (Osteolaemus tetraspis) and the smooth-fronted caiman (Paleosuchus), which grow to about 1.7 metres in length.

Distribution and ecology. North Temperate Zone. Although living reptiles, which number some 6,000 species, are primarily tropical animals, many inhabit the temperate zones. The northernmost ranges are those of the lacertid lizard (Lacerta vivipara) and the common viper (Vipera berus), both of Europe and Asia. These ovoviviparous (live-bearing, but the unborn snakes develop within eggs) species live north of the Arctic Circle, at least in Scandinavia. Two other lizards, the slowworm (*Anguis fragilis*) and the sand lizard (Lacerta agilis), and two snakes, the grass snake (Natrix natrix) and the smooth snake (Coronella austriaca), reach 60° N in Europe. Of the six northern species, all but the grass snake are ovoviviparous. Across Siberia only Lacerta vivipara and Vipera berus live north of 60".

In North America no reptile reaches the 60th parallel. Two species of garter snakes live as far north as 55° in western Canada. In North America and Eurasia the northern limit of turtles is about 55° N.

It is only south of 40° N that reptiles become abundant. In the eastern United States and eastern Asia, water snakes (Natrix), rat snakes (*Elaphe*), racers (Coluber), green snakes (Opheodrys), northern skinks (*Eumeces*), glass "snakes" (lizards of the genus Ophisaurus), and soft-shelled turtles (Trionyx) are common. One of the two living species of Alligator lives in southeastern United States; the other lives in China. Even though both regions are characterized by many species of emydid turtles (family Emydidae), the genera to which the species belong are found in only one region or the other. Many lizards of temperate Eurasia belong to the families Agamidae and Lacertidae, which do not occur at all in the Americas. On the other hand, many lizards of North America are in the families Iguanidae and Teiidae, which do not live in Eurasia.

The fauna of the eastern United States is almost as distinct from that of western United States and northern Mexico (which is faunistically part of the same region) as it is from that of eastern Asia. The eastern United States has many genera and species of emydid turtles; the western United States (defined by a diagonal line running southeast to northwest through Texas, then northward along the Continental Divide) has only four or five species. Few genera and species of iguanid lizards inhabit the eastern United States, whereas the western United States has many. Although the eastern United States has more species of water snakes, the western United States contains more garter snakes. More species of snakes appear in the eastern United States than in the western areas, while the converse is true of lizard species.

Ecological types of the North Temperate Zone

Reptiles of the North Temperate Zone include many ecological types. Aquatic groups are represented in both hemispheres by the water snakes, many emydid turtles, and the two alligators. Terrestrial groups include tortoises, ground-dwelling snakes, and many genera of lizards. Arboreal snakes are few, and arboreal lizards are almost nonexistent. Burrowing snakes are common. Specialized burrowing lizards are few.

Central and South America. In Central America the reptile fauna becomes richer. Besides several turtle families found in eastern United States, Central America has three genera of turtles (*Dermatemys, Claudius,* and *Stau*rotypus) not living elsewhere. Crocodilians become more numerous both in species and individuals. Lizards and snakes are particularly more abundant.

Many of the genera of iguanid lizards occurring in western United States have species in Mexico; one genus of spiny lizards (Sceloporus) reaches its peak of numbers of species in Mexico. South of Mexico the North American iguanid genera disappear and are replaced by tropical groups such as the black iguanas (*Ctenosaura*), the helmeted iguanids (*Corythophanes*), the casque-headed iguanids (Laemanctus), and the basilisks (Basiliscus). The lizard family Teiidae, though represented in the United States by the race-runner genus (*Cnemidophorus*), is tropical, and its real development begins in Central America with the large, conspicuous, and active ameivas (Ameiva) and several small genera that live in concealment.

Among snakes, the fer-de-lance genus Bothrops, the coral snakes (Micrurus), the rear-fanged snakes such as the cat-eyed snakes (Leptodeira), and nonvenomous genera such as the tropical green snakes (Leptophis) either appear for the first time or begin their proliferation of species in Central America.

Reptiles become increasingly numerous in northern South America. Vine snakes (Oxybelis and *Imantodes*), false coral snakes (Ery*throlamprus*), slender ground snakes (Drymobius), and the burrowing spindle snakes (*Atractus*) are most abundant there. Most of the genera of the lizard family Teiidae occur in this area. Iguanid lizards of the anole genus (Anolis) are represented in northern South America by approximately 165 species. Other iguanid genera—*e.g.*, the long-legged Polychrus—make their appearance.

Crocodilians, in terms of species, are more numerous in South than in Central America, and turtles are also abundant. Some of the North American groups—for example, the mud turtles (Kinosternon) and sliders (*Chrysemys*)—are represented, but the majority of species are members of genera and even families (*e.g.*, the side-necked turtles, families Pelomedusidae and Chelidae) unknown in temperate North America.

Several groups that form important, if not dominant, elements of the fauna of the Eastern Hemisphere are largely or completely absent from the American tropics: the lizard families Scincidae, Lacertidae, Chamaeleontidae, and Agamidae and the snakes of the cobra (Naja) and water snake (Natrix) genera.

South of the tropics, in the temperate zone of South America, the reptilian fauna diminishes rapidly. Crocodilians and turtles do not occur south of northern Argentina. An ovoviviparous pit viper reaches almost 50° S there; two iguanid lizards range almost to 55° S.

Asia. Apart from the genera of reptiles listed above as common to the eastern United States and eastern Asia, the temperate zone of Eurasia is noted for its many lizards of the families Agarnidae, Lacertidae, and to lesser degrees Gekkonidae and Scincidae. Most of the lizards are terrestrial; extremely specialized burrowers include desert-dwelling skinks (*Ophiomorphus* and *Scincus*). Most of the snakes characteristic of this vast area are also terrestrial. Arboreal snakes are represented almost exclusively by the rat snakes (*Elaphe*). The leaf-nosed snakes (Lyterkynchus) and the sand boas (Eryx) are the distinctive burrowing snakes of the region. Except for the Chinese alligator and the Indian gavial, temperate Eurasia lacks crocodilians. A few species of turtles are found.

A few types characteristic of the Oriental tropics extend into the temperate zone—*e.g.*, several rear-fanged snakes (Boiga *trigonata* and Psammodynastes), a cobra or two (*Naja*), several species of soft-shelled turtles (Trionyx), and some species of true chameleons (*Chamaeleo*).

In the Oriental tropics the reptilian fauna is extremely rich in species and diverse types. Aquatic groups are represented by snakes of various genera (*e.g.*, Natrix, *Enhydris*, Acrochordus), several groups of lizards (*Tropidophorus* among the skinks and Hydrosaurus among the agamids), many emydid and soft-shelled turtles, and five species of crocodiles. The numerous terrestrial reptiles include the small kukri snakes (*Oligodon*), the big Oriental rat snakes (Ptyas), cobras, monitor lizards (Varanus), many species and genera of skinks, some geckos, and several land turtles (Cuora, Geochelone). Specialized burrowing snakes (*e.g.*, the family Uropeltidae and the colubrid genus Calamaria) and lizards (*e.g.*, the family Dibamidae and the skink genus *Brachymeles*) are also abundant.

Arboreal reptiles of tropical Asia

The distinctive life-forms of reptiles in tropical Asia are arboreal. They include pythons and Oriental pit vipers (Trimeresurus), vine snakes (Ahaetulla), slug-eating snakes (Pareas), "flying" snakes (Chrysopelea), and tree racers (*Gonyosoma*). Some lizards climb only with the aid of claws (*e.g.*, the monitors), a few with the help of prehensile, or grasping, tails (*e.g.*, the deaf agamids, Cophotis), and many with the help of clinging pads under the digits (*e.g.*, many geckos). The most striking arboreal reptiles of this area are the flying lizards (Draco) and the parachuting gecko (Ptychozoon), which has fully webbed digits, a fringed tail, and wide flaps of skin along its sides.

Australia. Because New Guinea, although geographically part of the East Indies, has a reptilian fauna more akin to that of Australia, the two areas are considered here as one. The Australian region is the only area in the world in which venomous species of snakes outnumber harmless ones. The family Colubridae, comprising the majority of the nonvenomous or slightly venomous snakes of the world, is poorly represented in Australia, which has only 12 species. The Australian region has many snakes of the cobra family (family Elapidae), but no vipers. The fauna also include several pythons and minute blind snakes (family Typhlopidae); a variety of geckos, skinks, and agamid lizards; side-necked turtles; and three species of crocodiles.

Africa. The reptilian fauna of Africa forms two main divisions. The first, the fauna of the North African coast, is akin to that of central and southwestern Asia and southern Europe and is therefore mainly a temperate-zone fauna. The racers, the burrowing sand skink (*Scincus*), and the emydid turtle (Mauremys caspica) are elements of temperate fauna in North Africa. Some species of the great tropical fauna lying south of the Sahara Desert occur in North Africa and in Southwest Asia. Examples are the sand snakes (Psammophis), cobras, and chameleons (family Chamaeleontidae). As is true of the temperate fauna of Eurasia, the North African reptiles, though representing many families, are principally terrestrial and burrowing. Many lacertid and agamid lizards scamper over rocks and sand by day; they are replaced at night by small geckos and are preyed upon by the racers (*Coluber*) and sand snakes (*Psammophis*). In addition to cobras, the venomous snakes of North Africa include the common vipers, the saw-scaled viper (Echis carinatus), and the horned vipers (Cerastes). The last two are true desert animals. Land tortoises (Testudo) are common in the semi-arid land.

Reptiles of tropical Africa

The second and much larger division of the African fauna is the great tropical assemblage that ranges from the Sahara southward to the Cape of Good Hope. In common with tropical Asia, this vast area has cobras, many skinks, and many geckos. Its fauna differs from that of Asia in the absence of pit vipers (subfamily Crotalinae), the near absence of emydid turtles, and the poor representation of agamid lizards. These groups are replaced in tropical Africa by the many true vipers (subfamily Viperinae), the side-necked turtles (family Pelomedusidae), and the lacertid and cordylid lizards. Chameleons and land tortoises are abundant. Three species of crocodiles occur in Africa.

In Africa are found all of the diverse reptilian types characteristic of a tropical area: aquatic turtles, crocodiles, and snakes; terrestrial turtles, snakes, and lizards; burrowing snakes of the blunt-headed and auger types; limbless and virtually blind burrowing lizards; and a profusion of arboreal snakes and lizards.

Unique reptile fauna of Madagascar

The large island of Madagascar, off the eastern coast of Africa, has a peculiar fauna that appears in part as a collection of castoffs, groups that in Africa have not been able to meet the competition of more advanced forms. With few exceptions the reptiles of Madagascar belong to genera found only there.

NATURAL HISTORY

Reproduction and life cycle. Courtship. Courtship in some form is such a widespread prelude to mating among modern reptiles that it must have characterized many extinct groups as well. When courting, the male of some freshwater turtles, such as the red-eared turtle (*Chrysemys scripta* elegans) of the eastern United States, orients himself in the water so that he is directly in front of a female and facing her. With his forefeet close together, the male vibrates his claws against her head. If the female is receptive, she swims forward slowly while the male backs away. Finally the female sinks slowly down. The male then mounts her from behind, clutching her shell with all four feet. His tail is brought under hers and his penis introduced into her cloaca.

The male of some terrestrial turtles of North America (*e.g.*, the gopher tortoises, Gopherus) begins courtship by extending his neck and bobbing his head up and down. The courted female may bob her head in return. The male advances, nips at the female, and then circles her as she turns away from him. As soon as she shows signs of response, the male mounts her from the rear and begins a series of pumping movements that hump the rear of his shell against the ground. Finally the female extends her tail, and copulation begins. Males of the smaller box turtles (*Terrapene*) nip and butt at the female.

Mating behaviour of crocodilians

Male crocodilians bellow during the mating season, and in many cases the females respond with an answering call. The male American alligator, which copulates in water, grasps the female's neck with his jaws and slips the rear part of his body under hers to enable him to insert his penis into her cloaca.

Lizards have rather elaborate courtship patterns usually involving display and posturing by the males, who often have distinctive patches of colour on their throats or low on their sides. The male bobs up and down, thus exposing the patches of colour, which may be blue, orange, red, or black depending on the species. Males of some species, such as the green anole (Anolis carolinensis) of the southeastern United States, have brightly coloured folds of skin at the throat (dewlaps) that are expanded during courtship. If the female seems receptive, the male straddles her back, often gripping her back or legs with his jaws. Just before copulation, his tail is bent under hers.

The courtship patterns of snakes are simpler; they usually consist of the male's crawling over the back of the female and adopting every curve her body takes, then vibrating against her body or nudging it with waves of his own. Male boas and pythons stroke or scratch the female's body with their vestigial hind limbs. The male water snake rubs his chin against the female's back. The male rattlesnake (Crotalus) frequently nudges the female with his head. The male bull snake (*Pituophis*) grasps the female's neck with his jaws during copulation.

When the male king cobra (*Ophiophagus* hannah) crawls onto the back of a receptive female, he flicks his tongue against her repeatedly. The female raises her head and spreads her hood. The male nudges her neck and head with his snout and lifts the rear of her body with his tail. Copulation between cobras may last more than two hours.

The amnion

The embryo. The embryo of a reptile develops a thin sac, the amnion, that envelops the embryo and becomes filled with a watery fluid. The entire structure serves to protect the embryo from desiccation and mechanical injury. A parchment-like shell, which also contributes mechanical protection, is produced by the female parent and surrounds the amnion; between the shell and the amnion a second sac, the allantois, becomes inserted. The allantois, which is supplied with many fine blood vessels, serves as a respiratory organ, absorbing the oxygen and emitting the carbon dioxide that pass through the somewhat porous shell.

Egg laying. The typical mode of reptilian reproduction is oviparous (*i.e.*, the female lays eggs in which the young develop). The eggs are laid shortly after fertilization, and development of the embryos takes place largely after the eggs have been laid. This pattern characterizes crocodilians, turtles, the tuatara, most lizards and snakes, and many extinct reptiles. The size of eggs laid by lizards and snakes varies according to the size of the females. The banded rock lizard (Petrosaurus mearnsi) of the western United States, which ranges from 7.5 to ten centimetres (three to four inches) in length, lays eggs that are about one centimetre (0.4 inch) long; those of the 30- to 60-centimetre ringneck snake (Diadophis *punctatus*) from the eastern United States are 1.25 centimetres long. Eggs of the three-metre Komodo dragon lizard (Varanus komodoensis) and of the six-metre Indian python (Python *molurus*) are about 11.25 centimetres long.

A minority of modern and extinct reptiles are (and were) live-bearing, or viviparous. Strictly speaking, most of this minority are not truly viviparous but ovoviviparous, because the embryos develop with their shells or shell membranes intact and are nourished wholly by the yolk. In a few modern reptiles the embryonic membranes and the tissues lining the oviducts of the females come into close contact and are modified in one of several ways to provide a temporary organ, through which food and respiratory gases are exchanged; *i.e.*, a structure similar to the placenta of mammals. In the simplest reptilian "placenta," the most superficial layer (the epithelium) of the outer embryonic membrane and of the lining of the oviduct partially degenerates, thereby bringing the blood vessels of embryo and mother closer together. The approximation of the two bloodstreams facilitates the exchange of oxygen and carbon dioxide, this gas exchange being the only function of the organ at this stage of evolution. Several Australian snakes (Denisonia superba and D. *suta*) and a number of lizards—*e.g.*, the common East Indian brown-sided skink (Mabuya multifasciata) and the cylindrical skink (Chalcides ocellatus) of southern Europe and North Africa—are known to have this type of organ, as presumably do many ovoviviparous reptiles.

The best developed reptilian "placentas" consist of apposed, thickened, folded elliptical areas of the outer embryonic membrane and lining of the oviduct. The ridges of the oviductal areas are filled with blood vessels, and the epithelium between ridges is thickened and glandular. Usually, eggs developing with this type of "placenta" have less yolk; food and oxygen are transmitted from mother to embryo. Several species of Australian lizards, American water snakes, and the common European viper (Vipera berus) are known to provide this type of internal environment for their developing young.

Oviparity and ovoviviparity in lizards and snakes

The line between oviparity (egg laying) and ovoviviparity (hatching of eggs in the mother's body) is arbitrary. Females of some lizards and snakes retain the fertilized eggs in their bodies for a few days before laying them. Other species retain the eggs for most of the developmental period, hatching occurring shortly after laying. For the grass snake of England (Natrix natrix), the lapse of time between copulation and egg laying is usually two months; the young hatch six to ten weeks later. The interval between mating and egg laying is one month in the Texas horned lizard (Phrynosoma cornutum). A given species may be ovoviviparous in parts of its range and oviparous elsewhere.

The nest. The eggs of modern reptiles may be deposited in a nest prepared by the female or simply laid under some convenient cover, such as a rock or log. Crocodilians invariably prepare a nest, and the female invariably does the work. Most turtles dig their nests, scooping out a flask-shaped cavity in the ground with their hindfeet. When the hole has reached the proper size, the oval or spherical eggs, 1.5 to 3.75 centimetres (0.5 to 1.5 inches) in diameter depending on the species, are dropped into the nest from the female's cloaca. The female scratches soil over the eggs, usually obliterating the nest site. Crocodilians either dig a nest along the bank of a river or lake or heap together a mass of dead vegetation in which the eggs are laid; their oval eggs are usually about five centimetres (two inches) long.

Most oviparous lizards merely hide their eggs under some convenient cover such as under a rock or in a hole in a tree. Nest construction among lizards, though appearing in such diverse families as the iguanids, skinks, and true chameleons, is neither so elaborate nor so rigid in pattern as among turtles. The nest consists of a small hole made by either the snout or the limbs. Soil or leaves usually are pushed on top of the eggs to hide the nest, although the entrance to the nest cavity is kept open in a few species. Snakes, like lizards, usually lay their eggs under natural, pre-existing cover. The king cobra, one of the very few nest-building snakes, drags dead vegetation into a low heap by bending its body. The eggs are laid in a cavity at the centre. Other snakes deposit their eggs in holes they have scooped out of sand or soft earth with their snouts.

Number of offspring. The number of eggs in a clutch or offspring in a brood varies from one to 200 among living reptiles; presumably similar variation occurred among the extinct types. Crocodilians lay from 20 to 70 eggs, turtles from one to 200. In turtles, more so than in crocodiles, the number varies with the species and roughly with the size attained by the females. The big marine turtles have the largest clutches (usually more than 100); the smaller land and freshwater turtles have much smaller ones. The number of eggs or young is not so closely related to the size of mature females in species of lizards and snakes. With few exceptions, lizards of the family Gekkonidae lay two eggs at a time, regardless of the size of the female. Lizards of the family Scincidae have broods varying from two to about 30; one of the largest members of this family, the 30-centimetre-long (12-inch) stump-tailed skink (Tiliqua rugosa) of Australia, has only two young at a time, whereas the Great Plains skink (*Eumeces* obsoletus) of the United States usually lays between ten and 20 eggs in a clutch.

Relationship of brood size to adult size

Clutch size in the 1.5-metre (five-foot) bull snake (*Pi*tuophis catenifer) of western United States is usually ten or 12; in the grass snake of England and Europe, which measures 60 to 90 centimetres (two to three feet), it is 30 to 40; in the giant reticulated python of Southeast Asia and the East Indies, it may reach 100.

Parerttal care. Parental care of eggs and newborn young is neither well developed nor elaborated among reptiles. Female crocodilians generally remain in the vicinity of their nests and chase would-be predators from the site. In a few lizards the female returns to the nest between feeding excursions to coil around the eggs and turn them at intervals. The male and female king cobra remain in the vicinity of the nest, and one of the parents usually is coiled above the egg cavity.

Female pythons coil around their eggs and pull them into a heap. Females of some species remain with the eggs for the entire two-month incubation period; others leave the eggs only to drink. In at least one species (Python *molurus*) the female provides heat by muscular contraction to keep the eggs at incubation temperature on cool nights. Turtles and the majority of egg-laying lizards andsnakes abandon their eggs after they are laid.

Incubation period. The incubation (or gestation) period of reptilian eggs is affected by many factors and to such an extent that it is difficult to assign a figure characteristic of a given species. One source of complication is the combination of oviparous and ovoviviparous habits by certain species. The developmental period of the embryo, whether it occurs within the female's body or outside it, is referred to as the gestation period.

In general, the gestation period lasts from 60 to 105 days in most American and European reptiles. The eggs of the American alligator hatch about 63 days after they are laid, those of the small Eastern fence lizards (*Sceloporus* undulatus) in about the same time. The gestation period in the common European viper lasts from 60 to 90 days. Eggs of marine turtles hatch between 30 and 75 days after they are laid, depending on the site. The temperatures to which a brood is subjected shorten or lengthen gestation according to whether the temperatures are high or low.

Growth and longevity. Giant Galápagos tortoises kept under nearly ideal conditions have been known to increase their weight from 3.2 to six kilograms (seven to 13 pounds) to about 82 kilograms (180 pounds) in nine years. Smaller species also grow rapidly. The box turtle of the United States has a shell about 3.75 centirnetres long at the end of its first year; at the end of five years, the length has doubled.

Under favourable conditions a one-year-old American alligator is about 60 centimetres long and weighs about 1.8 kilograms (four pounds). At the end of six years, males average about 190 centimetres (about six feet) and about 36 kilograms (80 pounds). The red diamond rattlesnake is about 30 centimetres long at birth, grows to about *65* centimetres in its first year, reaches about 85 centimetres by the end of its second year, and grows more slowly after that. The pattern for lizards is much the same: rapid growth early in life and slow growth afterward. The significant difference between growth in reptiles and that in mammals is that a reptile has the potential of growing throughout its life, whereas a mammal reaches a terminal size and grows no more, even though it may subsequently live many years in ideal conditions.

The length of time needed to attain sexual maturity varies greatly among reptiles and, although roughly related to the size usually attained by the species, is even more closely related to the climate in which the animal lives. The red diamond rattlesnakes in southern California, for example, bear their first young when three years old; on the other hand, the northern Pacific rattlesnake (Crotalus viridis oreganus) bears its first litter when four years old. The much smaller common garter snake is sexually mature shortly before the age of two years.

Achievement of sexual maturity

The 12.5- to 15-centimetre (five- to six-inch) northwestern sagebrush lizard (Sceloporus graciosus gracilis), living in the Sierra Nevada range of California at an elevation (about 1,800 metres [6,000 feet]) where it has, at most, six months of activity each year, requires two years or more to reach sexual maturity. Another lizard, the green anole (Anolis carolinensis), similar in size to the sagebrush lizard but living in the lowlands of the southern United States, may reach maturity in four or five months in Florida.

Turtles mature at a slower rate. Females of the red-eared turtle of the central United States lay their first eggs when they are from three to eight years old, depending upon how long it takes them to reach a shell length of 15 centimetres. Females of the musk turtle, or stinkpot (Sternotherus odoratus), in Michigan require nine to 11 years to mature, at which time their shells are 7.5 to ten centimetres long. Presumably, turtles living in the tropics mature more rapidly.

The maximum age, meaning the potential longevity, of modern reptiles varies greatly and can be determined only from records of captive animals. Turtles as a group seem capable of Jiving longer than the others, and about 30 species have been kept in captivity more than 20 years. Several species, said to have lived 150 years or more, may be cases of two individuals whose periods of captivity overlapped. There is no reliable evidence for believing that the giant land tortoises live much longer than some smaller species. Two crocodilians (Alligator *mississippiensis* and A. sinensis) have survived in zoos for more than 50 years. Several species of pythons and boas have lived longer than 20 years. Lizards seem to have an upper limit near that of snakes. A slowworm, Anguis *fragilis,* has been kept in captivity for more than 30 years.

BEHAVIOUR

Defense. Avoidance and noise. Avoidance, the commonest form of defense in the animal kingdom, is also the commonest one in reptiles. At the first recognition of danger, most snakes and lizards crawl or scamper away into the undergrowth; turtles and crocodilians plunge into water and sink out of sight. But should the danger arise so suddenly and so close at hand that flight may be hazardous, other expedients are adopted.

Crocodiles, some lizards, turtles, and some snakes hiss loudly when confronted by an enemy. Rattlesnakes rapidly vibrate the tip of the tail, which consists of loose, dry, horny rings. A few snakes without rattles (*e.g.,* the fox snake, Elaphe vulpina, of the United States) vibrate the ends of their tails rapidly, and if, as often happens, the tail hits dry leaves, it makes a sound deceptively like the rattle of a rattlesnake.

Hissing and tail rattling

Body form and posturing. Change in body form, which is relatively common in snakes, usually involves spreading the neck, as in cobras (family Elapidae), or the whole body, as in the harmless hognose snakes (*Heterodon*) and DeKay's snake (Storeria dekayi) of the United States. Some snakes inflate the forward parts of their bodies; inflation is one of the defensive actions of the large South American tree snake Spilotes and of the African boomslang (Dispholidus).

Threatening postures may be assumed by snakes as they change their body form. A cobra raises the forepart of its body and spreads its hood when endangered. The typical defensive posture of a viper is with the body coiled and the neck held in an S-curve, the head poised to strike.

Some lizards flatten their bodies, puff out their throats, and turn broadside to the enemy. The helmeted iguanids (Corythophanes) of Central America and the chameleons of Africa increase their apparent size in this way when approached by snakes. The Australian bearded lizard (Amphibolurus barbatus) spreads its throat downward and outward. The Australian frilled lizard (*Chlamydosaurus* kingi) suddenly raises a wide membrane, or frill, which extends backward from the throat. Many lizards and snakes open their mouths when threatened, but do not strike. A common African lizard, Agama atricollis, faces an enemy with head held high and mouth open to show the brilliant orange interior.

Display of colour. Display of colour in Agama *atricollis* may not be part of a threatening mechanism, but it is so in the instances of certain red- or yellow-bellied snakes that turn over or curl up their tails, exposing the brightly coloured undersurface. This behaviour, known in harmless (*e.g.*, the American ring-necked snake, *Diadophis*) as well as venomous snakes (*e.g.*, the coral snake, Micrurus frontalis), is displayed only by snakes having red, orange, or yellow undersides. These colours must have some significance, as yet not fully understood, to predacious animals, for they are also the common colours in insects having warning coloration.

The defense mechanism of camouflage involving form and colour is common. Many arboreal snakes and lizards (*e.g.*, chameleons) are green; some of the green snakes (*e.g.*, the vine snakes of South America, Oxybelis, and of southern Asia, Ahaetulla) are very slender, resembling plants common in the habitat. Lizards of semi-arid and rocky country frequently are pale in colour and blotched in pebble fashion—*e.g.*, the leopard lizard (Crotaphytus *wislizeni*) of the southwestern United States.

Mimicry

Mimicry of dangerous species by harmless snakes is a passive defense. Its validity as an actual mechanism of defense is, however, sometimes challenged. The venomous coral snakes (Micrurus) of the Western Hemisphere are ringed with bright red, yellow, and black. A series of relatively harmless snakes, such as *Erythrolamprus* and *Anilius* of South America and the scarlet king snake, Lampropeltis *triangulum* doliata, of southeastern United States, have similar colours and patterns that may confer some protection against predators.

Striking and biting. If a threatening posture does not succeed in driving off an enemy, many reptiles become more aggressive. Some snakes (*e.g.*, DeKay's snake) strike, but with their mouths closed. Others (*e.g.*, the hognose snakes) strike with their mouth open but do not bite. Still others strike and bite viciously. Among the nonvenomous snakes of North America, few are as quick to bite as the water snakes (Natrix). The sole danger from the bites of these snakes is infection of the wound.

Most of the dangerously venomous snakes (vipers, pit vipers, and cobras) bite in self-defense. Vipers and pit vipers usually strike from a horizontally coiled posture. From this position the head can be shot forward, stab the enemy, and be as rapidly pulled back in readiness for the next strike. From the typical raised posture a cobra sweeps its head forward and downward to bite. To strike again it raises its head and neck once more; such aggressive, defensive movements of cobras are slower than those of pit vipers.

Many lizards, regardless of family and size, bite in defense. Gekko gecko of Southeast Asia bites if sufficiently threatened. Although small lizards have a bite effective against only the smallest predators, a large monitor lizard (Varanus) can inflict a painful wound with its large teeth and strong jaws. Some turtles, particularly the soft-shelled turtles (Trionyx), bite frequently, vigorously, and effectively.

Spitting. The spitting of venom by certain African cobras, the ringhals (Hemachatus haemachatus), and the black-necked cobra (*Naja nigricollis*) is a purely defensive act directed against large enemies. A fine stream of venom is forced out of each fang, which, instead of having a straight canal ending in a long opening near the tip as in most cobras, has a canal that turns sharply forward to a small round opening on the front surface well away from the tip. At the moment of ejection the mouth is opened slightly, and venom is forced out of the fangs by contraction of the muscle enveloping the poison gland. Usually a spitting cobra raises its head and the forepart of its body in the characteristic cobra defensive posture prior to spitting, but venom can be ejected from any position. The effect on skin is negligible; the eyes, however, may be severely damaged, and blindness can result unless the venom is washed out quickly.

Use of the tail. A few lizards, representing different families, have in common thick tails covered by large, hard, spiny scales. Such a tail swung vigorously from side to side is an effective defense against snakes, especially when the head and body of the lizard are in a burrow or wedged between rocks.

Voluntary tail shedding

Lizards' tails are useful in defense in another way. When captured, many lizards voluntarily shed their tails, which wriggle violently, temporarily confusing the predator and allowing the lizard to escape. Each vertebra of the tails of lizards with this capacity has a fracture line and can be split on that line when tail muscles contract violently. Simultaneous stimulation of the nerves in the severed portion keeps it twitching for a few seconds after separation. Usually the tail is broken in only one place, but a few lizards, particularly the so-called glass snakes (*Ophisaurus*), break their tails into several pieces. The stump heals quickly, and a new tail grows; often, however, the regenerated tail is not so long as the original and has simpler scales.

Snakes, turtles, and crocodiles may have their tails bitten off by predators, but they cannot break them voluntarily or regenerate them. Some snakes use their tails in diversionary tactics by raising them and moving them slowly. Species with this habit commonly have thick, blunt, brightly coloured tails. The small African python Calabaria and the Oriental venomous snake Maticora wave their tails in the air as they move slowly away from a threat.

Balling. Many snakes, both harmless and venomous, attempt to hide their heads under coils of their bodies. The body may be coiled loosely, as it is in most species with this habit, or tightly so that it forms a compact ball with the head in the centre. Balling, as the latter habit is called, is a characteristic response of Calabaria and another African python, Python regius.

The African armadillo lizard (*Cordylus* cataphractus), a species with heavy scales on its head and hard spiny scales covering its body and tail, rolls on its back and grasps its tail in its mouth. It thus presents a ring of hard spines to a predator.

Odours. Some reptiles use musk-secreting glands when other defensive measures fail. The water snakes (Natrix), the garter snakes (*Thamnophis*), the alligator lizards (Gerrhonotus), and the musk turtles (Sternotherus) emit a foul-smelling substance from anal glands.

Feeding habits. With few exceptions, modern reptiles feed on some form of animal life: insects, mollusks, birds, frogs, mammals, fishes, or other reptiles. Land tortoises are vegetarians, eating leaves, grass, and in some cases even cactus. The big green iguana (Iguana iguana) of Central and South America, its relative the chuckwalla (Sauromalus obesus) of southwestern United States and northern Mexico, and the spiny-tailed agamids (*Uromastix*) of North Africa and southwestern Asia also are herbivorous. The marine iguana (*Amblyrhynchus cristatus*) of the Galapagos Islands dives into the sea for seaweed.

The majority of carnivorous reptiles have nonspecial-

Man-eating reptiles

ized diets, feeding on a variety of animals. In general, the smaller the reptile, the smaller is its prey. Only the very largest of living snakes—the reticulated python (*Python* reticulatus), the Indian python (P. *molurus*), and the anaconda (*Eunectes murinus*)—are capable of eating large mammals such as small pigs and deer. Among crocodiles the largest species—the Nile crocodile (*Crocodylus* niloticus), the East Indian saltwater crocodile (C. *porosus*), and the Orinoco crocodile (C. *intermedius*)—have been known to attack and to eat men. Presumably, even larger prey was devoured by the great carnivorous dinosaurs such as *Allosaurus* and *Tyrannosaurus*, which were almost certainly capable of killing the largest of their herbivorous contemporaries.

Locomotion. Walking and crawling. The majority of reptilian orders are quadrupedal—*i.e.*, four-legged. Among the land vertebrates, the limbs gradually shifted from a lateral to a ventral position. In most amphibious reptiles the limbs projected out to the side and then bent downward to the ground at the knee and elbow. With few exceptions, the quadrupedal reptiles have the same awkward position. With such an orientation, the centre of gravity of the body is not in the same axis as the hands and feet, resulting in a sideways as well as a forward component of thrust when the animal walks. The typical reptile throws its body into a slight horizontal curve to progress straight forward. In mammals the limbs are directly underneath the body, the centre of gravity is in the axis of the limbs, and all of the thrust of the limbs is directed forward. The latter position and type of motion are more efficient. The lateral orientation of the limbs in amphibians and reptiles also makes it more difficult to raise the body off the ground.

Despite the awkwardness of the orientation of their limbs, some reptiles are (and many extinct forms probably were) capable of moderate speeds. Crocodilians raise their bodies off the ground and make short, fast rushes. Short-bodied lizards also can move fast for short distances; longer bodied lizards have greater difficulty in raising their bodies. They usually have short legs and proceed in a serpentine fashion, with the body, thrown in horizontal curves, doing much of the work.

A snake moves by pushing backward against rocks, sticks, or any relatively fixed point—a lump of earth or a small depression in uneven ground—with the rear surface of the horizontal curves of its body. Each joint of the body passes through the same curves, pressing against the same object and thrusting the forepart of the body forward. Heavy-bodied snakes such as pythons and certain rattlesnakes can move forward without throwing their bodies into curves. This rectilinear movement depends on the ability of snakes to stretch or contract their bodies in the longitudinal axis. By raising a part of its belly, stretching that part forward, lowering it to the ground, and repeating the process alternately with other parts of the body, a heavy snake moves forward smoothly in a straight line.

Bipedal locomotion

Some modern lizards have adopted semi-bipedal locomotion. The collared lizard (Crotaphytus collaris) of the United States and the frilled lizard (*Chlamydosaurus kingi*) of Australia show the early stages of bipedalism, a phenomenon widespread among the dinosaurs and therefore important in reptilian history. These lizards run on their long hindlegs with the forward parts of their bodies at an angle of about 60° off the horizontal.

Presumably, bipedalism among the dinosaurs began as it did among modern lizards, as an occasional means of obtaining bursts of speed. Because the centre of gravity is in front of the hips, modern bipedal lizards must move forward continuously in order to maintain a semi-erect posture; they can stand still in that position only for very short periods.

The awkward sideways orientation of the limbs forces bipedal lizards to swing each leg outward as it is brought forward and to push the body sideways and forward when each leg thrusts backward against the ground. Bipedal dinosaurs eliminated this inefficient rocking motion, for during the course of evolution their hind limbs were rotated forward so that they were directly under their bodies. Thus, they delivered their full force in the forward direction. So successful was this mode of locomotion that dinosaurs utilizing it dominated terrestrial life for millions of years.

Clinging and climbing. Associated with arboreal life are groups of anatomical features mainly concerned with clinging. The commonest clinging structures in vertebrates are claws; they seem to be the only arboreal adaptations of some lizards, such as the common iguana (*Iguana* iguana). Similar structures appear in many lizards of the family Gekkonidae, in the anoles (Anolis) of the family Iguanidae, and in some skinks of the family Scincidae.

Pads on the feet consist of wide plates or scales under the fingers and toes. The outer layer of each plate or scale is composed of innumerable tiny hooks formed by the free, bent tips of cells. These minute hooks catch in the slightest irregularities and enable geckos to run up apparently smooth walls and even upside down on plaster ceilings. Because the hooklike cells are bent downward and to the rear, a gecko curls its toes upward to disengage them. Thus, when walking or running up a tree or wall, a gecko must curl and uncurl its toes at every step.

Prehensile tails

The giant Solomon Islands skink (Corucia), true chameleons (Chamaeleontidae), arboreal vipers, boas, and pythons use prehensile tails—that is, tails capable of supporting most of the weight of the animal or used habitually for grasping—for clinging to their aerial supports. For this purpose, however, true chameleons rely mainly on a tonglike arrangement of their digits, which are ucited into two opposed bundles on each foot—three on the inside and two on the outside of the front foot, and two on the inside and three on the outside of the hindfoot.

Slender vine snakes of several genera of the family Colubridae are capable of extending half the body length in a horizontal plane without support; they do so habitually in bridging the gap between branches. Most snakes can reach across an open space, but all except the vine snakes can extend only a short length of the body, and that portion invariably sags like a cable. The vine snakes bridge an open space like an I-beam. This ability is based partly on reduced body weight and partly on deepened and strengthened vertebrae.

Swimming. In water, of course, neither bipedal nor quadrupedal locomotion is very effective. Aquatic reptiles, with few exceptions, use the same means of propulsion as do fish and whales—that is, powerful beats of the tail. Crocodilians and aquatic lizards such as some monitors (Varanidae) lash their tails from side to side while holding the limbs against the body. The same method was used by the ancient mesosaurs (Mesosauria) and ichthyosaurs (Ichthyosauria). The marine ichthyosaurs, which were the reptilian counterpart of the porpoises, may have used their very short limbs for steering.

A fishlike method of swimming requires a flexible body and at least a moderately long tail. Turtles propel themselves by using their feet as paddles—the hindfeet, which have webbed toes, in the case of freshwater turtles, and the forefeet, which are modified into large paddles, in the case of marine turtles.

The extinct marine plesiosaurs (suborder Plesiosauria), with their short bodies and tails and their large paddle-like limbs, swam the way marine turtles do, although they may have used their hindlimbs for more than just steering. Both pelvic and pectoral (shoulder) girdles were modified in the plesiosaurs into structures having small upper portions and very large lower portions. As the upper element, especially that of the pelvic girdle, has the important function of transferring the weight of the body to the limbs, it is likely that the limbs of plesiosaurs could not support the body weight on land and that the plesiosaurs never came out of water.

Most plesiosaurs had long necks. By moving toward their prey with the neck curved, they probably could strike suddenly. The heavy trunk would provide the inertia against which the neck could move, thus preventing a significant backward shift of the animal as the head shot forward. The modern sea snakes (Hydrophidae) show the same adaptation. Though they swim with an eel-like

Locomotion of sea snakes

undulation of the body, the sea snakes have relatively small heads, slender necks, and very heavy middle and rear sections. With most of the body mass concentrated in the second half of the animal, almost all of the force of the strike is used to drive the head forward.

Flying. Three groups of reptiles have experimented with flight. Thecodontia, a group of Archosauria (the so-called ruling reptiles, which included dinosaurs and crocodilians), became highly successful at this means of locomotion and evolved into birds.

A second group of archosaurs, the Pterosauria, developed wings that were supported along the front margin by the arm and an extremely elongated finger. The pterosaur wing was made of skin; since it lacked both internal supports and feathers, it probably lacked the flexibility or durability of a bird wing. Flight of the pterosaurs presumably amounted to soaring and gliding. It is not understood how they moved when not flying and how they managed to take off if they happened to land on level ground. Since most remains have been found in marine deposits, it is assumed that they lived along ocean shores, probably roosting on cliffs from which takeoff would have been easy.

"Flying" lizards

The third experiment with flight was made by a group of modern lizards *(Draco)*. The "wing" of these small lizards consists of skin supported by five or six elongated ribs between the arm and leg. At rest the ribs and the wings are folded against the sides of the body. In flight the wings form broad semicircles from arm to leg on each side. These flying lizards, which live in the forested country of Southeast Asia and the East Indies, are capable only of gliding. A flying lizard launches itself from a tree into the air and glides toward another tree, turning upward sharply just before lighting on the new perch. Since the arms and legs are not modified, this lizard is capable of scampering about like any strictly arboreal lizard.

FORM AND FUNCTION

External covering. The external covering of reptiles is characteristically dry. It bears few glands or none at all and differs in this respect from the skin of amphibians and mammals. The so-called malpighian layer of the epide mis secretes the outer layer, which is tough and horny. Bony plates develop in the dermis, which lies just below the epidermis. The arrangement of scales is usually characteristic for each species.

Internal features. *Skeletal system and dentition.* The skeleton of reptiles fits the general pattern of vertebrates. They have a bony skull, a long vertebral column that encloses the spinal nerve cord, ribs that form a bony basket around the viscera, and a framework of limbs.

Each group of reptiles developed its own particular variations on this major pattern in accord with the general adaptive trends of the group. Snakes, for example, have lost the limb bones, although a few retain vestiges of the hindlimb. The limbs of several types of marine reptiles became modified into fins or flippers with obvious functional significance. In groups such as the extinct ichthyosaurs and plesiosaurs, the bones of the limbs, no longer supporting the weight of the body against the pull of gravity, became much shortened. At the same time the bones that in other reptiles composed the digits multiplied in number, forming a long flipper.

Groups of reptiles whose modes of life came to depend heavily on passive defense also developed specializations of the skeleton. The bony and horny shell of turtles and rows of bony plates on the back of ankylosaurs (Cretaceous dinosaurs) are cases in point.

The skulls of the several subclasses and orders vary in the ways mentioned below. In addition to differences in openings on the side of the skull and in general shape and size, the most significant variations in reptilian skulls are those affecting movements within the skull.

Reptilian skulls as a group differ from those of early amphibians, the vertebrates from which reptiles arose, in lacking on otic notch (an indentation at the rear of the skull) and several small bones at the rear of the skull roof. The skulls of modern reptiles are sharply set off from those of mammals in many ways, but the clearest differences are in the lower jaw and adjacent regions. Reptiles have a number of bones in the lower jaw, only one of which, the dentary, bears teeth. Behind the dentary a small bone, the articular, forms a joint with the quadrate bone near the rear of the skull. In mammals the lower jaw consists of a single bone, the dentary, and the articular and quadrate have become part of the chain of little bones in the middle ear. An almost complete transition between these two very different arrangements is known from fossils of mammal-like reptiles (order Therapsida).

Differences between reptile and mammal skulls

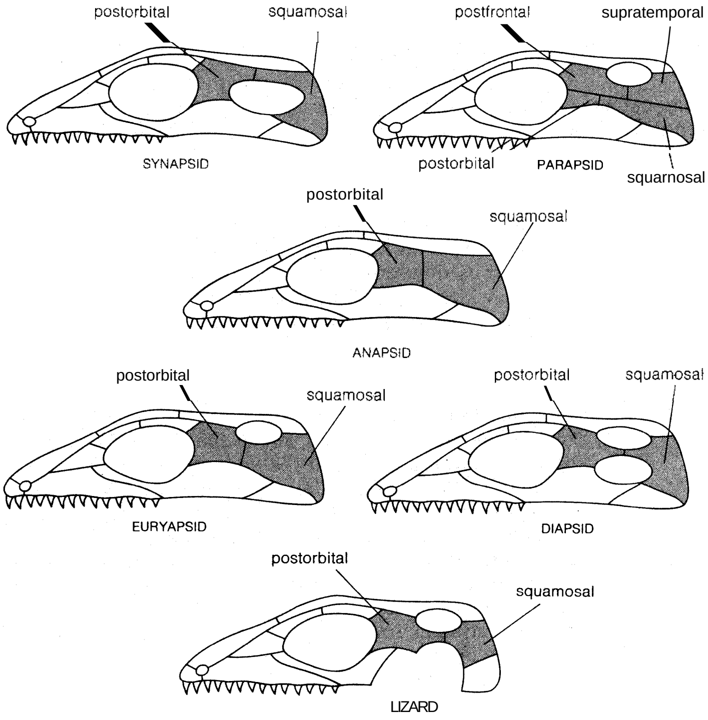


Figure 2: Diagrammatic reptilian skulls.

The dentition of most reptiles shows little specialization along the tooth row. **A** division into distinctive bladelike incisors, tusklike canines, and flat-crowned molars, such as characterize mammals, does not occur in reptiles. Instead, the entire tooth row usually consists of long conical teeth. Venomous snakes have one or several hollow or grooved fangs, but they have the same shape as most snake teeth. The principal differences between species lie in the number, length, and position of the teeth. Crocodilians among the living forms and dinosaurs among the extinct forms have but a single upper and a single lower tooth row. Snakes and many extinct reptilian groups have teeth on the palatal bones (vomer, palatine, pterygoid) and on the bones of the upper jaw (premaxilla, maxilla); only one row of teeth is present on the lower jaw.

Lizards have conical or bladelike bicuspid or tricuspid teeth. Some species have conical teeth at the front of the jaws and cuspid teeth toward the rear, but the latter are not comparable to the molars of mammals in either form (they are not flat-crowned) or function (they do not grind food). Turtles, except for the earliest extinct species, lack teeth, having instead upper and lower horny plates that serve to bite off chunks of food.

The teeth of reptiles are also less specialized in function than are mammalian teeth. The larger carnivorous reptiles (*e.g.*, crocodilians and certain dinosaurs) are equipped only to tear off or bite off large pieces of their prey and to bolt them without chewing. Insectivorous lizards (the majority of lizards) usually crack the exoskeleton of their insect prey, and then swallow the prey without grinding it up. Snakes simply swallow their prey whole without any mechanical reduction.

Skull and joint structures. Many reptiles developed joints (in addition to the hinge for the lower jaw) within the skull, permitting at least slight movement of one part relative to others. The capacity for such movement with-

Kinesis

in the skull, called kinesis, enables an animal to increase the gape of the mouth and thus is an adaptation for swallowing large objects. Apparently some of the large carnivorous theropod dinosaurs (*e.g., Allosaurus*) had a joint between the frontal and parietal bones in the roof of the skull. All reptiles of the subclass Lepidosauria (lizards, snakes, rhynchocephalians, and the extinct eosuchians) have had kinetic skulls, but they differ from the dinosaurs in having the joint on the floor of the skull at the juncture of basisphenoid and pterygoid bones.

The skulls of the lepidosaurians became increasingly kinetic as new groups evolved. The Rhynchocephalia (which include the living tuatara) and their antecedents, the Eosuchia, had only the basisphenoidal–pterygoidal joint. The lizards lost the lower temporal bar, thereby freeing the quadrate bone and allowing greater movement to the lower jaw, which is hinged to the quadrate. Finally, in the snakes, this trend culminates in the most kinetic skull among the vertebrates—a skull having the ancestral basisphenoidal–pterygoidal joint, a highly mobile quadrate (which gives even greater mobility to the lower jaw), upper jaws capable of rotating on their longitudinal axes and of moving forward and backward, and often a hinge on the roof of the skull between the nasal and frontal bones that allows the snout to be raised slightly. In short, the only part of a snake's skull incapable of movement is the braincase.

Nervous system. As in all vertebrates, the nervous system of reptiles consists of a brain, a spinal nerve cord, nerves running from the brain or spinal cord, and sense organs. Reptiles have small brains compared with mammals. The most important difference between the brains of these two vertebrate groups lies in the size of the cerebral hemispheres, the principal associative centres of the brain. In mammals these hemispheres make up the bulk of the brain and, when viewed from above, almost hide the rest of the brain. In reptiles the relative and absolute size of the cerebral hemispheres is much smaller. The brain of snakes and alligators forms less than $\frac{1}{1,500}$ of the total body weight, whereas, in mammals such as squirrels and cats, the brain accounts for about $\frac{1}{100}$ of the body weight. A stegosaur (Stegosauria), roughly the size of an elephant, had a braincase no larger than that of a 2.4-metre crocodile, about large enough to contain a brain the size of a large walnut.

Circulatory system. Modern reptiles do not have the capacity for rapid sustained activity found in birds and mammals. It is generally accepted that this lower capacity is related to differences in the circulatory and respiratory systems. Before the origin of lungs, the vertebrate circulatory system had a single circuit: in the fishes, blood flows from heart to gills to body and back to the heart. The heart consists of four chambers arranged in a linear sequence.

Evolution of the reptile heart

With the evolution of lungs in amphibians, a new and apparently more efficient circulatory system evolved. Two chambers of the heart, the atrium (or auricle) and ventricle, became increasingly important, and the beginnings of a double circulation appeared. An early stage in this evolution can be seen in amphibians today, where one of the main arteries from the heart (the pulmonary artery) goes directly to the lungs, whereas the others (the systemic arteries) carry blood to the general body. The blood is aerated in the lungs and carried back to the atrium of the heart. From the left side of the atrium, which is at least partially divided for the first time, the aerated blood is pumped into the ventricle and there mixes with the nonaerated blood from the body that was returned to the heart via the right half of the atrium. Then the cycle begins again. One of the features of the amphibian system is that the blood leaving the heart for the body is only partially aerated; part of it is the deoxygenated blood returned from the body.

All groups of modern reptiles have a completely divided atrium; it is safe to assume, therefore, that this was true of most, if not all, extinct reptiles. In reptiles, the ventricle for the first time becomes at least partially divided in the four major living groups.

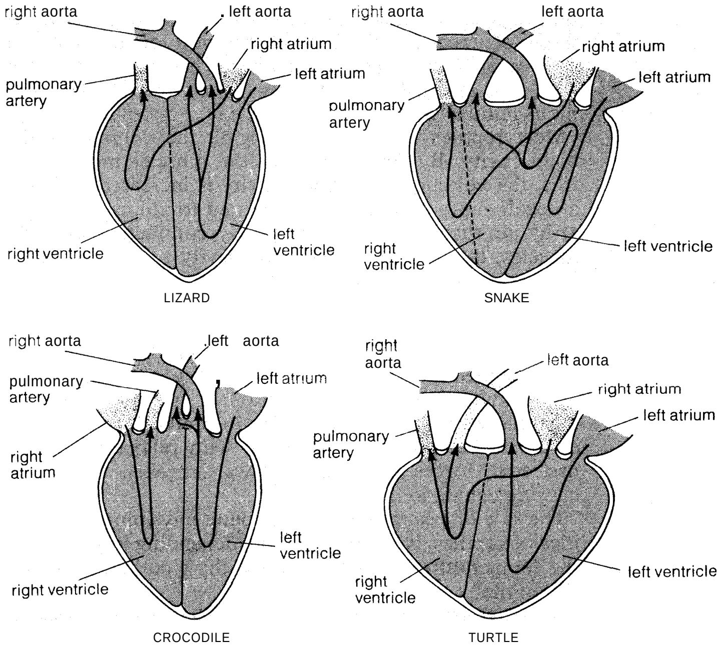


Figure 3: Types of reptilian hearts.

When the two atria of a lizard's heart contract, the two streams of blood (aerated blood from the lungs in the left atrium and nonaerated blood from the body in the right atrium) flow into the left chamber of the ventricle. As pressure builds up in that chamber, the nonaerated blood is forced through the gap in the partition into the right chamber of the ventricle. Then, when the ventricle contracts, nonaerated blood is pumped into the pulmonary artery and thence to the lungs, while aerated blood is pumped into the systemic arteries (the aortas) and so to the body.

In snakes all three arterial trunks come out of the chamber of the ventricle that receives the nonaerated blood of the right atrium. During ventricular contraction, a muscular ridge forms a partition that guides the nonaerated blood into the pulmonary artery, while the aerated blood received by the other chamber of the ventricle is forced through the opening in the ventricular septum and out through the aortas.

In crocodiles the ventricular septum is complete, but the two aortas come out of different ventricular chambers. A semilunar valve at the entrance to the left aorta prevents nonaerated blood in the right ventricle from flowing into the aorta. Instead, part of the aerated blood from the left ventricular chamber pumped into the right aorta flows into the left by way of an opening.

The ventricle of the turtle is not perfectly divided, and some slight mixing of aerated and nonaerated blood takes place.

Oxygen capacity of reptile blood

Despite the peculiar and complex circulation, a double system has been achieved by lizards, snakes, and crocodilians. Tests of the blood in the various chambers and arteries have shown that the oxygen content in both systemic aortas is as high as that of the blood just received by the left atrium from the lungs and is much higher than that of the blood in the pulmonary artery. Except for the turtles, limitation of activity in reptiles cannot be explained on the basis of imperfect heart circulation.

An explanation may lie in the chemistry of the blood. Apparently, the blood of reptiles has less hemoglobin and thus can carry less oxygen that that of mammals.

Respiratory system. The form of the lungs and the methods of irrigating them may also influence activity by affecting the efficiency of respiration. In snakes the lungs are simple saclike structures having small pockets, or alveoli, in the walls. In the lungs of many lizards and turtles and of all crocodilians the surface area is increased by the development of partitions that, in turn, have alveoli. Because exchange of respiratory gases takes place across surfaces, an increase of the ratio of surface area to volume leads to an increase in respiratory efficiency. In this regard, the lungs of snakes are not so effective as those of crocodilians. The elaboration of the internal surface of lungs in reptiles is simple, however, compared with that reached by mammalian lungs with their enormous number of very fine alveoli.

Most reptiles breathe by changing the volume of the body cavity. By contractions of the muscles moving the ribs, the volume of the body cavity is increased, creating a negative pressure, which is restored to atmospheric level by air rushing into the lungs. By contraction of body muscles, the volume of the body cavity is reduced, forcing air out of the lungs.

This system applies to all modern reptiles except turtles, which, because of the fusion of the ribs with the rigid shell, are unable to breathe by this means; they do use the same mechanical principle of changing pressure in the body cavity, however. Contraction of two flank muscles enlarges the body cavity, causing inspiration. Contraction of two other muscles, coincident with relaxation of the first two, forces the viscera upward against the lungs, causing exhalation.

The rate of respiration, like so many physiological activities of reptiles, is highly variable, depending in part upon the temperature and in part upon the emotional state of the animal.

Digestive and urogenital systems. The digestive system of modern reptiles is similar in general plan to that of all higher vertebrates. It includes the mouth and its salivary glands, esophagus, stomach, and intestine, ending in a cloaca. Of the few specializations of the reptilian digestive system, the evolution of one pair of salivary glands into poison glands in the venomous snakes is the most remarkable.

During development the embryos of higher vertebrates (reptiles, birds, and mammals) use three separate sets of kidneys consecutively; these are arranged in longitudinal sequence in the body cavity. The first set, the pronephroi, are vestigial organs left over from the evolutionary past that soon degenerate and disappear without having had any function. The second set, the mesonephroi, are the functional kidneys of adult amphibians, but their only contribution to the lives of reptiles is in providing the duct (the wolffian duct) that forms a connection between the testis and the cloaca. The operational kidneys of reptiles, birds, and mammals are the last set, the metanephroi, which have separate ducts to the cloaca. The principal function of the kidney is the removal of nitrogenous wastes resulting from the oxidation of proteins. Vertebrates eliminate three kinds of nitrogenous wastes: ammonia, urea, and uric acid. Ammonia and urea are highly soluble in water; uric acid is not. Ammonia is highly poisonous, urea slightly so, and uric acid not at all. Among reptiles the form taken by the nitrogenous wastes is closely related to the habits and habitat of the animal. Aquatic reptiles tend to excrete a large proportion of these wastes as ammonia in solution. This method, involving a great loss of body water, is no problem for an alligator, which eliminates between 40 and 75 percent of its nitrogenous wastes as ammonia. Terrestrial reptiles, such as most snakes and lizards, which must conserve body water, convert their nitrogenous wastes to insoluble, harmless uric acid, which forms a more or less solid mass in the cloaca. In snakes and lizards these wastes are eliminated from the cloaca together with wastes from the digestive system.

Discharge routes for eggs and sperm

Prior to the evolution of the metanephric kidney, the products of the male gonad, the testis, travelled through the same duct with the nitrogenous wastes from the kidney. But with the appearance of the metanephros, the two systems became separated. The female reproductive system never shared a common tube with the kidney. Oviducts in all female vertebrates arise as separate tubes with openings usually near, but not connected to, the ovaries. The oviducts, like the wolffian ducts of the testes, open to the cloaca. Both ovaries and testes lie in the body cavity near the kidneys.

With the evolution of the reptilian egg, internal fertilization became necessary. The males of all modern reptiles, with the exception of the tuatara, have copulatory organs. The structures vary from group to group, but all include erectile tissue as an important element of the operating mechanism, and all are protruded through the male's cloaca into that of the female during copulation. Unlike the penis of turtles and crocodilians, the copulatory organ of lizards and snakes is paired, each unit being called a hemipenis. The hemipenes of lizards and snakes are elongated tubular structures lying in the tail. The penis of a crocodile or turtle is protruded through the cloacal opening wholly by means of a filling of blood spaces (sinuses) in the penis; protrusion of a lizard's or snake's hemipenis, however, is begun by a pair of propulsor muscles. Completion of the erection is brought about by blood filling the sinuses in the erectile tissue. Only one hemipenis is inserted into a female, but which one is a matter of chance. Unlike the penis of mammals, the copulatory organs of reptiles do not transport sperm through a tube. The ducts from the testes, as already mentioned, empty into the cloaca, and the sperm flow along a groove on the surface of the penis or hemipenis.

Sense **organs.** *Sight.* In general construction the eyes of reptiles are like those of other vertebrates. Accommodation for near vision in all living reptiles except snakes is accomplished by pressure being exerted on the lens by the surrounding muscular ring (ciliary body), which thus makes the lens more spherical. In snakes the same end is achieved by the lens being brought forward under pressure built up on the vitreous humour by contractions of muscles at the base of the iris. The pupil shape varies remarkably among living reptiles, from the round opening characteristic of all turtles and many diurnal lizards and snakes to the vertical slit of crocodilians and nocturnal snakes and the horizontal slits of a few tree snakes. Undoubtedly the most bizarre pupil shape is that of some geckos, in which the pupil contracts to form a series of pinholes, one above the other. The lower eyelid has the greater range of movement in most reptiles. In crocodilians the upper lid is more mobile. Snakes have no movable eyelids, their eyes being covered by a fixed transparent scale. The tuatara and all crocodilians have a third eyelid, the nictitating membrane, a transparent sheet that moves sideways across the eye from the inner corner, cleansing and moistening the cornea without shutting out the light.

Visual acuity

Visual acuity varies greatly among living reptiles, being poorest in the burrowing lizards and snakes (which often have very small eyes) and greatest in active diurnal species (which usually have large eyes). Judging by the size of the skull opening in which the eye is situated, similar variation existed among the extinct reptiles. Those that hunted active prey (*e.g.*, the ichthyosaurs) had large eyes and presumably excellent vision; many herbivorous types (*e.g.*, the horned dinosaur Triceratops) had relatively small eyes and weak vision. Colour vision has been demonstrated in few living reptiles.

Hearing. The power of hearing is variously developed among living reptiles. Crocodilians and most lizards hear reasonably well, but snakes are not sensitive to airborne waves; the hearing capability of turtles is not known. The auditory apparatus in reptiles typically consists of a tympanum, a thin membrane located at the rear of the head; a small bone, the stapes, running between the tympanum and the skull in the tympanic cavity (the middle ear); the inner ear; and a eustachian tube connecting the middle ear with the mouth cavity. In reptiles that can hear, the tympanum vibrates in response to sound waves and transmits the vibrations to the stapes. The inner end of the stapes abuts against a small opening (the foramen ovale) to the cavity in the skull containing the inner ear. The inner ear consists of a series of hollow interconnected parts: the semicircular canals; the ovoidal or spheroidal chambers called the utriculus and sacculus; and the lagena, a small outgrowth of the sacculus. The tubes of the inner ear, suspended in a fluid called perilymph, contain another fluid, the endolymph. When the stapes is set in motion by the tympanum, it develops vibrations in the fluid of the inner ear; these vibrations activate cells in the lagena, the seat of the sense of hearing. The semicircular canals are concerned with equilibrium.

Most lizards can hear; details of the acuity of hearing, however, are largely unknown. The majority have a tympanum, tympanic cavity, and eustachian tube. The tympanum, usually exposed at the surface of the head or at the end of a short open tube, may be covered by scales or may be absent. In general the last two conditions are characteristic of lizards that lead a more or less completely subterranean life and presumably do not hear airborne sounds. The middle ear of these burrowers is usually degenerate as well, often lacking the tympanic cavity and eustachian tube.

Snakes have neither tympanum nor eustachian tube, and the stapes is attached to the quadrate bone on which the lower jaw swings. It is unlikely that snakes can hear airborne sounds, although they are obviously sensitive to vibrations in the ground. A loud sound above a snake does not elicit any response provided the object making the sound does not move or, if it does, the movements are not seen by the snake. On the other hand, the same snake will raise its head slightly and flick its tongue in and out rapidly if the ground behind it is tapped or scratched. Snakes undoubtedly "hear" these vibrations by means of bone conduction. Sound waves travel more rapidly and strongly in solids than in the air and are probably transmitted first to the inner ear of snakes through the lower jaw, which is normally touching the ground, thence to the quadrate bone, and finally to the stapes. Burrowing lizards presumably hear ground vibrations in the same way.

Crocodilians, all of which have an external ear consisting of a short tube closed by a strong valvular flap and ending at the tympanum, have rather keen hearing. The American alligator (Alligator *mississippiensis*) can hear sounds within a range of 50 to 4,000 cycles per second. The hearing of crocodilians is involved not only in detection of prey and enemies but also in their social behaviour, for males roar or bellow either to threaten other males or to attract females.

Although turtles have well-developed middle ears and usually large tympana, their ability to hear airborne sounds is still an open question. Measurements of the impulses of the auditory nerve between the inner ear and the auditory centre of the brain show that the inner ear in several species of turtles is sensitive to airborne sounds in the range of 50 to 2,000 cycles per second, but this does not prove that the animals are aware of the sounds.

Chemoreception. Chemical-sensitive organs, used by many reptiles to find their prey, are located in the nose and in the roof of the mouth. Part of the lining of the nose consists of cells subserving the function of smell and corresponding to similar cells in other vertebrates. The second chemoreceptor is Jacobson's organ, originally an outpocketing of the nasal sac in amphibians and remaining so in the tuatara and crocodilians. It has been lost by turtles. Jacobson's organ is best developed in lizards and snakes, in which its connection with the nasal cavity has been closed and is replaced by an opening into the mouth. The nerve connecting Jacobson's organ to the brain is a branch of the olfactory nerve.

Jacobson's organ

The use of Jacobson's organ is most obvious in snakes. If *a* strong odour or vibration stimulates a snake, its tongue is flicked in and out rapidly. Wtih each retraction the forked tip touches the opening of Jacobson's organ in the roof of the mouth, transmitting any chemical fragments adhering to the tongue. In effect, Jacobson's organ is a supplement to taste and is a short-range chemical receptor, as contrasted with the long-range testing of the true sense of smell located in the nasal tube.

Some snakes (notably the large vipers) and lizards (especially skinks and burrowing species of other farnilies) rely upon the olfactory tissue and Jacobson's organ to locate food, almost to the exclusion of other senses. Other reptiles, such as certain diurnal lizards and crocodilians, appear not to use scent in searching for prey, though they may use their sense of smell for locating a mate.

Some snakes, notably pit vipers, boas, and pythons, have special heat-sensitive organs on their heads as part of their food-detecting apparatus. Just below and behind the nostril of a pit viper is the pit that gives the group its common name. The lip scales of many pythons and boas have depressions (labial pits) that are analogous to the viper's pit. The labial pits of pythons and boas are lined with skin thinner than that covering the rest of the head and are supplied with dense networks of blood capillaries and nerve fibres. The facial pit of the viper is relatively deeper than the boa's labial pits and consists of two chambers separated by a thin membrane bearing a rich supply of fine blood vessels and nerves. In experiments using warm and cold covered electric light bulbs, pit vipers and pitted boas have been shown to detect temperature differences of less than 0.6" C (1.1° F).

Since many pit vipers, pythons, and boas are nocturnal and feed largely on mammals and birds, the facial sense organs enable them to direct their strikes accurately in the dark, once their warm-blooded prey arrives within range. The approach of the prey to that point is probably detected by the chemical receptors — either the nose, Jacobson's organ, or both.

Thermal relationships. Reptiles are often described as being cold-blooded, which is not always true. Their body temperatures are not always low, but they have no internal mechanism for regulating body temperature and thus approximate closely the temperature of their surroundings. This condition is termed poikilothermy. Mammals and birds maintain their relatively high body temperatures at a fairly constant level by physiological means that are independent of the external environment, a condition called homoiothermy. When the body temperature of a dog or a man falls below the normal range, he begins to shiver, blood vessels in the skin contract, muscular activity generates heat, and the contraction of the superficial blood vessels, by reducing the volume of blood flow at the surface, reduces heat loss by radiation. By contrast, a reptile, when its body temperature falls below the optimum, must move to some portion of the environment having a higher temperature; in less than optimal temperatures, its activity drops, its movements become sluggish, its heartbeat slows, and its rate of breathing drops. In short, it becomes incapable of the normal activities required for growth, reproduction, and survival.

Mammals and birds have some physiological means of cooling their bodies (*e.g.*, panting and sweating, expansion of superficial blood vessels), but a reptile must ordinarily move away from a spot in which the temperature is too high or it will perish very quickly. Some reptiles also pant, but most of their temperature accommodations are behavioral (*e.g.*, orienting to sun or wind, raising body from the ground).

Range of body temperatures

Each group of reptiles has its own characteristic thermal range. One genus of lizards, for example, may require temperatures of 29"–32" C (84°–90° F) for maximum efficiency, and another may require 24"–27" C (75°–81° F). As a result of such physiological differences, lizards of the two groups will be active at different times of the day or occupy slightly different habitats.

In general the normal activity temperatures of reptiles are lower than those of most mammals; however, a few sun-loving (heliothermic) lizards (*e.g.*, the greater earless lizard, Holbrookia texana, of southwestern United States) have average activity temperatures above 38° C (100" F), several degrees higher than the average human body temperature. Such high temperatures are exceptional, and the majority of lizards have normal activity temperatures in the 27°–35° C (81°–95° F) range.

EVOLUTION AND PALEONTOLOGY

Historical development. Reptiles occupy an evolutionary position between amphibians, on the one hand, and the birds and mammals on the other, the last two classes having evolved from reptilian ancestors. Reptiles first appear in the fossil record of the Carboniferous Period, over 280,000,000 years ago. By the Triassic, about 50,000,000 years later, they began to dominate the terrestrial life of the world and continued that dominance through the Mesozoic Era (65,000,000–225,000,000 years ago). Reptiles succeeded in adapting to deserts, swamps, forests, grasslands, rivers, lakes, and even the air and the seas. Coincident with the rise of mammals at the end of the Mesozoic, most reptilian groups became extinct.

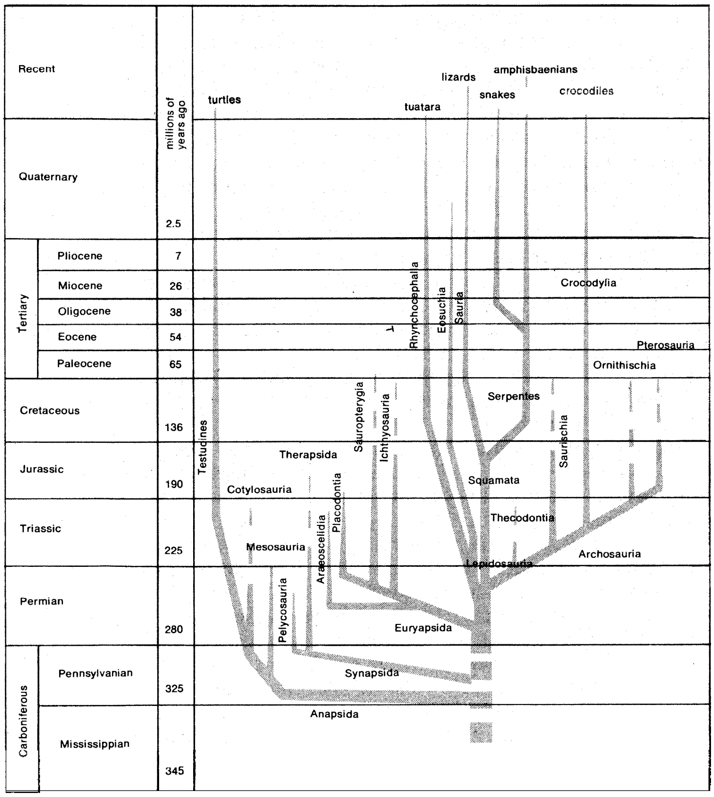


Figure 4: Family tree of the reptiles.

The big evolutionary step made by reptiles was the final emancipation from life in water, for, until that step was taken, vertebrates could not exploit all of the Earth's surface. That breakthrough required two basic changes, the first of which took place in the skin. Modem amphibians have naked skins that lack horny scales, hair, and other protective devices. One small amphibian group, the caecilians, has small fishlike scales embedded in the skin; similar scales occurred in certain extinct amphibians. Because of their thinness and position, such amphibian scales are no protection against desiccation, one of the principal hazards of life for all animals, vertebrate and invertebrate. This susceptibility to drying out forces amphibians to remain in water or in very humid places, thus limiting their exploitation of the terrestrial environment. Reptiles evolved a different type of scale consisting of keratin (or horn) deposited in the outermost layer of the skin. This type of scale was in a position and of a thickness to prevent desiccation.

Consequences of amniote egg

The second basic change made by the reptiles was the development of the amniote egg. This development expanded the possibilities of exploiting terrestrial environments. The egg could be laid under rocks or logs, in holes in the ground, in deserts or forests—in fact, almost anywhere except in water.

The development of the reptilian egg had several other consequences. An egg that is enclosed by a shell must be fertilized before that shell is deposited, thus necessitating internal fertilization. The evolution of a land egg also increased the efficiency of the life cycle of terrestrial vertebrates. An amphibian, hatching from an aquatic egg, must develop and grow in water in a larval form, the tadpole. The history of a reptile, on the other hand, is one of development and growth of adult structures adapted for a terrestrial life. It need not develop gills or a lateral line system (series of sense organs), which are needed by the aquatic tadpole and which must be resorbed and reworked into other structures. This important change in the type of development was made possible by a great increase in the amount of yolk in the reptilian egg.

The large amount of yolk also permitted the lengthening of the embryonic period, which in turn allowed the development of all structures that are necessary for successful existence on land. When a reptile hatches, it is ready to carry out all of the activities of the adult (with the exception of reproduction), and in the same environment and manner.

After reptiles acquired scaly protection for their skin and an egg that did not have to be laid in water, they were free to move over most of the Earth's surface. That freedom set the stage for the evolution of the many varied types of reptiles and, ultimately, for the evolution of birds and mammals.

Fossil distribution. What is known of the fossil record of reptiles shows that most of the major groups, or orders, were worldwide or nearly so at some time in their individual histories. Few orders are known from South America and Australia; the absence of most major groups from these areas more likely is explained on the basis of lack of preservation or lack of discovery of fossil beds than on the basis of a genuine absence of the animals throughout such a long interval as the Mesozoic. In the following discussion the names of present-day continents are used, though it should be understood that continental outlines in the past did not always coincide with those of today.

Few orders of reptiles are known from Paleozoic times —*i.e.*, Carboniferous (280,000,000–345,000,000 years ago) and Permian (225,000,000–280,000,000 years ago). The stem reptiles, the Cotylosauria, have been found in Carboniferous deposits of eastern and western North America and western Europe and in Permian beds of the Soviet Union and Africa. Presumably, they also lived in Asia during this interval of over 100,000,000 years, but their remains have yet to be found there. In the same period the Pelycosauria lived in North America and Europe, where their fossils are well-known, and possibly in Africa and Asia. The related mammal-like Therapsida and perhaps other forms were fossilized in Africa and

Drawing by J. Helmer based on (*Hadrosaurus, Brontosaurus*) photographs courtesy of Field Museum of Natural History, Chicago; (others) Colbert. *Age of Reptiles*, Weidenfeld & Nicolson

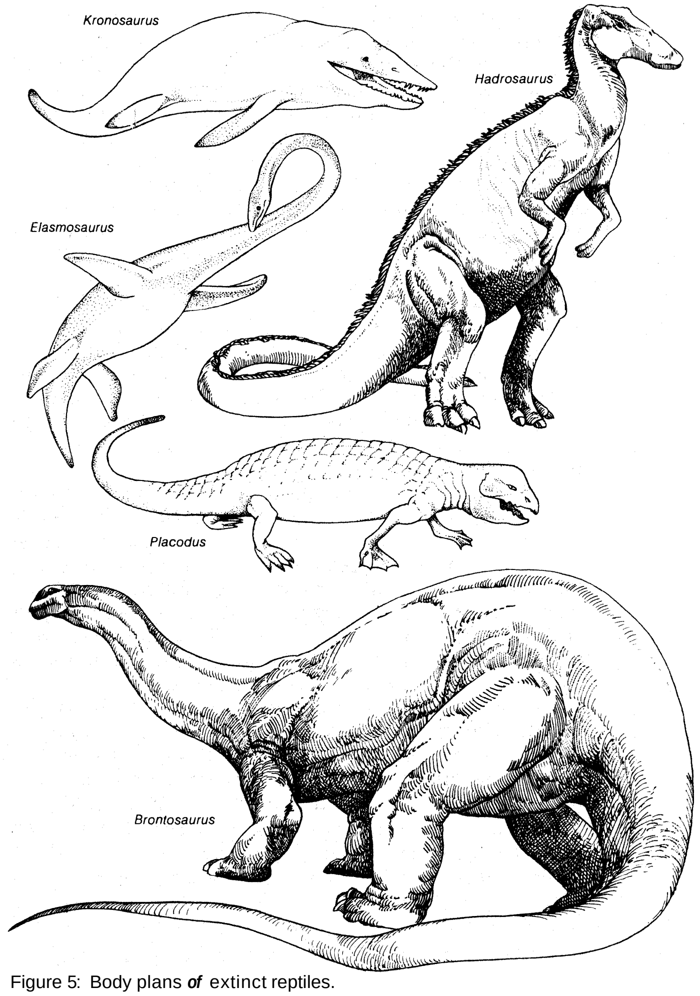


Figure 5: Body plans of extinct reptiles.

Europe. The probable ancestors of turtles appeared in Africa, and the first diapsids (reptiles having two-arched temporal structures) appeared in Africa and Europe in the Permian.

By the Triassic (190,000,000–225,000,000 years ago), the earliest portion of the Mesozoic, the mammal-like reptiles had spread to all of the continents except Australia. Turtles were still in Africa and had spread at least as far as Europe. The Ichthyosauria were living in seas covering what is now western North America and western Europe and may have been much more widely distributed, considering their oceanic habitat. North America at that time was also the home of primitive diapsids (Thecodontia), phytosaurs (suborder Phytosauria), and the earliest dinosaurs (Saurischia) and crocodiles. Besides the mammal-like therapsids, Eurasia and Africa in the Triassic had phytosaurs, the first Rhynchocephalia, and the early dinosaurs. The major groups of reptiles, therefore, were essentially worldwide in distribution by the Triassic.

Appearance of giant dinosaurs

The giant dinosaurs began their efflorescence in the Jurassic (136,000,000–190,000,000 years ago). The big carnivorous types, such as *Allosaurus*, roamed the landscapes of the major continents, presumably preying on even larger herbivorous dinosaurs (Brontosaurus, Diplodocus, etc.), whose remains have been found in North America, Europe, Africa, and Australia. Marine ichthyosaurs and plesiosaurs (order Sauropterygia) still swam in the shallow seas of both hemispheres. The ornithischian dinosaurs (Ornithischia), the ancestors of the duck-billed and horned dinosaurs, became widely distributed in the Jurassic. One group, the armoured stegosaurians (suborder Stegosauria), left fossils in North America, Europe, and Africa. The flying reptiles (Pterosauria) also made their appearance at least in Africa and Europe during the Jurassic.

The culmination of dinosaur evolution occurred in the Cretaceous (65,000,000–136,000,000 years ago), when every part of the world had herbivorous ornithischian dinosaurs: the ankylosaurs (suborder Ankylosauria), medium-sized dinosaurs armoured with heavy plates and large spines, ranging from South America to Africa; the duck-billed dinosaurs (suborder Ornithopoda), ranging from North America to Africa; and the horned dinosaurs (suborder Ceratopsia) in North America and Asia. Everywhere they were preyed upon by the big carnivorous types, which culminated in North America in the gigantic carnivore Tyrannosaurus. The skies over North America, Africa, and Europe (and probably Asia and South America as well) were the province of the flying reptiles (Pterosauria). There were also inconspicuous groups that later inherited the reptilian world. Lizards (suborder Sauria) appeared in most continents, and snakes (suborder Serpentes) appeared in some places. Birds, having splintered off from reptilian ancestors in the Jurassic, became more numerous in the Cretaceous; mammals, which ultimately replaced most of the reptiles, were represented on most continents by small creatures.

From The Procession of *Life* by Alfred S. Romer (1968); Universe Books, New York, and Weidenfeld & Nicolson Ltd.

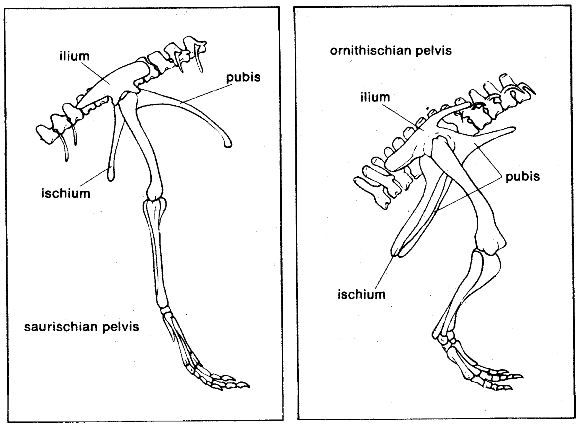


Figure 6: The types of dinosaur pelvis.

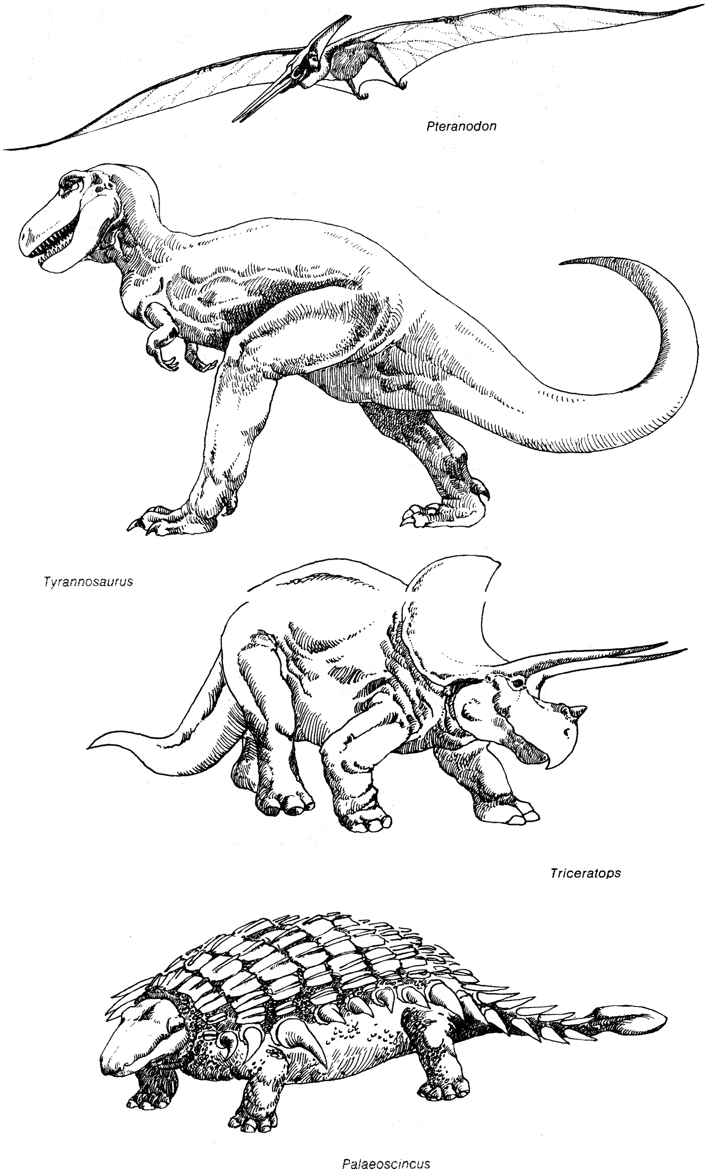


Figure 7: Body plans of extinct reptiles.

Drawing by J. Helmer based on (Tyrannosaurus, *Palaeoscincus*) photographs courtesy of Field Museum of Natural History. Chicago

Through all this evolutionary activity, the conservative turtles continued their plodding evolutionary pace, changing but little, yet lasting through all. (Ed.)

CLASSIFICATION

Distinguishing taxonomic features. The major reptile groups are distinguished on the basis of vertebral and skull features, particularly the number and positions of the temporal fenestrae (*i.e.*, large openings in the temporal bone). Beyond these, the pelvic structure and that of the teeth and limbs are important. Any extraordinary structural development (*e.g.*, the wings of pterosaurs, the shell of turtles) is also a major factor in the system of classification. In recent reptiles (four out of some 17 orders), the structure of the heart, the male secondary sex organs, the extent and kind of dermal armour, and the structure of other soft organs are used. Discoveries in serology (the study of blood serum) and karyology (the study of cell nuclei) have had little effect, as yet, on the systems of classification.

Annotated classification. The following classification of the reptiles is based on that of A.S. Romer (1956, 1966), the American vertebrate paleontologist, as later modified by himself (1968) and in minor respects by E.H. Colbert (1965), B.W. Halstead (1969), and H.G. Dowling (1971). No single system of reptile classification is acceptable to all herpetologists, and widely differing views are held by various authorities. Considerable change in the recognition and content of both major and minor categories is to be expected. Groups marked with a dagger (†) are extinct and known only from fossils.

CLASS REPTILIA

Air-breathing, vertebrate animals without hair or feathers, the body usually covered with infolded epidermal scales. The occipital condyle (a protuberance where the skull attaches to the 1st vertebra) is single (except in transitional forms such as Therapsida), and representatives with well-developed limbs have 2 or more sacral vertebrae. The single auditory bone, the columella, is equivalent to the mammalian stapes. The lower jaw is made up of several bones and connects to the braincase by way of the quadrate bone and often by way of the supratemporal, the squamosal, or both. The systemic arch (part of the aorta) is paired; the respiratory and systemic portions of the circulatory system are incompletely separated. The red blood cells contain nuclei. Reproduction typically is by leathery-shelled or calcareous-shelled eggs with specialized membranes (chorion and amnion) that help to protect the embryo. In some lizards and snakes and in some extinct reptiles (*e.g.*, ichthyosaurs) the eggs are retained in the oviducts of the mother, sometimes with a placental connection, and the young are born alive. There are about 6,000 living species.

Subclass Anapsida

Pennsylvanian to present. Skull typically without temporal openings.

†*Order* Cotylosauria (cotylosaurs)

Lower Pennsylvanian (300,000,000–325,000,000 years ago) to Upper Triassic (190,000,000–210,000,000 years ago). Vertebrae amphicoelous (*i.e.*, concave at both ends) with small intercentra (crescentic elements between vertebrae); neural arches (the portion of the vertebra that encloses the nerve cord from above) convex. Skull with large toothed flange on pterygoid (a bone of the lower part of the skull); a lateral temporal fenestra (between jugal and squamosal) in some advanced forms. Mostly small and lizard-like.

?Order Mesosauria (mesosaurs)

Upper Pennsylvanian (280,000,000–300,000,000 years ago) or Lower Permian (250,000,006–280,000,000 years ago) of South Africa and South America. A small fish-eating aquatic reptile with cotylosaur-like vertebrae. The temporal region of the skull is poorly known, but there may have been a lateral fenestra. Slender, long-jawed reptiles 40–90 cm long.

Order Testudines (or Chelonia: turtles)

Upper (possibly Middle) Triassic to present. Skull without pineal (on the midline of the "forehead") or temporal fenestrae, though the temporal region may be emarginated, or indented. Jaws toothless (though there were palatine teeth in some extinct forms), with a horny beak. A shell (dorsal carapace and ventral plastron) covers the body and encloses the pectoral and pelvic girdles. Most modern turtles have shells less than 60 cm long, but oceanic forms have larger ones (2.7 m), and those of some extinct turtles exceeded 3.6 m in length.

Subclass Lepidosauria (lepidosaurians)

Upper Permian *(225,000,000–250,000,000* years ago) to present. Primitive forms had 2 openings in the temporal region of the skull; most of the descendants have lost the lower (jugal–quadratojugal) temporal arch. The earliest known forms already had a jaw that was shortened. Usually a pineal eye (degenerate, median eyelike structure), no trend toward bipedalism. Ribs single-headed in advanced forms.

†*Order* Eosuchia (eosuchians)

Upper Permian to Eocene (38,000,000–54,000,000 years ago). Primitive lizard-like reptiles with 2 temporal arches and without a beaked snout.

Order Rhynchocephalia (beaked reptiles)

Lower Triassic to present. Scaled reptiles with 2 temporal arches. The premaxillary overhangs the lower jaw as a beak; the teeth are acrodont (*i.e.,* attached to the edge of the jaw rather than inserted in sockets). The vertebrae are amphicoelous, and the ribs are single-headed. Mostly lizard-like forms 30–90 cm long; one group (rhynchosaurs) attained lengths of up to 180 cm. One living species.

Order Squamata (scaly reptiles)

Lizards, snakes, and amphisbaenians. Upper Triassic to present. The quadrate is freed by loss of the lower (jugal–quadratojugal) arch and reduction of the squamosal to allow some movement at both ends. Vertebrae procoelous (*i.e.*, with the centre part concave on the anterior side, convex on the posterior side) except in a few geckos (Sauria), in which they are amphicoelous. Living species possess unique paired copulatory structure (hemipenes).

Suborder Sauria (lizards). Upper Triassic to present. Most generalized suborder; most species with well-developed limbs, an external ear opening, movable eyelids, or some combination of these structures. The skull typically has a pineal opening, and epipterygoid, lacrimal, and jugal bones. About 3,000 living species (see SAURIA).

Suborder *Amphisbaenia* (amphisbaenians). Eocene to present. Highly specialized, limbless. burrowing reptiles with the eyes hidden under the skin, no pineal opening, and the body scales fused into annuli, or rings. The skull is solidly constructed as a burrowing wedge. About 140 living species.

Suborder Serpentes (snakes). Upper Cretaceous (65,000,000–100,000,000 years ago) to present. A highly specialized group, without pectoral limbs or girdle, pelvic limbs rudimentary when present. Without external ear or movable eyelids. Upper temporal arch (postorbital–squamosal) absent, leaving quadrate movable at both ends. No pineal opening; no epipterygoid, lacrimal, or jugal bones. About 2,500 living species (see SERPENTES).

Subclass Archosauria (ruling reptiles)

Upper Permian to present. Reptiles with 2 temporal openings (diapsid) and a tendency toward bipedalism. Most have long hindlegs and short forelimbs. Typically without a pined opening in the skull, but with an antorbital fenestra and one on the outer surface of the lower jaw. Ribs typically 2-headed, at least anteriorly. Ischium and pubis (bones of pelvis) elongated. Teeth in deep sockets (thecodont). Most with some armour.

†*Order* Thecodontia (primitive archosaurs)

Upper Permian to Upper Triassic. Lightly built bipedal or more heavily armoured reptiles, some (phytosaurs) crocodile-like and presumably amphibious, but with nostrils far back on the snout. Most had at least 2 rows of bony plates along the spine.

Order Crocodylia (crocodilians)

Upper Triassic to present. Aquatic or amphibious reptiles, rather generalized in body form, but with a flattened skull, the nostrils on the tip of the snout, and a well-developed secondary palate. Typically, the pubis is excluded from the acetabulum, or hip socket, and the 5th toe is reduced to a stump; 21 living species (see CROCODILIA)

†Order Saurischia (carnivorous dinosaurs and giant herbivorous dinosaurs)

Middle Triassic (200,000,000 years ago) to Upper Cretaceous. Pelvis triradiate (*i.e.*, 3-branched). Some reduction in digits. Forelimbs usually distinctly shorter than hind. Three to 7 sacral vertebrae. Some herbivorous forms were more than 24 m long.

†*Order* Ornithischia (herbivorous dinosaurs)

Upper Triassic to Upper Cretaceous. Pelvis tetraradiate (*i.e.*, 4-branched). Typically with a beaklike structure in the front part of the mouth and grinding teeth in the rear. Toes often with hooflike structures. Many with heavy armour and horns. Largest about 9 m long.

†*Order* Pterosauria (pterydactyls)

Lower Jurassic (150,000,000–190,000,000 years ago) to Upper Cretaceous. Highly specialized flying reptiles with hollow bones; 4th digit of the forelimb greatly elongated to support the flying membrane of the wing. Early forms toothed and with long tails; later forms tendedto be larger and to have lost both teeth and tail.

†**Subclass Euryapsida** (plesiosaurs and relatives)

Lower Permian to Upper Cretaceous. Mainly aquatic reptiles with an upper temporal opening (between postorbital, squamosal, and parietal bones), and a broad plate of bone below.

†*Order* Araeoscelidia (primitive euryapsids)

Lower Permian to Upper Triassic. Primitive terrestrial reptiles with lizard-like proportions, some with a specialized cervical region.

?Order Sauropterygia (nothosaurs and plesiosaurs)

Middle Triassic to Upper Cretaceous. Aquatic reptiles with strongly developed ventral ribs, dorsally placed nostrils, and a highly modified palate (pterygoids and often palatines are joined in the midline). Limbs paddle-like in advanced forms.

†*Order* Placodontia (placodonts)

Lower to Upper Triassic. A side branch of euryapsids, apparently mollusk eaters. In some the body was armoured and turtle-like in form.

†*Order* Ichthyosauria (ichthyosaurs)

Middle Triassic to Upper Cretaceous. Highly aquatic reptiles with porpoise-like bodies, a dorsal fin, and a reversed-heterocercal tail (*i.e.*, with the lower lobe longer than the upper). Limbs paddle-like; snout often elongated and beaklike.

†**Subclass Synapsida** (mammal-like reptiles)

Lower Pennsylvanian to Middle Jurassic (160,000,000 years ago). A single lateral temporal opening with the postorbital and squamosal bones joining above it in primitive forms; the opening extends upward to the parietal in later forms. Pineal present. Teeth differentiated. Two coracoids in pectoral girdle.

†*Order* ***Pelycosauria*** (primitive synapsids, pelycosaurs)

Lower Pennsylvanian to Middle Permian (250,000,000 years ago), especially in Europe and North America. Neural arches higher and less swollen than in the contemporary cotylosaurs. Abdominal ribs present in most.

†*Order* ***Therapsida*** (advanced synapsids, therapsids)

Middle Permian to Middle Jurassic, mainly in South Africa. Temporal opening expanded in advanced forms, a secondary palate formed. Occipital condyle double and dentary bone much enlarged.

Critical appraisal. A natural classification of reptiles is more difficult than that of many animals because the main evolution of the group was during Mesozoic time; 13 of 17 recognized orders are extinct. The consequent reliance on osteological (bone) characters may have obscured some important evolutionary trends, and there is little agreement among herpetologists and paleontologists on reptile taxonomy. Even the major categories of reptile classification are still in dispute. Although the ideas of Watson, Colbert, and Romer have dominated the field, some authorities question most of the basic elements—from subclass to suborder--on which the framework of their classification depends. Halstead, for example, would discard the entire system of reptile subclasses and recognize a "superclass Reptilia" with seven reshuffled "classes."

On the other hand, there is general agreement that the base reptilian stock is the Cotylosauria, which evolved from an amphibian labyrinthodont stock (the captorhinomorphs) at about the Mississippian–Pennsylvanian transition. It is also quite clear that the cotylosaurs early divided into two lines, one of which (the pelycosaurs) represented the stock that gave rise to the mammals. Another branch led to all of the other reptiles, and, later, to the birds as well. Thus, most of the questions of reptilian evolution and classification deal with inter-reptilian relations, rather than with their relationships with other animals.

BIBLIOGRAPHY

General works: ANGUS D'A. BELLAIRS, ***The Life of Reptiles*** (1970); ROGER COWANT, ***A Field Guide to Reptiles and Amphibians (Eastern U.S. and Canada)*** (1958); P.J. DARLINGTON, ***Zoogeography*** (1957); CARL OANS ***et al.*** (eds.), ***Biology of the Reptilia,*** 3 vol. (1968–70); C.J. and O.B. GOIN, ***Introduction to Herpetology,*** 2nd ed. (1971); ARTHUR LOVERIDGE, ***Reptiles of the Pacific World*** (1945); ROBERT MERTENS, ***La Vie des amphibiens et reptiles*** (1959; Eng. trans., ***The World of Amphibians and Reptiles,*** 1960); J.A. OLNER, ***The Natural History of North American Amphibians and Reptiles*** (1955); J.A. PETERS, ***Dictionary of Herpetology*** (1964); C.H. POPE, ***The Reptile World*** (1955); C.L. PROSSER and F.A. BROWN, JR., ***Comparative Animal Physiology,*** 2nd ed. (1961); A.S. ROMER, ***Osteology of the Reptiles*** (1956); K.P. SCHMIDT and R.F. INGER, ***Living Reptiles of the World*** (1957); M.A. SMITH, ***The British Amphibians and Reptiles*** (1951); R.C. STEBBINS, ***A Field Guide to Western Reptiles and Amphibians*** (1966).

Paleontology: E.H. COLBERT, ***The Dinosaur Book*** (1945) and ***The Age of Reptiles*** (1965); A.S. ROMER, ***Vertebrate Paleontology,*** 3rd ed. (1966).

Snakes: L.M. KLAUBER, ***Rattlesnakes: Their Habits, Life Histories, and Influence on Mankind*** (1956); C.F. KAUFFELD, ***Snakes and Snake Hunting*** (1957); A.H. and A.A. WRIGHT, ***Handbook of Snakes of the United States and Canada,*** 2 vol. (1957).

Lizards: H.M. SMITH, ***Handbook of Lizards: Lizards of the United States and Canada*** (1946).

Turtles: A.F. CARR, ***Handbook of Turtles: The Turtles of the United States, Canada, and Baja California*** (1952).

(H.G.Do.)

Research and Development, Industrial

Research and development, a phrase unheard of in the early part of the 20th century, became in the latter part a universal watchword in industrialized nations. The concept of research is as old as science, whereas the concept of the intimate relationship between research and subsequent development was not generally recognized until after World War II.

Basic and applied research

The distinction between basic and applied research was established in the first half of the present century; basic research was defined as the work of scientists and others who pursued their investigations without conscious goals apart from the desire to unravel the secrets of nature. In modem industrial research and development programs, basic research, though sometimes called pure research, is usually not entirely pure; it is commonly directed toward a generalized goal, such as the investigation of a newly discovered frontier of technology that promises to relate to the problems of a given industry. Applied research carries the findings of basic research to a point at which they can be exploited to meet a specific need. Development includes the steps necessary to bring a new or modified product or process into production. In Europe, the U.S., Japan, and, to a lesser extent, elsewhere, the unified concept of research and development has been an integral part of economic planning, both by government and by private industry.

HISTORY

The origins of industrial research. The first organized attempt to harness scientific skill to communal needs took place in the 1790s, when the young revolutionary government in France was defending itself against most of the rest of Europe. The results were remarkable. Explosive shells, the semaphore telegraph, the captive observation balloon, and the first method of making gunpowder with consistent properties all were developed during this period.

The lesson was not learned permanently, however, and another half-century was to pass before industry started to call on the services of scientists to any serious extent. At first the scientists consisted only of a few gifted individuals. Robert W. Bunsen, in Germany, advised on the design of blast furnaces. William H. Perkin, in England, showed how dyes could be synthesized in the laboratory and then in the factory. William Thomson (Lord Kelvin), in Scotland, supervised the manufacture of telecommunication cables. In the U.S., Leo H. Baekeland, a Belgian, produced Bakelite, the first of the plastics. There were inventors, too, such as John B. Dunlop, Samuel Morse, and Alexander Graham Bell, who owed their success more to intuition, skill, and commercial acumen than to scientific understanding.

Early industrial laboratories. While industry in the U.S. and most of western Europe was still feeding on the ideas of isolated individuals, in Germany a carefully planned effort was being mounted to exploit the opportunities that scientific advances made possible. Siemens, Krupp, Zeiss, and others were establishing laboratories and, as early as 1900, employed several hundred people on scientific research. In 1870 the ***Physicalische Technische Reichsanstalt*** was set up to establish common standards of measurement throughout German industry. It was followed by the Kaiser Wilhelm Gesellschaft (later renamed after Max Planck), which provided facilities for scientific cooperation between companies.

In the U.S., the Cambria Iron Company set up a small laboratory in 1867, as did the Pennsylvania Railroad in 1875. The first case of a laboratory that spent a significant part of its parent company's revenues was that of the Edison Electric Light Company, which employed a staff of 20 men in 1878. The U.S. National Bureau of Standards was established in 1901, 31 years after its German counterpart, and it was not until the years immediately preceding World War I that the major American companies started to take research seriously. It was in this period that General Electric, Du Pont, American Telephone & Telegraph Company, Westinghouse, Eastman Kodak, and Standard Oil set up laboratories for the first time.

Except for Germany, progress in Europe was even slower. When the National Physical Laboratory was founded in England in 1900, there was considerabie pubiic comment on the danger to Britain's economic position of German dominance in industrial research, but there was little action. Even in France, which had an outstanding record in pure science, industrial penetration was negligible.

The effect of World War I. World War I produced a dramatic change. Attempts at rapid expansion of the arms industry in the belligerent as well as in most of the neutral countries exposed weaknesses in technology as well as in organization and brought an immediate appreciation of the need for more scientific support. The Department of Scientific and Industrial Research in the U.K. was founded in 1915, and the National Research Council in the United States in 1916. These bodies were given the task of stimulating and coordinating the scientific support to the war effort, and one of their most important long-term achievements was to convince industrialists, in their own countries and in others, that adequate and properly conducted research and development were essential to success.

The inter-war years

At the end of the war the larger companies in all the industrialized countries embarked on ambitious plans to establish laboratories of their own; and, in spite of the inevitable confusion in the control of activities that were novel to most of the participants, there followed a decade of remarkable technical progress. The automobile, the airplane, the radio receiver, the long-distance telephone, and many other inventions developed from temperamental toys into reliable and efficient mechanisms in this period. The widespread improvement in industrial efficiency produced by this first major injection of scientific effort went far to offset the deteriorating financial and economic situation.

Unfortunately, the economic pressures on industry reached crisis levels by the early 1930s, and the major companies started to seek savings in their research and development expenditure. It was not until World War II that the level of effort in the U.S. and the U.K. returned to that of 1930. Over much of the European continent the Depression had the same effect, and in many countries the course of the war prevented recovery after 1939. In Germany, Nazi ideology tended to be hostile to basic scientific research, and effort was concentrated on short-term work.

Expansion after World War II. The picture at the end of World War II provided sharp contrasts. In large parts of Europe, industry had been devastated, but the U.S. was immensely stronger than ever before. At the same time, the brilliant achievements of the men who had produced radar, the atomic bomb, and the V-2 rocket had created a public awareness of the potential value of research that ensured it a major place in postwar plans. The only limit was set by the shortage of trained men and the demands of academic and other forms of work.

Since 1945 the number of trained men has increased each year. In most industrial countries, the increase has continued at about 5 percent each year as new graduates have become available. The U.S. effort has stressed aircraft, defense, and space. Indirectly, U.S. industry in general has benefitted from this work, a situation that compensates in part for the fact that in specifically nonmilitary areas the number of men employed in the U.S. is lower in relation to population than in a number of other countries.

Outside the air, space, and defense fields the amount of effort in different industries follows much the same pattern in different countries. An important point is that countries that, like Japan, have no significant aircraft or space industries have substantially more manpower available for use in the other sectors.

Research in the U.S.S.R.

It is unfortunate that few figures are available for the U.S.S.R. The high quality of Soviet work on aircraft and space exploration is evident, and most visitors to research laboratories have been impressed by the facilities and by the calibre of the staff. What is known of Soviet military and naval equipment indicates a satisfactory level of effort and competence. The general impression gained by some observers is that the rest of the industrial effort has suffered from the concentration of necessarily limited resources in particular fields. Without figures that command more confidence than any that are available, this must remain conjecture.

TYPES OF LABORATORIES

Company laboratories. Company laboratories fall into three clear categories: research laboratories, development laboratories, and test laboratories.

Research laboratories. Most of these laboratories carry out both basic and applied work. They usually support a company as a whole, rather than any one division or department. They may be located at a considerable distance from any other part of the company and have managements reporting independently to the board of directors. Bell Telephone Laboratories is an outstanding example. There the transistor and coaxial cable were developed, and pioneer work in satellite communications was carried out. In Frankfurt, Germany, the research centre of Farbwerke-Hoechst has successfully synthesized several new dyestuffs.

Development laborafories. These installations are specifically committed to the support of particular processes or product lines. They are normally under the direct control of the division responsible for manufacture and marketing and are often within walking distance of the manufacturing area. In Germany, for example, the Bayer Company's development laboratories have constructed car body shells made from a plastic material up to 55 percent lighter than steel plate but with the same strength.

Test laboratories. These laboratories may serve a whole company or group of companies or only a single manufacturing establishment. They are responsible for monitoring the quality of output. This often requires chemical, physical, and metallurgical analyses of incoming materials, as well as checks at every stage of a process. These laboratories may be part of a manufacturing organization, but many companies give them an independent status.

Government laboratories. ***United States government.*** The pattern followed by different countries varies widely. The general policy of the U.S. government has been not to set up laboratories of its own, even for military work, but to offer research and development contracts, usually on the basis of competitive bidding, to private companies. The most important reason for this has been a belief that the right place to develop equipment is very close to the place at which it will eventually be manufactured.

There are exceptions to the rule. One is the type of laboratory represented by the National Bureau of Standards, a central authority on problems of measurement and standardization. Another is the type of laboratory supported by the Department of Agriculture, set up by the government in the belief that research in this field is necessary but that the industry had neither the finances nor the organization to maintain it. The continuing support of successive administrations has resulted in a large and authoritative body carrying out research over a wide field for the benefit of the farming community and thus, indirectly, of the whole nation.

Atomic Energy Commission

A third type of government laboratory is represented by the Atomic Energy Commission. In this case the U.S. government recognized a situation of potential danger and also opportunity of such a nature that it was not practicable for it to be handled by private individuals. It therefore set up a body to deal with the situation, allocating funds directly and maintaining close control of the objectives and timing of research. A similar challenge is faced by the National Aeronautics and Space Administration. Although much of the detailed research and development work is contracted to private industry, overall control, as well as much of the most important work, is handled directly by the central organization.

U.K. and other countries. A different type of policy has been followed in the United Kingdom. A chain of government laboratories supports the requirements of the armed forces. The Royal Radar Establishment; the Signals Research and Development Establishment; and the Atomic Energy Research Establishment, Weapons

Group are three of the best known. These carry out a great deal of the basic and applied research from which new weapons and military techniques emerge, and they play a major part in negotiating and monitoring the contracts placed with private industry for the eventual development and production of equipment for the armed forces.

In addition to these military laboratories, the United Kingdom government supports a number of civilian establishments, such as the National Engineering Laboratory, the Road Research Laboratory, and the Hydraulics Research Laboratory. These have a considerable degree of independence in selecting projects that will bring the greatest benefit to industry as a whole, and their results are made available to all. They maintain close liaison with the research associations (see below, ***Research associations)*** and with private industry and attempt to concentrate their work in areas that for one reason or another are not covered elsewhere.

In Germany, as in the U.K., defense research is the responsibility of a chain of government laboratories, but they are much smaller. Most of the work is done for them on contract by the research associations. They place very little research with private industry and call upon it only in the later stages of development.

In Japan there is a chain of laboratories, such as the National Aerospace Laboratory, the National Institute of Resources, and the Technical Research and Development Institute for Defence, which serve the needs of government departments. They work closely with the research associations that support particular industries. The military laboratories carry out the bulk of defense research and development themselves, and they are also responsible for the placing of contracts with private industry. These are usually confined to the later stages of development and are expected to lead almost directly to production.

The French system is similar, but the directly controlled government laboratories are even smaller and do little more than direct and coordinate work done by the research associations.

Factors common in all countries

In spite of differences in organization, the day-to-day conduct of government-sponsored research and development in all countries has much in common. In every case a comparatively small number of government employees keep in constant touch with the whole of the scientific and technical community and dispense contracts in the way that they consider will make the best use of the resources available in the broad national interest. The fact that in some countries it is done in laboratories under direct governmental control, in others in those under private control, and in yet others in those in which responsibility is split, is of secondary importance. In every case, the men who hold the purse strings determine how the work will be done.

Independent laboratories. The concept of a laboratory that maintains itself solely by selling research originated with the Mellon Institute in Pittsburgh before World War I. The difficulties that have to be faced are formidable, for a great deal of research work yields no immediate or obvious reward and it is extremely difficult to satisfy a customer that he is getting value for his money. Nevertheless, a number of such bodies, including the Battelle Memorial Institute, Columbus, Ohio, and the Stanford Research Institute, Menlo Park, California, have become large and successful. These organizations offer the services of workers of high professional standing who cover between them a wide range of disciplines. They undertake studies and investigations on any subject within their competence for fees that are negotiated with each customer; and, although they do not expect to make profits, they are required to be self-supporting.

Another type of organization is represented by Arthur D. Little, Inc., Cambridge, Massachusetts, which is run on strictly commercial lines, seeking to make a commercially viable profit from the resources employed. Only one or two organizations of similar type have been established in western Europe, and they have not grown to a size comparable with those in America.

Both in Europe and in the U.S., there are a great number of small laboratories providing specialist analytical, spectographic, metallurgical, and similar services to industry. Most of their clients are companies that lack adequate facilities of their own and that in course of time either learn to stand on their own feet or go out of business. But the constant appearance of new companies and the increasing need for technical understanding in established companies results in a slow but steady increase in the number of independent specialist laboratories serving them.

Research associations. A more important part of the industrial research and development effort in western Europe and in Japan is represented by the research associations. Each is concerned with a single industry. Germany, France, and the U.K. each have about 50 of these associations, and Italy and Japan about half as many. The smallest employ 30 or 40 scientists, and the largest as many as 600. Examples are the British Glass Industry Research Association in Sheffield, the Institut Français du Pétrole, des Carburants et Lubrifiants in Paris, the Max-Planck-Institut für Eisenforschung in Diisseldorf, the Centro di Studio per l'Elettronica e le Telecomunicazioni in Milan, and the Textile Research Institute in Yokohama. These laboratories are mainly concerned with the long-term problems of the industries they serve, but they are on occasion called in to help with immediate technical difficulties beyond the powers of local staff.

In European countries other than the U.K., they carry out substantial work under contract to the defense departments.

Financing and control

The method of financing and controlling work varies slightly in different countries. In the U.K., costs are shared by government and industry, and control lies with councils elected by the companies served. In France and Germany the laboratories are government financed but largely independent in their activities. In Japan they are directly controlled by the appropriate ministries but are industry oriented, unlike the U.K. government civil establishments previously mentioned, which are applications oriented.

University laboratories. In principle, university laboratories are completely independent and free to investigate anything that interests them. In practice, many of them are anxious to keep in touch with industry and to focus their research effort on problems wtih direct applications. Similarly, industrial scientists wish to maintain contact with advanced academic research. The result is a constant interchange between universities and industry; industrialists suggest problems for university research and provide funds to support it, and university staffs act as consultants and advisers to industry. In addition, government may play a direct role by funding university research, especially in defense-related fields.

PROPRIETARY RESEARCH AND DEVELOPMENT

Product planning. The basic purpose of the laboratories of private industry is to provide new products for manufacture. One difficulty facing those who plan these projects is the relationship between development costs and predicted sales. In the early stages of development, costs are low. They increase to a maximum and decline slowly, disappearing as early production difficulties are overcome.

Similarly, production rises slowly at first, then more rapidly, and finally reaches a plateau. After a time, production starts to fall, sales declining gradually as the product becomes obsolete or abruptly as it is replaced by a new one.

At any particular time, a company may have a number of products at different stages of the cycle. Laboratory managers must ensure that the total development effort required is neither greater nor less than that available, and the production managers must be satisfied that the eventual demands upon their resources will be sufficient to keep them fully loaded but not overloaded.

New product proposals

To maintain such an optimum condition, a steady flow of new product proposals is required. Each must be studied by technical, commercial, financial, and manufacturing experts. Planning, then, consists of selecting for development new products that promise to employ the resources available in the most profitable manner. The laboratories have a key part to play in proposing projects as well as in carrying them out.

Types of project. Basic research, applied research, and development have been described earlier. The way in which projects are carried out can best be illustrated by examples typical of each, taken from different industries.

Basic research. In the mid-1960s the scientists of a company manufacturing memory systems for electronic computers reported to their management that the magnetic core matrices then in general use could eventually be superseded by optical means. They were instructed to carry out a basic research project in this field. A small team spent several weeks studying the available literature and produced a plan for a specific project.

Their method involved printing information in the form of black and white marks on a plate, illuminated by a very finely focussed beam of light. The light was designed to scan the plate by being passed through crystals made of a material that deflects light by an amount that depends upon the electric potenial applied to it. As it falls upon a black or white mark, the change in reflected light can be indicated by a photosensitive detector.

The project was broken down into a number of tasks. One group studied the printing of information on plates; another studied the problem of focussing a light beam. A third team studied the problem of the crystals that were to deflect the beam.

Eventually there emerged a system that demonstrated that an optical memory store was technically possible. Its performance was then compared with the requirements determined by market research. The improvements and additions necessary to make possible the design of a salable product were listed, and in due course these became the subject of applied-research projects.

Applied research. A classical applied-research project was carried out by an English pharmaceutical firm in 1938.

Basic research work had shown that a benzene ring with an amino group in one position and a sulfonamido group in the opposite one had antibiotic properties. Substitution of the hydrogen atoms in the groups changed both the effectiveness of the substance in killing bacteria and its toxicity to the patient. The problem was to find the substitution that gave the best combination. A team of chemists set out to synthesize every possible derivative of the original substance, while teams of bacteriologists and biologists determined the effect of each one on bacteria and on mice.

Gradually a pattern emerged. Certain types of substitution increased one effect; others increased the other. As the more promising substances were identified, medical workers tested them on human patients, and eventually one compound was found to give the best ratio of antibiotic power to toxicity.

Testing for the optimum compound

This completed the applied-research stage, but development was still required. The synthesis of the drug had to be transferred from the laboratory bench to the factory, and this had to be associated with marketing plans. Finally, the checks and tests by which manufacture could be controlled and the final product monitored had to be established.

Development. The problems of development are illustrated by a program that resulted in a new model of a transistor radio. Technical uncertainties had been resolved before the start of the project. The components to be used had been established, and new techniques required had been explored in applied-research projects. Each stage in the development program had to result in a set of drawings and a list of materials that would enable production engineers who knew very little about the final product to set up lines upon which it could be manufactured. The whole effort was aimed at a particular level of output at a chosen date, which was linked with a sales plan. Each stage in the design had a firm date for completion, and each part of the equipment was required to pass performance and environmental tests specified at the start of the project.

Value engineering and cost analysis. In the areas in which technology advances fastest, new products and new materials are required in a constant flow, but there are many industries in which the rate of change is gentle. Although ships, automobiles, telephones, and television receivers have changed over the last quarter of a century, the changes have not been spectacular. Nevertheless, a manufacturer who used methods even ten years old could not survive in these businesses. The task of laboratories working in these areas is to keep every facet of the production process under review and to maintain a steady stream of improvements. Although each in itself may be trivial, the total effect is many times as large as the margin between success and failure in a competitive situation.

These efforts to improve existing products and processes have been formalized under the titles of value engineering and cost analysis.

In value engineering every complete product and every component has its primary function described by a verb and a noun. For example, an automobile's dynamo, or generator, generates electricity. The engineer considers all other possible methods of generation, calculates a cost for each, and compares the lowest figure with that for the existing dynamo. If the ratio is reasonably close to unity, he can accept the dynamo as an efficient component. If not, he examines the alternatives. The same treatment is applied in tum to each of the parts out of which the chosen component is built, until it is clear that the best possible value is being obtained.

Cost analysis approaches the same fundamental problem from a different angle. It takes each part of an assembly and calculates the cost of obtaining it, taking full account of purchased material, labour, and other factors. This focusses attention upon the most expensive items and makes it possible to apply the principal effort in seeking economies at the points of maximum reward.

These two processes are unending. Every new material, every new manufacturing technique, every new way of carrying out an operation gives the engineer a chance to improve his product, and it is from these continuing improvements that the high degree of economy and reliability of most modem equipment derives.

Management of research. There are wide differences in the organization and management of industrial laboratories. Many of these are due to the fact that, unlike the majority of human activities in which it is desirable that the participants should concentrate on their work without excessive discussion, it is crucial to a research and development program that there should be continual discussion and mutual criticism among the participants. The organization must encourage this by keeping the appropriate people in close contact. At the same time, the chain of command must be such that decisions are made promptly and implemented without confusion or misunderstanding.

A system of organization common in research departments is subdivision according to scientific specialty. The laboratory manager allocates each project to a particular area or splits it among several. Each worker is surrounded by colleagues of similar background who can give him the help and stimulation that he needs. In another type of structure, the subdivisions are stages in the research process, such as basic research, applied research, development, and design. This type of structure is appropriate to establishments in which the problems are of a more straightforward nature and the chief requirement is to maintain time schedules and engineering standards. It is frequently used by companies that produce a large number of fairly similar products.

Organizing by project

Still another type of organization is common in laboratories handling a small number of large projects. Here the individual project is the basic organizational unit. Each project leader reports directly to top management and controls a staff that fluctuates with the requirements of the project. This requires that staff be able to move freely from one area to another and involves situations in which project leaders direct the activities of technical experts whose status and remuneration may be higher than their own.

Whatever management organization is chosen, it faces the task of controlling projects that contain unknown and unpredictable factors. When they are reasonably simple, this can be done by the use of bar charts. Such charts break the project down into discrete activities, each of which is estimated individually, and make it possible to determine the manpower requirements week by week. As the work goes forward, progress can be marked on the chart and an estimate made of the effect of any delay or difficulty upon the date of completion.

Control systems. In the mid-1950s, sophisticated methods of project control began to appear. The original systems were known by initials such as PERT (program evaluation and review technique) and CPA (critical path analysis). Many variations and extensions are in use. They have proved particularly valuable in the development of large military networks in which the work of hundreds of separate contractors has to be coordinated, and their use is now common in all types of research and development projects, as well as in civil engineering and construction work.

Monitoring costs. It is necessary to monitor not only technical progress and time but also costs. The simplest method is to calculate the cost of each man in terms of direct salary and overhead and then to record the number of hours that he spends each week on a particular project. To this is added the cost of materials and services used.

Certain overhead items, such as the cost of lighting, heating, floor space, and depreciation of equipment, are simple to assess, but there is considerable variation of practice in the allocation of the costs of individuals. In some laboratories, senior staff, who spend much of their time in administrative and managerial activities, are treated as overhead and not charged directly to specific projects. In others an attempt is made to allocate their cost in as realistic a manner as is possible among the projects with which they are concerned.

Similar differences exist in the treatment of laboratory assistants, print-room services, consultants, and computer time. Every increase in the number of items charged directly to projects increases the accuracy of management costing and control, but it also increases the amount of nonproductive and paper-generating work; thus, a balance must be struck.

A similar balance must be struck in the charge made against a project for each hour spent on it. If each man has his own rate calculated in terms of total remuneration per hour of work, including salary and fringe benefits, much time is spent in unfruitful calculation. A common compromise is to divide staff into two or three grades and to have a flat charge for each grade. Even here there is a discrepancy between the laboratories that charge the same rate of overhead for each grade and those that charge an amount proportional to the average salary.

THE ROLE OF GOVERNMENT

World War I brought home to every government involved the importance of having its armed forces supported by an industry using the most advanced scientific techniques. Since then it has been generally accepted that it is frequently desirable to encourage research and development for reasons of economic growth as well as national security. This has resulted in massive support from public funds for many sorts of laboratories.

Through World War II this support was limited to research and development of direct military significance, but in more recent years the types of equipment used by the armed forces have become so extensive and so complicated that it is no longer practicable to distinguish between the requirements of an efficient armament industry and those of an efficient civilian industry. Advanced communication systems, aircraft engines, computers, and nuclear-power generators have been just as important to one as to the other. This fact has led governments to become the greatest single sponsors of industrial research, and by 1970 the U.S. government alone was supporting $10,000,000,000 worth of work.

"Spin-off"

During the 1960s it became clear that the "spin-off," or civilian and commercial application of work done under defense contracts, was giving the industries who benefitted a crucial advantage over their competitors, particularly over those in countries in which comparable assistance was not available. The dominance of U.S. firms in computer development and in microelectronics was generally attributed to this cause, and the outstanding success of the British aeroengine industry could hardly have been achieved without it. There were obvious examples, such as the communication satellites, which derived from work on military rocket propulsion, and more subtle ones, such as the highly reliable electronic components, developed to make weapons safe, which made it possible to produce television sets with far longer life between failures than before. The reaction of most industrial countries was to increase government support of private research. In the U.K. the Ministry of Technology, set up in 1964, took responsibility for allocating funds to private industry for research projects with no direct military application. The usual practice has been to contribute 50 percent of the cost of the work, the private company providing the balance.

Placing of research contracts. In the U.S. and in most western European countries, research contracts placed by government departments originate in the decision of a scientifically or technicallv oriented executive of the department that certain work should be done. This leads to the preparation of a specification of the work, which is then offered to industry, to private research institutes, and to universities for competitive bidding.

The terms of contract have varied widely. During the years of World War II it was common to offer contracts on a cost-plus basis. The contractor kept records of the hours worked by his staff and the materials used; these were checked by government auditors and paid for at a negotiated rate together with a fixed percentage as profit. Criticisms of this system led to fixed-price contracts, but these have the drawback that it is often so difficult to define the end point of a research contract that the contractor can treat a fixed-price agreement as if it were cost-plus. Another problem is that when the end point can be exactly defined, but there are genuine uncertainties in the program, the most attractive bid may come from a contractor who, through ignorance, takes too light a view of the difficulties. Yet another formula that has been tried is to offer contracts on a cost-plus-fixed-profit (rather than cost-plus-percentage) basis.

Fixed-price contracts

In all these cases the main concern of the officer who sponsors the contract is to get the work done as efficiently as possible. With the many uncertainties of research and development, true economy is more likely to lie in high-quality work than in low pricing. Consequently, in every country in which the government is a substantial supporter of private research and development, the departments concerned have set up elaborate systems of monitoring work and of keeping in touch with the performance and capabilities of the companies willing to undertake it. In negotiating contracts, the sponsors attempt to place them where they will be handled most successfully. At the same time, they are concerned to keep together teams that are likely to do good work for them in the future. Within this framework the struggle of the customer to negotiate the best price that he can for the work that he wants and that of the contractor to get a good return for the commitment of his resources follows normal commercial practice.

Patent rights. There are wide variations in the way patent rights that arise from research contracts are treated. In some cases the rights are the exclusive property of the government, and in others they belong to the contractor. A common compromise is for the government to retain all rights when anyone uses the patents to supply a government department but for the contractor to retain them when another party is involved. Thus, the government can place production orders with any contractor that it chooses, and the company that carried out the development is obliged to release information to him. If, however, the new contractor wishes to sell in the open market, he is obliged to negotiate a license and pay a royalty to the original development laboratories.

BIBLIOGRAPHY. Very little was written on the subject of industrial research and development until the end of World War II. Since then an increasing number of books have appeared, and articles in technical and management journals have become too numerous for most people to read. The list that follows is reasonably representative of literature in the English language. Most of the books mentioned give numerous references to other works.

History: K. BIRR, *Pioneering in Industrial Research: The Story of the General Electric Research Laboratory* (1958); W.R. MacLAURIN and R.J. HARMAN, *Invention and Innovation in the Radio Industry* (1949), a detailed account of the technical problems solved in one area; H. MELVILLE, *The Department of Scientific and Industrial Research* (1962), a history of a government research department; OECD DIRECTORATE FOR SCIENTIFIC AFFAIRS, *International Statistical Year for Research and Development: The Overall Level and Structure of R and D Efforts in* OECD *Member Countries* (1963–64).

Organization: A.O. STANLEY and K.M. WHITE, *Organizing the R & D Function* (1965), descriptions of the organizational structures of leading U.S. companies; I.D.L. BALL (ed.), *Industrial Research in Britain* (1968), details and statistics of U.K. industry; J. COCKCROFT (ed.), *The Organization of Research Establishments* (1965), essays by directors of 15 leading U.K. and U.S. laboratories; G.S. MONTEITH, *R. and D. Administration* (1969), a detailed description of administrative procedures in the U.K. and U.S., with an extensive bibliography.

Government: Report of the Committee on the Management and Control of Research and Development (HMSO 1961); *Basic Research and National Goals* (1965), a report of the National Academy of Sciences.

General: D. ALLISON (ed.), *The R and D Game* (1969), essays by 15 authors covering the whole field from creativity to administration, with an extensive bibliography; E.D. REEVES, *Management of Industrial Research* (1967), a discussion of the integration of R and D into corporate structure; T.S. McLEOD, *Management of Research, Development and Design in Industry* (1969), basic problems of technical management; C. HEYER (ed.), *Handbook of Industrial Research Management*, 2nd ed. (1968), essays by American managers on R and D problems as seen from the Board Room level; E. MORISON, *Men, Machines and Modern Times* (1966), an account of resistance encountered by new inventions.

Project management: K.G. LOCKYER, *An Introduction to Critical Path Analysis*, 2nd ed. (1967); J.J. MODER and C.R. PHILLIPS, *Project Management with CPM and PERT* (1964).

(T.S.McL.)

Resnais, Alain

Resnais.
By courtesy of the French Film Office, New York

With *Hiroshima mon amour*, his first feature-length motion picture, Alain Resnais attained pre-eminence among the innovative New Wave directors (who appeared in France in the late 1950s) particularly for the finesse with which he combined a traditional form and a radical content. In his preference for elaborate rehearsals of his actors, Resnais belongs to the classical cinema; but his social, political, and spiritual milieu is that of the "Left Bank" school of film makers, so called for their political philosophy as well as for a formidable intellectuality related to the cosmopolitan bohemianism of the Saint-Germain-des-PrCs district, on the Left Bank of the Seine, in Paris. The orientation of this group is opposite to that of the *Cahiers du Cinéma* subsection of the New Wave, which tended toward a politically quietist anarchism and drew upon a deliberately conventional, often Catholic, bourgeois culture—and whose editorial offices were in the fashionable Champs-ElysCes across the Seine. Though less well publicized and far less prolific, the "Left Bank" group (of whom Resnais was the spearhead) anticipated the political upheavals of Paris in 1968 and has dominated French cinema culture from the late 1960s.

Alain Resnais was born in Vannes, Brittany, on June 3, 1922, the son of a well-to-do pharmacist. A victim of chronic asthma, the young Resnais became solitary and intensely interested in creative activity, characteristics that would remain salient through his adulthood. While still a boy, he was given a movie camera, and at the age of 14 he directed his classmates in a film version of a popular thriller, *Fantômas.*

His illness exempted him from military service in World War II; and in 1940 he went to Paris, where he studied cinema at the Institut des Hautes Études Cinématographiques. During the German occupation of France, he became interested in theatre; later he would reproach himself for having become too immersed in it to join the underground resistance movement, but his brief stage career helped develop his sensitivity to actors and his technique of rehearsing them for his films.

Despite his interest in theatre, films remained his first love (along with comic strips, which he considers a kindred medium); and in 1947 he initiated a series of short films devoted to the visual arts with *Chateaux de France*, which he made by cycling and camping through the country. Having little interest in the French commercial film industry of the time, he continued making shorts—on van Gogh, Gauguin, and Picasso's painting "Guernica," among others—for the next nine years. Even in such documentary-like works, Resnais's profound vision of man's ominous alienation from his own humanity began to be expressed. He received commissions for political and propaganda films, whose immediate purpose he fulfilled but also transcended artistically. Thus, his documentary about concentration camps, *Nuit et brouilfard* ("Night and Fog"), with a commentary by a former inmate, the contemporary poet Jean Cayrol, stressed "the concentrationary beast slumbering within us all." *Le Chant du styrène*, written by author and critic Raymond Queneau, nominally publicizing the versatility of the plastic polystyrene, became a meditation on the transformation of matter from amorphous nature into bright, banal household implements.

Early documentaries

The postponement of popular success in his career permitted his art to mature all the more intensively. The solitude he experienced in childhood reappeared thematically in his sensitivity to the evanescence of experience, to the passing of time, and to the discrepancies between individual consciousnesses—themes that inspired comparisons to the philosophy of Henri Bergson and to the novels of Marcel Proust. His youthful fascination with comic-strip fantasies matured into a realization of the quietly monstrous forces within man and society. In his films Resnais shows man at his most sensitive, confronting his own devious barbarism—in the form of the atom bomb in *Hiroshima mon amour*, of a sumptuous but chilling dreamworld in *L'Année dernière à Marienbad (Last Year at Marienbad)*, of police torture in *Muriel*. He repeatedly presents human relationships that are .characterized by reticence, modesty, immaculate courtesy, and a stimulating respect for others, together with overtones of solitude. Resnais has regularly worked with such distinguished French literary figures as Marguerite Duras and Alain Robbe-Grillet, encouraging them to write the script as a piece of literature rather than as a screenplay. He then

Collaboration with literary figures

transposes their vision into cinematic terms, in a style richly impregnated with his own sensibility.

A bachelor, Resnais lives quietly in Paris. His close friends include many lesser known actors and technicians he has worked with. His films epitomize the unusual blend of circumspection and commitment in his own personality. Although he has dealt regularly with problems of personal and political action, his radical commitment is often underestimated by critics who are mesmerized by his immaculate style. His short films have had several brushes with government censorship. *Les Statues rneurent aussi* ("Statues Also Die"), his study of African art, was banned for 12 years for references to colonialism that he refused to alter. Some critics condemned *Hiroshima mon amour* for its sympathetic treatment of the heroine, once a wartime collaborationist and now an interracial adulteress who advocates internationalism and the "New Morality." Even when Resnais deals explicitly with political figures, however, as in *La Guerre est finie* ("The War Is Over"), his scrupulosity and tragic humanism are so much in evidence that his work transcends partisan feelings.

MAJOR WORKS

Van Gogh (1948); Gauguin (1950); Guernica (1950); Les Statues meurent aussi (1953); Nuit et brouillard (1955); Toute la mémoire du monde (1956); Le Chant du styrène (1958); Hiroshima mon amour (1959); L'Année dernihre à Marienbad (1961; Last Year at Marienbad); Muriel (1963); La Guerre est finie (1966); Loin de Vietnam (1967; Far from Vietnam), an episode; *Je t'aime, je t'aime (1968).*

BIBLIOGRAPHY. ROY ARMES, *The Cinema of Alain Resnais* (1968), a useful general survey; JOHN WARD, *Alain Resnais; or, the Theme of Time* (1968), a thoughtful study, preoccupied with Resnais as Bergsonian. *(Scripts): Hiroshima mon anzour (1960;* Eng. trans., *1966); L'Année dernihre à Marienbad (1961); La Guerre est finie (1966).*

(R.Du.)

Respiration, Disorders of

The prime function of the respiratory system is the exchange of gas between the blood and the external atmosphere—the removal of carbon dioxide (CO_2) from the blood and the enrichment of the blood with oxygen. The functioning units for this purpose are the alveoli in the lungs, minute air sacs whose membranous walls are filled with a network of microscopic blood vessels, the capillaries; through the alveolar walls and thin walls of the capillaries the gases are diffused. The rest of the system provides passageways for ventilation, the removal of air laden with carbon dioxide from the lungs and its replacement with fresh air. (See also RESPIRATORY SYSTEM, HUMAN; RESPIRATION, HUMAN.)

Oxygen and the lung structure

Each of the basic disorders of respiration is a disturbance of one aspect of the normal functioning of the lung, or arises from interference with one of the physiological processes concerned with delivery of oxygen to the tissues or removal of carbon dioxide from them. Oxygen, on which man's metabolic processes entirely depend, is, by comparison with CO_2, a relatively insoluble gas, so that it passes through membranes or water barriers considerably less quickly than carbon dioxide. Clearly, it is this relative insolubility of oxygen that accounts for some of the aspects of lung structure in man. The internal surface area of the human lung is about 750 square feet, this size being achieved within the chest cage by the extreme smallness of the alveoli of the lung. This smallness in turn would not be feasible except for the presence of a surface-active material that lines the alveoli to reduce the high tension which would otherwise exist at the gas-liquid interface of such a small volume. The tissues separating gas and blood are exceedingly fine, and the blood flowing through the lung flows through very small channels. Not more than about 100 cubic centimetres (cc) of blood—about half a cupful—is in the pulmonary capillaries at any one time, although the flow through the lung at rest is approximately five litres per minute.

BASIC DISORDERS OF RESPIRATION

Interference with ventilation. Obviously gross interference with ventilation as in asphyxiation and drowning prevents effective gas exchange. Respiration may be halted also by severe electric shock or by sudden brain hemorrhage. In all these instances recovery may occur if ventilation can be quickly reestablished. The quick clearance of any material that may be in the airway of the affected person and the prompt application of mouth-to-mouth resuscitation are techniques that may save lives. Older methods of artificial respiration were relatively ineffective, and the direct method of mouth-to-mouth breathing has largely supplanted all other first-aid techniques. In this method, the nose of the subject is held, and his lungs are inflated by expiration of the person giving the treatment. Approximately ten inflations a minute are made, and the lungs are allowed to empty spontaneously. There is little danger of overinflation of the lung, and effective ventilation can be maintained for a considerable period by this method. A small rubber bag and valve system with a mouthpiece, apparatus commonly found wherever there is danger of asphyxiation or drowning, provides equally effective ventilation. Police and first-aid workers are normally trained in use of the apparatus. If the victim's heart has not completely stopped at the time ventilation is begun by this method, there is good hope that resuscitation may be achieved. In drowning, there is more hazard from fresh than from salt water, because very considerable dilution of the blood occurs if much fresh water is aspirated into the lung, whereas the salts in seawater prevent this blood dilution. There is thus a better chance of resuscitation after apparent drowning in salt water than in fresh.

Paralysis or artificial relaxation of respiratory muscles

Since the body contains practically no stored oxygen, the brain can withstand deprivation of oxygen for only a few minutes. Interference with normal ventilation may occur as a result of disease of the respiratory muscles, of which acute poliomyelitis is a severe example. Other diseases that cause chronic loss of muscle power may progress to involve inadequate ventilation and respiratory failure. Respiration may be depressed by the administration of such drugs as morphine or by the accidental ingestion of poisons of many kinds. In most instances the drug acts on the respiratory centre, causing a severe depression of its activity and consequent failure to maintain normal ventilation. Curare and similar compounds are used in anesthesia to cause total muscle relaxation and, when this is done, ventilation is maintained artificially during the surgical procedure. In severe cases of asthma, the obstruction of airflow may be so severe that the total ventilation of the victim is endangered; in these circumstances acute respiratory failure may occur and may be dangerous to life. In all of these instances, mechanisms have been brought into play that have prevented effective ventilation of the lungs.

Oxygen pressures at high altitudes

Reduction of oxygen pressure. The transport of gas across the lungs depends on the process of diffusion, the rate of transfer of oxygen from gas to blood across the alveolar membrane of the lung being dependent on the difference in partial pressure that exists on the two sides. The partial pressure of oxygen in the alveoli is maintained by the process of ventilation, but the active process of oxygen uptake from the alveoli means that the average alveolar oxygen pressure is bound to be less than that in outside air. At sea level the oxygen pressure, expressed in millimetres (mm) of mercury (Hg), is about 150 (approximately one-fifth of the barometric pressure of 760 mm mercury, since oxygen is about one-fifth of the atmosphere); a normal man, breathing at sea level, has an average alveolar oxygen pressure of about 100 mm mercury. As man goes higher above sea level, the barometric pressure drops continuously; when he has reached 18,000 feet (about 5,500 metres [m]) it is only about half of that at sea level. Consequently, ascent to any altitude means that the oxygen pressure in the environment is lower than it is at sea level. A population living at an altitude of 6,000 feet (about 1,830 m), in Johannesburg, South Africa, or in Denver, Colorado, is bound to have a slightly lower arterial oxygen pressure at rest than does the sea level population, and it will be found to have, in addition, a slightly lower carbon dioxide pressure. This is because the subject at altitude is stimulated by the low oxygen pressure to ventilate rather more than the dweller at sea level; and, since the carbon dioxide in the blood is a direct reflection of alveolar ventilation, it consequently is lower in those living at high altitude. The processes of adjustment to altitude are complex, but they include the following important physiological mechanisms:

1. The oxygen content of blood is not much reduced when the pressure to which it is exposed falls from 100 mm mercury to 70 mm mercury, so that the quantity of oxygen carried by the blood is relatively well maintained down to the lower pressure. This relationship is of material assistance to man living at altitudes up to 10,000 feet (about 3,050 m). The curve relating oxygen content to oxygen pressure falls steeply after a pressure of approximately 70 mm Hg is reached.

2. The arterial oxygen tension of blood is continuously measured and controlled by chemoreceptors, small bodies situated at the division of the carotid artery in the neck, which stimulate ventilation when a fall in oxygen pressure in arterial blood occurs.

3. During the first three weeks at higher altitudes, a man becomes progressively more sensitive to carbon dioxide. The exact nature of this process is not fully understood, but it is known to depend on changes in the concentration of hydrogen ion occurring between the cerebrospinal fluid and the arterial blood. Although the carbon dioxide tension in cerebrospinal fluid is bound to be the same as that of blood (because equilibrium is readily achieved with a gas as soluble as CO_2), there is an active process of transport of bicarbonate ion (HCO_3) from cerebrospinal fluid, so that a hydrogen ion difference is established between cerebrospinal fluid and blood (*i.e.*, cerebrospinal fluid becomes more acid than blood) as the heavy breathing and hyperventilation of acute exposure to altitude occurs. The result is that centrally placed chemoreceptors in the medulla of the brain trigger a more sensitive setting of the respiratory control system to carbon dioxide.

4. Long-term residence at high altitude occasions changes in the blood that enable more oxygen to be carried. This process occurs slowly, but with long-term residence at altitudes of 15,000 and 16,000 feet (about 4,570 and 4,875 m) represents an important part of adaptation to the environment. Residents at high altitude, and more particularly those who have been born there, are believed to develop or to have richer capillaries in muscle tissue to facilitate oxygen exchange between blood and tissue cells.

Acute altitude sickness

These adaptive processes may break down in a number of ways. The unacclimatized person who is abruptly moved to 16,000 feet or so and made to do physical work suffers from acute altitude sickness with headache, nausea and vomiting, inability to sleep, and extreme fatigue. These acute effects result from hypoxia (deficiency of oxygen reaching the tissues of the body), but the exact mechanism is not understood in detail. In some instances, fluid will accumulate in the lungs of unacclimatized persons working at high altitudes. This dangerous complication was encountered on a large scale in the 1960s when Indian troops were flown from sea level to altitudes of about 16,000 feet to repel the invasion of India by China. Acclimatization is important for mountaineers and consists essentially of three weeks of graded exercise to progressively higher altitudes, permitting the regulatory processes of the body, particularly the chemoreceptors, to adjust the ventilation to the correct level. The kidney also requires time to adjust the bicarbonate level of blood to the lowering of CO, tension within it, so that the acidity-alkalinity relation of arterial blood may be maintained constant.

The disease of chronic mountain sickness (Monge's disease) occurring in residents at high altitude is characterized by severe hypoxia and secondary congestive heart failure. The increase in red blood cells (polycythemia) may be extreme. Chronic mountain sickness appears to result from failure of the stimulus of oxygen lack to have sufficient effect on ventilation. Prompt relief occurs when the affected person is brought to a lower altitude, but the disease is likely to recur if he returns to high altitude. The reason for the breakdown in ventilatory stimulus is not well understood.

Other circumstances in which oxygen intake may be inadequate include being trapped in a space from which oxygen has been removed by combustion or in which oxygen is deficient so that victims may suffer acute hypoxia before they are aware of the danger.

Inadequate oxygenation of the blood. Arterial tension may be depressed even though the environmental oxygen tension is normal, as when ventilation is grossly reduced by some interference with breathing, as noted above, or when there is severe lung dysfunction. Most commonly, although ventilation is maintained, the mismatch between the blood perfusion of the lung and the ventilation of it results in large volumes of blood flowing through parts of the lung that are very poorly ventilated. A variety of clinical conditions give rise to this not uncommon condition. Prompt treatment with enriched oxygen mixtures usually restores the arterial oxygen tension to normal without difficulty.

Another circumstance in which there is deficient oxygenation of blood, although the inspired oxygen tension is normal, is when fluid accumulates in the lungs. In this condition, callea pulmonary edema, leakage of fluid from the pulmonary capillaries into pulmonary tissue causes diminishing ventilation of the affected part of the lung and may lead to considerable hypoxia. Pulmonary edema occurs either when the pulmonary capillary bed has been damaged or when the lungs are subjected to an acute elevation of blood pressure, a condition most commonly occurring when function of the left ventricle of the heart is failing.

Quantity of oxygen. The carriage of oxygen by the blood is primarily dependent on the amount of hemoglobin in it. When the hemoglobin is grossly reduced, as in anemia, the amount of oxygen that may be carried by any given amount of blood is proportionately reduced. Blood travelling through the lung, however, will still equilibrate completely with the alveolar gas, and so the arterial oxygen tension is normal. In these circumstances of anemia, the reduced amount of oxygen that each given amount of blood is carrying leads to a compensatory increase in the output of the heart to permit maintenance of a normal delivery of oxygen per minute to the tissues.

Effects of anemia

Carbon monoxide poisoning

Carbon monoxide poisoning has unique physiological features that make this gas extremely dangerous to those who do not know they may be exposed to it. The odourless gas, carbon monoxide (CO), commonly produced by incomplete combustion of any hydrocarbon, is a major part of automobile exhaust. Hemoglobin has several hundred times the affinity for CO that it does for oxygen, which means that when CO is breathed, the hemoglobin travelling through the lung will combine with CO in preference to oxygen. Thus when CO is bound in this way to hemoglobin (Hb), the hemoglobin cannot pick up oxygen. In addition, the presence of carbon monoxide hemoglobin (COHb) in the blood affects the release of oxygen by the remaining unaffected hemoglobin in such a way that the oxygen tension within working'tissues becomes grossly reduced, although the arterial oxygen tension has remained normal. This shift in the release of oxygen by hemoglobin is a major reason why a saturation of half the hemoglobin with carbon monoxide (COHb percentage = 50 percent) causes severe and dangerous incapacity in an individual, whereas removal of half the hemoglobin (hemoglobin = 50 percent normal) as in anemia causes much less severe effects. CO is additionally dangerous because it does not stimulate ventilation and thus gives little warning of its presence. The cherry pink colour of carbon monoxide hemoglobin causes the victim to appear to be well oxygenated, and may mislead observers as to the dangerously low level of tissue oxygenation. The effects of carbon monoxide are insidious. If men are unaware that they are being progressively poisoned by this gas, they may only realize the hazard of their situation when too weak to take action to correct it.

SPECIAL SITUATIONS AFFECTING RESPIRATION

Respiration during swimming and diving. There may be considerable danger to a swimmer if he greatly increases his ventilation just before he jumps into the water. Hyperventilation at rest will reduce the carbon dioxide tension of the blood, but because of the shape of the oxygen dissociation curve, it does not increase significantly the amount of oxygen in the body. If the swimmer then jumps into the water and swims hard with his face under water, he will not be stimulated to take a breath until the oxygen tension in his blood has reached such low levels that he may well lose consciousness. Hyperventilation before he jumps thus postpones the moment at which a rising CO_2 tension will force him to breathe, but danger arises because his arterial oxygen pressure may have fallen to very low levels.

Scuba diving, with apparatus supplying gas to the diver at a pressure equal to the pressure exerted by the water on his chest wall, permits underwater diving to depths of up to 100 feet (30.5 m). Thus the diver can breathe without difficulty. An important hazard of scuba diving, however, is that a rupture of lung alveoli may occur during a rapid ascent from depth. If the diver's airway is closed during ascent, the diminishing pressure on the chest leads to rapid gas expansion with possible rupture of alveoli and the occurrence of air embolism, the presence of air bubbles in the bloodstream. It is important that subjects be trained to ascend with the glottis open so that the airway through the larynx is open.

In diving in which the diver is enclosed in an airtight costume including a brass headpiece, and air is pumped down to him from the surface, since the diver is breathing at the pressure of the water in which he is working, the pressure of nitrogen and oxygen in the air he is breathing will be greatly increased over its pressure at sea level.

Problems of breathing air

Two problems thus arise. First, oxygen at increased pressure has toxic effects. Although pure oxygen at sea-level pressure can be breathed for two hours safely, even at this pressure if it is breathed for long it causes changes in the lung and has secondary effects in the liver and brain. With the increased pressure under water, the acute elevation of oxygen pressure to several thousand millimetres of mercury may cause acute oxygen poisoning that leads to muscular twitching and, if unchecked, to dangerous convulsions. Thus the oxygen pressure when diving at depth has to be most carefully controlled. For example, in a simulated dive in a pressure chamber to a pressure equivalent to a depth of 1,000 feet under water, the inspired oxygen concentration was reduced from the normal 20 percent to 0.9 percent to attain the desired oxygen partial pressure of 150 mm Hg. The second problem of breathing air when under water is that the pressure of nitrogen is increased, thus increasing nitrogen tension through the body, swiftly in those regions with a large blood supply and slowly in organs with a poor blood supply. The diver may descend to considerable depths as fast as he wishes; but, having remained at such depths, he is unable to ascend quickly because a too rapid ascent from depth causes nitrogen to bubble out of the blood and bring on the syndrome known as "bends." In mild form "bends" may amount to no more than transient joint pains and some skin itching, but severe "bends" cause gas embolization and death. Because of this hazard, divers breathing air and working at considerable depths have to ascend gradually; after working for two hours at a depth of 300 feet (about 90 m), a diver may have to take five or six hours to ascend safely to the surface.

Advantages of helium-oxygen mixture

Because of the requirement of slow ascent, divers going to great depths use helium and oxygen mixtures. Helium is much less soluble in body tissues than nitrogen and has the added advantage of being easier to breathe at depth than nitrogen. The economics of men working at a depth of 300 feet, in terms of their working time as against their decompression time, have led to the development of a laboratory on the sea floor ("sealab") in which men may live for up to three weeks. Once saturation has occurred with gas at the higher pressure, the decompression time is the same if one has spent three weeks at depth as if one had spent only eight or nine hours at depth; thus it is economic to keep men at depth for much longer periods under these conditions. Divers leave the "sealab" for an hour or so to work and return for rest and food.

Abnormalities of rhythm. Under normal resting conditions, the volume of air taken in with each breath is approximately constant and the respiratory rhythm is regular. Where there is disease of blood vessels in the brain, enlargement of the heart, and a considerable slowing of the circulation, the breathing may oscillate with periods of increased tidal volume alternating with periods of very shallow breathing. This condition, known as Cheyne-Stokes respiration, is thought to result from the oscillation of the feedback circuitry that controls respiration. In Biot's respiration, the breathing frequency and tidal volume are irregularly irregular as a consequence of depression of the respiratory centre in the midbrain. This condition results from brain damage and also occasionally from overdose of a sedative drug.

Respiratory problems in space, A major problem posed by space exploration is that of withstanding the acceleration necessary to leave the Earth's atmosphere and the deceleration caused by reentry into the Earth's atmosphere. Under conditions of severe acceleration, the blood flow through the lung becomes inevitably restricted to the most dependent part of the lung in the gravitational field; and as this part of the lung may become collapsed, there may be a serious disorder of gas exchange. Some degree of minor atelectasis (collapse of small lung unit) probably commonly occurs in these circumstances but is quickly rectified by the taking of a few deep breaths once the accelerative force is over. The stage of weightlessness (zero gravitation) does not seem to cause problems for the normal function of the lung, and indeed the distribution of ventilation and perfusion in the resting situation within the lung may be slightly more uniform under these conditions than in the presence of the normal gravitational pull of the Earth, It is not known whether readjustment to Earth's gravitational field after a prolonged period of weightlessness can be easily achieved, but the reverse adaptation from the Earth's gravitational field to the state of weightlessness does not seem to cause difficulty. Under conditions of considerable acceleration (three times gravity or 3g) the ability of a man to perform physical work may be affected by the limitations imposed on gas exchange across the lung.

BIBLIOGRAPHY. W.O. FENN and H. RAHN (eds.), "Respiration," in ***Handbook of Physiology,*** sect. 3 (1965), a comprehensive and detailed analysis of every aspect of respiratory physiology; J.H. COMROE, ***Physiology of Respiration: An Introductory Text*** (1965), an elementary account of the principles of lung function; V.B. MOUNTCASTLE (ed.), ***Medical Physiology, 2*** vol., 12th ed. (1968), a general text that emphasizes respiratory problems of diving.

(D.V.B.)

Respiration, Human

Respiration is the process by which sufficient oxygen for their needs is delivered to body cells and most of the carbon dioxide that these cells form is eliminated into the atmosphere. This process is carried out by (1) a physiologic pump that moves oxygen-rich air and carbon dioxide-laden air between the atmosphere and the surface for gas exchange with blood; (2) a transport system between this surface and the body cells; and (3) a control system matching pumping and transport to the requirements of the cells. The pumping of air is done by the chest, abdomen, and lungs. The surface for exchange of oxygen and carbon dioxide consists of the walls of the alveoli—minute air sacs in the lungs—and of the walls of the tiny blood vessels in the alveolar walls. Transport of oxygen and carbon dioxide is carried out by the blood as it is pumped by the heart through the blood vessels. Control is maintained by the nervous system, which activates muscles in response to impulses from receptors sensitive to cellular needs. Thus, many structures and organs besides those of the respiratory system are involved in respiration: the lungs, the bronchi, the trachea, the throat, and the nose.

While a presentation of respiration divides itself logically into a consideration of each part of the process individually, all organs and structures work interdependently in creating an optimal oxygen and carbon dioxide environment for the cell. This article is intended to deal primarily with the processes of breathing and the transport of oxygen and carbon dioxide. Cellular respiration is dealt with in the article METABOLISM, breathing of organisms other than man in RESPIRATION AND RESPIRATORY SYSTEMS.

Mechanics of breathing. Air moves in and out of the lungs in response to differences in pressure. When the air pressure within the alveolar spaces falls below atmospheric pressure, air enters the lungs (inspiration), provided the larynx is open; when the air pressure within the alveoli exceeds atmospheric pressure, air is blown from the lungs (expiration). The flow of air is rapid or slow in proportion to the magnitude of the pressure differences. Because atmospheric pressure remains relatively constant, flow is determined by how much above or below atmospheric pressure the pressure within the lungs rises or falls.

Alveolar pressure fluctuations are caused by expansion and contraction of the lungs resulting from tensing and relaxing of the muscles of the chest and abdomen. Each small increment of expansion transiently increases the space enclosing lung air. There is, therefore, less air per unit volume within the lungs and pressure falls. A difference in air pressure between atmosphere and lungs is created, and air flows in until equilibrium with atmospheric pressure is restored at a higher lung volume. When the muscles of inspiration relax, the volume of chest and lungs decreases, lung air becomes transiently compressed, its pressure rises above atmospheric pressure, and flow into the atmosphere results until pressure equilibrium is reached at the original lung volume. This, then, is the sequence of events during each normal respiratory cycle: lung volume change, leading to pressure difference, resulting in flow of air into or out of the lung, and establishment of a new lung volume.

Elastic properties of lung

The lung–chest system. The forces that normally cause changes in volume of the chest and lungs stem not only from muscle contraction but from the elastic properties of both the lung and the chest. A lung is similar to a balloon in that it resists stretch, tending to collapse almost totally unless held inflated by a pressure difference between its inside and outside. This tendency of the lung to collapse or pull away from the chest is measurable by carefully placing a blunt needle between the outside of the lung and the inside of the chest wall, thereby allowing the lung to separate from the chest at this particular spot. The pressure measured in the small pleural space so created is substantially below atmospheric pressure at a time when the pressure within the lung itself equals atmospheric pressure. This negative (below atmospheric) pressure is a measure, therefore, of the force required to keep the lung distended. The force increases (pleural pressure becomes more negative) as the lung is stretched and its volume increases during inspiration. The force also increases in proportion to the rapidity with which air is drawn into the lung and decreases in proportion to the force with which air is expelled from the lungs. In summary, the pleural pressure reflects primarily two forces: (1) the force required to keep the lung inflated against its elastic recoil and (2) the force required to cause airflow in and out of the lung. Because the pleural pressure is below atmospheric pressure, air is sucked into the chest and the lung collapses (pneumothorax) when the chest wall is perforated, as by a wound or by a surgical incision.

The force required to maintain inflation of the lung and to cause airflow is provided by the chest and diaphragm (the muscular partition between chest and abdomen), which are in turn stretched inward by the pull of the lungs. The lung-chest system thus acts as two opposed coiled springs, the length of each of which is affected by the other. Were it not for the outward traction of the chest on the lungs, these would collapse; and were it not for the inward traction of the lungs on the chest and diaphragm, the chest would expand to a larger size and the diaphragm would fall from its dome-shaped position within the chest.

The role of muscles. The role of respiratory muscles is to displace the equilibrium of elastic forces in lung and chest in one direction or the other by adding muscular contraction. During inspiration, muscle contraction is added to the outward elastic force of the chest to increase the traction on the lung required for its additional stretch. When these muscles relax, the additional retraction of lung returns the system to its equilibrium position.

Contraction of the abdominal muscles displaces the equilibrium in the opposite direction by adding increased abdominal pressure to the retraction of lungs, thereby further raising the diaphragm and causing forceful expiration. This additional muscular force is removed on relaxation and the original lung volume is restored. During ordinary breathing, muscular contraction occurs only on inspiration, expiration being accomplished "passively" by elastic recoil of the lung.

At total relaxation of the muscles of inspiration and expiration, the lung is distended to a volume—called the functional residual capacity—of about 40 percent of its maximum volume at the end of full inspiration. Further reduction of the lung volume results from maximal contraction of the expiratory muscles of chest and abdomen. The volume in these circumstances is known as the residual volume; it is about 20 percent of the volume at the end of full inspiration (known as the total lung capacity). Additional collapse of the lung to its "minimal air" can be accomplished only by opening the chest wall and creating a pneumothorax.

The membrane of the surface of the lung (visceral pleura) and on the inside of the chest (parietal pleura) are normally kept in close proximity (despite the pull of lung and chest in opposite directions) by surface tension of the thin layer of fluid covering these surfaces. The strength of this bond can be appreciated by the attempt to pull apart two smooth surfaces, such as pieces of glass, separated by a film of water.

The respiratory pump and its performance. The energy expended on breathing is used primarily in stretching the lung–chest system and thus causing airflow. It normally amounts to 1 percent of the basal energy requirements of the body but rises substantially during exercise or illness. The respiratory pump is versatile, capable of increasing its output 25 times, from a normal resting level of about six litres (366 cubic inches) per minute to 150 litres (9,150 cubic inches) per minute in adults. Pressures within the lungs can be raised to 130 centimetres of water (about 1.8 pounds per square inch) by the so-called Valsalva manoeuvre—*i.e.,* a forceful contraction of the chest and abdominal muscles against a closed glottis (*i.e.,* with no space between the vocal cords). Airflow velocity, normally reaching 30 litres (1,830 cubic inches) per minute in quiet breathing, can be raised voluntarily' to 400 litres (24,400 cubic inches) per minute. Cough is accomplished by suddenly opening the larynx during a brief Valsalva manoeuvre. The resultant high-speed jet of air is an effective means of clearing the airways of excessive secretions or foreign particles. The beating of cilia (hairline projections) from cells lining the airways normally maintains a steady flow of secretions toward the nose, cough resulting only when this action cannot keep pace with the rate at which secretions are produced.

An infant takes 33 breaths per minute with a tidal volume (the amount of air breathed in and out in one cycle) of 15 millilitres, totalling about 0.5 litre—approximately one pint—per minute as compared to adult values of 14 breaths, 500 millilitres, and seven litres (14 pints), respectively.

If the force of surface tension is responsible for the adherence of parietal and visceral pleurae, it is reasonable to question what keeps the lungs' alveolar walls (also fluid-covered) from sticking together and thus eliminating alveolar air spaces. In fact, such adherence occasionally does occur and is one of the dreaded complications of premature births. Normal lungs, however, contain a substance—a phosphalipid surfactant—that reduces surface tension and keeps alveolar walls separated.

Composition of lung air and arterial blood. Lung, or pulmonary, air contains less oxygen, more carbon dioxide, and more water vapour than atmospheric air. Oxygen is lost to the blood of the pulmonary capillaries (the smallest blood vessels), carbon dioxide is added from venous blood entering the pulmonary capillaries, and the water vapour content is raised by virtue of saturation at body temperature (37" C, 98.6" F). The Table lists representative values for room air, tracheal air at the end of inspiration, and alveolar air in terms of both concentration and partial pressure.

Concentrations and Partial Pressures of Respiratory Gases at Inspiration

key: F—concentration; P—partial pressure (mm Hg)

	dry room		trachea		alveoli	
	F	P	F	P	F	P
Nitrogen (N_2)	.7900	600	.7408	563	.7408	563
Oxygen (O_2)	.2096	159	.1960	149	.1447	110
Carbon dioxide (CO_2)	.0004	1	.0004	1	.0527	40
Water vapour (H_2O)	.0	0	,0618	47	.0618	47
	1.000	760	1.000	760	1.000	760

Composition of air in alveoli

The alveoli. Of the four gases in alveolar air, two (water vapour and nitrogen) have fixed concentrations, whereas the concentration of the other two (oxygen and carbon dioxide) is variable within limits. When the volume of air entering and leaving the alveoli per minute (ventilation) is increased, carbon dioxide is "blown off" from the lungs and body and its partial pressure in the alveoli falls; when ventilation is decreased, the alveolar partial pressure of carbon dioxide rises. Any such change in partial pressure of carbon dioxide is attended by an opposite variation in the concentration of oxygen. This occurs because the total pressure of the air remains relatively constant, as does that of two of its components, water vapour and nitrogen; any change in concentration of one of the two remaining components must be accompanied by an opposite change in the other.

The management of the pressure of oxygen in blood and in tissue fluid hinges, therefore, on eliminating carbon dioxide by ventilation of the alveoli. One of the remarkable attributes of the respiratory system is the delicacy with which alveolar ventilation is modulated to meet varying rates of carbon dioxide production and of oxygen utilization by the body, as during exercise or fever. When alveolar ventilation does fail, carbon dioxide retention (hypercapnia) and a deficit of oxygen (hypoxia) ensue. In such an event, the pressure of oxygen in alveoli and blood can be maintained only by adding oxygen to inspired air.

The airways and "dead space." Air in the airways (bronchi, trachea, throat, and nose) at the end of inspiration differs from room air only in that it is saturated with water vapour at body temperature. At the end of expiration, on the other hand, the airways are filled with alveolar air in equilibrium with the gases in the pulmonary capillary blood.

The volume of the airways, approximately 150 millilitres (about 0.3 pint) in adults, is referred to as dead-space volume because no exchange of gases with blood occurs in the airways. To the extent that alveoli also fail to exchange gases with blood effectively, as when pulmonary capillaries are absent, they add to the airway, or anatomic dead space, an increment referred to as physiologic dead space. The sum of the anatomic and physiologic dead spaces is normally not more than one-third of a normal breath (one tidal volume).

Inasmuch as total ventilation is the sum of alveolar ventilation plus dead-space ventilation, ventilatory effort for a given alveolar ventilation is reduced or augmented as dead space is reduced or increased. Economy of respiratory effort requires, therefore, that both anatomic and physiologic dead spaces be as low as possible. One of the effects of reducing the blood flow through capillaries that are nevertheless ventilated, as in certain diseases of the lung, is to increase physiologic dead space. In any case, expired air is a mixture originating in parts of the lung that are imperfused (anatomic dead space) and perfused to varying degrees.

Anatomic and physiologic bypasses

The arterial blood. Just as air leaving the lungs is a mixture, so also is the blood returning to the heart from the lungs. This is true for two reasons: (1) Some blood bypasses the gas-exchanging surface entirely (anatomic shunt). (2) Some alveoli are ventilated less than is required in order to maintain an optimal concentration or partial pressure of oxygen (physiologic shunt); the blood returning from such alveoli is, therefore, more like venous blood. The mixture of circulating arterial blood accordingly reflects the magnitude both of the anatomic shunts (increased by some forms of congenital heart disease) and of the physiologic shunts (increased by some lung diseases), as well as reflecting the general level of alveolar ventilation.

In summary, both expired air and arterial blood are mixtures. The mix of expired air reflects in part the nonuniformity of the perfusion of blood through the alveoli; the amount of oxygen in the arterial blood is a measure of the nonuniformity of alveolar ventilation.

Diffusion of oxygen and carbon dioxide between air and blood. Gases are exchanged between alveolar air and pulmonary capillary blood by diffusion in response to differences in partial pressure (the pressure of one of the components of a mixture of gases). Since the partial pressure of oxygen in alveolar air is higher than that of the venous blood entering the lung, oxygen enters the fluid and red blood cells in the capillaries until its pressure there virtually equals its pressure in the alveoli. And since carbon dioxide has a higher partial pressure in venous blood than in alveolar air, it diffuses from the capillaries into the alveoli until there is an equilibrium of pressures. Because the rate of diffusion between a liquid and a gas depends in part on the solubility of the gas in the liquid and, further, because carbon dioxide is about 20 times more soluble than oxygen in lung fluid, equilibrium in the exchange of carbon dioxide is reached more rapidly than for oxygen.

A given red blood cell normally remains within a pulmonary capillary approximately 0.7 second, substantially longer than the time required for oxygen-exchange equilibrium in normal circumstances. When some abnormality in the capillary wall or the red cells slows down the passage of oxygen, a deficit of oxygen in the blood results unless the alveolar partial pressure of oxygen is raised either by hyperventilation with air low in carbon dioxide or by breathing an oxygen-enriched mixture.

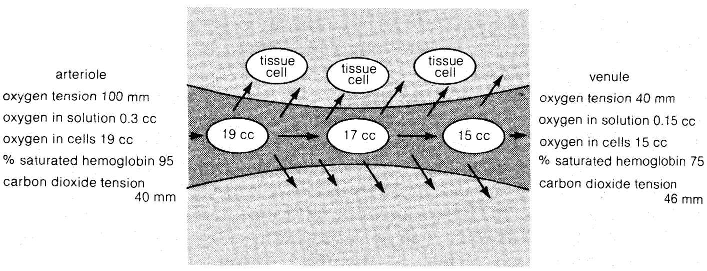
The effectiveness of the lung in carrying out diffusion (pulmonary diffusing capacity) can be measured in terms of how much oxygen or some other gas, usually carbon monoxide, diffuses from alveoli to capillaries per unit difference in the partial pressures at those two sites. Diffusing capacity increases from infancy to adulthood and with exercise, suggesting an increase in surface area of the alveolar capillary bed. Diffusing capacity is lowered when the number of alveoli and capillaries has been reduced or when the structures between alveoli and capillary channels are thickened.

Gas transport. Oxygen and carbon dioxide are transported between tissue cells and the lungs by the blood. The quantity transported is determined both by the rapidity with which the blood circulates and by the proportion of gas delivered that is removed by the tissues.

The rapidity of circulation is determined by the output of the heart, which in turn is responsive to overall body requirements. Local flows can be increased selectively, as occurs, for example, in the flow through skeletal muscles during exercise. The performance of the heart and circulatory regulation are, therefore, important determinants of gas transport.

Oxygen in solution and in red blood cells

Transport of oxygen. Oxygen is transported both in solution within the plasma (the blood except for its blood cells) and in combination with hemoglobin within the red blood cells. The oxygen dissolved in plasma is approximately 0.3 millilitre per 100 millilitres of blood. In the same quantity of blood (100 millilitres) there are approximately 14 grams of hemoglobin, which carries


arteriole
oxygen tension 100 mm
oxygen in solution 0.3 cc
oxygen in cells 19 cc
% saturated hemoglobin 95
carbon dioxide tension 40 mm
tissue cell
tissue cell
tissue cell
19 cc
17 cc
15 cc
venule
oxygen tension 40 mm
oxygen in solution 0.15 cc
oxygen in cells 15 cc
% saturated hemoglobin 75
carbon dioxide tension 46 mm

Passage of oxygen from blood to tissues in a resting organ (see text).
From Samson Wright's Applied Physiology, **revised by** Cyril A. Keele and Eric Neil, 11th ed. (1965); Oxford University Press

18.76 millilitres of oxygen, more than 60 times the amount in solution in the plasma.

Not all of the oxygen transported is transferred to the tissue cells. The amount they extract depends in part on their rate of energy expenditure. Changes in oxygen requirements of the tissue cells are met by changes in either the rate at which oxygen is delivered by the blood (how much blood perfuses a tissue per unit of time) or the degree to which oxygen is extracted from a given quantity of blood or by both processes.

Oxygen escapes from arterial blood into tissue fluid in response to a difference in partial pressure. The partial pressure at equilibrium between tissue and arterial blood becomes the partial pressure of oxygen in venous blood returning to the lung for replenishment of oxygen and completion of the transport cycle.

The amount of oxygen held combined with hemoglobin varies with the partial pressure to which the hemoglobin or blood is exposed. The relationship between amount held and partial pressure is such that hemoglobin is practically fully saturated by exposure to alveolar partial pressure of oxygen and holds oxygen relatively poorly at tissue partial pressures. It arrives at the tissue "fully loaded," therefore, and normally, when it leaves, has released about one-fifth of its oxygen (see the illustration). Delivery to the tissues is favoured also by the fact that carbaminohemoglobin (see below) formed in the tissue capillaries has a relatively poor affinity for oxygen.

Hemoglobin in the fetus differs from that in the adult in having greater affinity for, or tendency to combine with, oxygen at a given partial pressure of oxygen. This characteristic makes it possible for fetal blood to carry sufficient oxygen despite its low partial pressure of oxygen (about one-third that of maternal blood). The normal low oxygen level of fetal blood is caused by the dilution of arterial maternal blood by deoxygenated fetal venous blood in the placenta.

In a recent discovery, red blood cells were found to contain a substance—2,3-diphosphoglycerate—capable of increasing the availability of oxygen to tissues by reducing the tendency of hemoglobin to hold oxygen at its low partial pressure. It was found that the quantity of the substance in the red blood cells increases with gains in altitude, with occurrences of the diseases emphysema (abnormal inflation of the lungs) and anemia (characterized by a low number of normal red blood cells or low quantity of normal hemoglobin), and in other circumstances in which delivery of oxygen to tissues is in jeopardy.

Four methods of CO_2 transport

Carbon dioxide transport. Venous blood leaving the tissues contains about 52 millilitres of carbon dioxide per 100 millilitres; arterial blood leaving the lungs contains about 48 millilitres. The four-millilitre difference, representing the amount removed in the expired air, is transported in four forms: (1) a negligible amount as carbonic acid (H_2CO_3); (2) about 0.3 millilitre as free gas in solution; (3) about 1.2 millilitres combined with protein (carbaminohemoglobin); and (4)about 2.5 millilitres as bicarbonate ions (HCO_3^-). Almost all of the gas, therefore, is transported in two forms, either as carbaminohemoglobin (30 percent of the total) or bicarbonate (63 percent).

Hemoglobin's ability to transport carbon dioxide from the tissues to the lungs and release it there is enhanced because hemoglobin lean in oxygen (reduced hemoglobin) combines more readily with carbon dioxide than hemoglobin rich in oxygen (oxyhemoglobin). Thus when oxygen-laden hemoglobin arrives at the tissues and releases oxygen, it then tends to pick up carbon dioxide, becoming carbaminohemoglobin. Conversely, when hemoglobin laden with carbon dioxide arrives in the lung and takes up oxygen, it tends to release carbon dioxide, which is diffused into the lung alveoli and breathed out.

The major part (63 percent) of the carbon dioxide that has diffused from the tissues with the blood of the capillaries and has been transported as bicarbonate combines with water to form carbonic acid (H_2CO_3). The reaction is slow when the enzyme carbonic anhydrase is absent from the red blood cells; usually, however, the enzyme is present, and carbonic acid forms some 200 times faster than it does outside the cells. Carbonic acid within the red cells quickly dissociates to hydrogen ions (H^+) and bicarbonate ions (HCO_3^-), most of the hydrogen ions being taken up by reduced hemoglobin, a weak acid. The bicarbonate ions are matched by intracellular cations (positive ions), primarily potassium.

In the lung, the equilibrium shifts oppositely. As oxyhemoglobin, a more ionized acid, is formed, hydrogen ions are released, resulting in carbonic acid, which is dissociated within the red cell to carbon dioxide for diffusion into the plasma and then to alveolar gas.

Oxygen and carbon dioxide reciprocate in effecting changes in hemoglobin, each fostering the release and transport of the other. The change in affinity for hydrogen ions, characteristic of reduced and oxygenated hemoglobins, makes it possible for carbon dioxide transport to occur with minimal change in the hydrogen ion concentration of the blood.

Regulation. The movements of breathing continue independently of volition or consciousness and yet are responsive to the will and to changes in the organism or environment. Regulation of breathing, therefore, includes provision for automaticity, for optional volitional control, and for appropriate responsiveness.

Automaticity of rhythm and depth of breathing

Automaticity is incompletely understood. Cells in the brainstem in front of the fourth ventricle (a fluid-filled cavity near the base of the brain) are closely implicated in establishing the rhythm and depth of breathing. These cells are sensitive to fluctuations in carbon dioxide partial pressure and the hydrogen ion concentration of their surrounding fluid, producing changes in alveolar ventilation in keeping with the maintenance of normal carbon dioxide and hydrogen ion concentrations. The cells are also affected by impulses from higher stations in the central nervous system that mediate volitional and "psychogenic" respiratory drives.

Finally, the cells in the brainstem receive impulses from chemoreceptors (structures responding to chemical stimuli) sensitive to carbon dioxide, hydrogen ion concentrations, and oxygen pressures; these receptors are located above the heart (in the arch of the aorta) and at the base of the neck (in the fork, or bifurcation, of the common carotid artery). Impulses also come from other types of receptors: stretch receptors in the lung and chest that signal both stretch-inspiration and relaxation-expiration; receptors that are in the vicinity of joints and detect activity there; receptors variously scattered, as in the skin, that do not have specific relevance to respiration but that nevertheless influence breathing through medullary centres. Clearly, therefore, the system is designed to provide feedback to the centre from peripheral sensors responsive to a variety of changes.

The partial pressure of carbon dioxide in the tissues is the variable most delicately preserved by the regulatory mechanisms. This is assured primarily by the sensitivity of medullary cells themselves (the cells in the medulla, that part of the brainstem directly above the spinal cord) and secondarily through the aortic and carotid chemoreceptors. Alveolar ventilation is about doubled when arterial carbon dioxide pressure is increased from 40 to 43 millimetres of mercury and is quadrupled by an increase from 40 to 47 millimetres.

Increased acidity (hydrogen ion concentration) of either blood or medullary tissue fluid also increases alveolar ventilation. Indeed, evidence suggests that the effect of carbon dioxide on alveolar ventilation is indirect, through an attendant rise in hydrogen ion concentration.

A low oxygen level is a relatively ineffectual respiratory drive, a fall in arterial oxygen pressure all the way to 60 millimetres of mercury (from a normal of about 100 millimetres) being required to induce a ventilatory response. When the respiratory centre loses its sensitivity to carbon dioxide, and alveolar ventilation fails, however, stimulation through the chemoreceptors sensitive to oxygen becomes critical.

In summary, the regulation of breathing is one of the most impressive examples of complex physiologic regulation in the interest of preserving an optimal cellular environment.

BIBLIOGRAPHY. Textbooks of physiology are one important source of additional information. Three particularly recommended are: A.C. GUYTON, *Basic Human Physiology,* (1971); V.E. MOUNTCASTLE (ed.), *Medical Physiology,* 12th ed. (1968); and C.H. BEST and N.B. TAYLOR, *The Physiological Basis of Medical Practice,* 8th ed. (1966). Monographs devoted entirely to either the subject as a whole or specific subdivisions are: J.H. COMROE, *The Physiology of Respiration* (1965); J.E. COTES, *Lung Function,* 2nd ed. (1968); and E.J.M. CAMPBELL, E. AGOSTINI, and J. NEWSOM DAVIS, *The Respiratory Muscles,* 2nd ed. (1970). An annual review of developments in the field may be found in the *Annual Review of Physiology.*

(A.A.S.)

Respiration and Respiratory Systems

Man failed to understand–the significance of his own breathing until late in the 18th century. In 1777 Lavoisier reported that the air in a closed jar containing a breathing bird showed an increasing amount of a gas that was absorbed by soda lime; that as the animal continued to respire there was a decrease in a life-sustaining gas he called *oxygdne.* The gas absorbed by soda lime was carbon dioxide.

The life processes of animals involve the expenditure of energy, and it is through the oxidation of molecules containing carbon that this energy, along with carbon dioxide, is released. Respiration is the process in which animal organisms consume oxygen and release carbon dioxide. This exchange of oxygen taken in for carbon dioxide given out stands in a remarkable relationship with plant life, for plants take in carbon dioxide and yield up oxygen, which can then enter once again into animal metabolism (see Figure 1). Although the acquisition of oxygen and elimination of carbon dioxide are essential requirements for all animals, the amounts vary according to the kind of animal and its state of activity. In relation to its weight, a butterfly requires more oxygen than a man, considerably more when it is flying. In Table 1 the oxygen consumption of various animals is expressed in terms of millilitres per kilogram of weight per hour. The butterfly of the genus *Vanessa,* for example, consumes 600 millilitres of oxygen per kilogram of weight per hour when resting, but during flight its consumption rises to 100,000 millilitres per kilogram of weight per hour—or 25 times as much, in relative terms, as a man doing hard physical work.

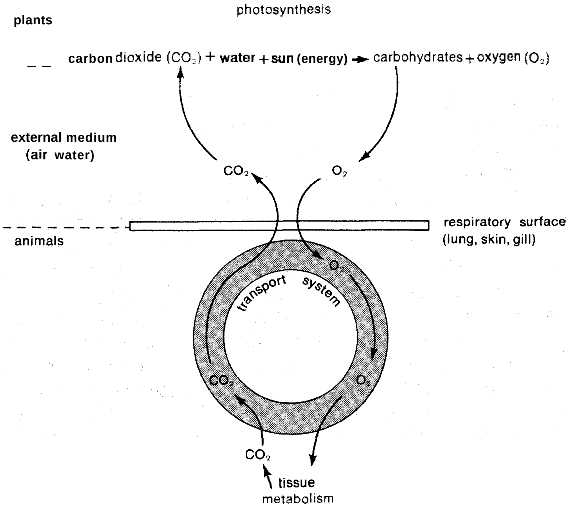


Figure 1: Relationship between photosynthesis and respiration (see text).

Table 1: Oxygen Consumption of Various Animals and Its Variation with Rest and Activity

	weight (grams)	oxygen consumption (millilitres per kilogram of weight per hour)
Paramecium	0.000001	500
Mussel (*Mytilus*)	25	22
Crayfish (*Astacus*)	32	47
Butterfly (Vanessa)	0.3	
resting		600
flying		100,000
Carp (*Cyprinus*)	200	100
Pike (*Esox*)	200	350
Mouse	20	
resting		2,500
running		20,000
Man	70,000	
resting		200
maximal work		4,000

Source: A. Krogh, The Comparative Physiology of Respiratory Mechanisms (1959).

Very small organisms such as Protozoa require no special respiratory apparatus but rely on the diffusion of oxygen and carbon dioxide across the general surface area of their cell membranes. The evolution of larger animals has been accompanied by the development of specialized mechanisms for extracting oxygen from the environment and releasing carbon dioxide. These include gills, lungs, and various modifications of the outer surface of the body; all have in common the presentation to the external environment of a large surface area, often well-supplied with blood vessels, across which diffusion of the gases can occur. Since the tissues that ultimately consume oxygen may be quite remote from the gas-exchange surfaces, transport systems are employed to deliver gases across relatively great distances. The circulatory system in vertebrate animals is controlled by delicate mechanisms that regulate local blood flow in accordance with the needs of the various tissues. Many animals have specialized blood-borne pigments (hemoglobins in vertebrates, hemoglobins and other pigments in invertebrates) that convey most of the oxygen from the respiratory exchange surface to the tissues. The blood plasma itself is an aqueous fluid that holds only a limited quantity of oxygen.

Since an animal's requirements for oxygen vary over a wide range, its capacity for gas exchange with the environment must be adjusted accordingly. Increases in respiratory depth and rate result from strenuous exercise; quiet sleep or rest is correspondingly accompanied by a reduction in respiratory activity. In many animals the nervous system senses changes in the chemical composition of the body fluids reflecting the gas demands; the central nervous system then responds by exciting or depressing the machinery of external respiration.

The article is divided into the following sections:

I. General features

THE GASES IN THE ENVIRONMENT

Varying composition of air

The range of respiratory problems faced by aquatic and terrestrial animals can be seen from the varying composition and physical characteristics of water and air. Air contains about 20 times the amount of oxygen found in air-saturated water. In order to extract an equivalent amount of oxygen as an air breather, an aquatic animal may find it necessary to pass across the respiratory surfaces a relatively larger volume of the external medium. Moreover, the diffusion rate of oxygen is much lower in water than in air. The problem is further compounded by the higher density (1,000 times ail) and viscosity (100 times air) of water, which impose on the machinery of aquatic respiration a much greater work load. Thus it is not surprising to find that fish may expend about 20 percent of their total oxygen consumption in running the respiratory pump, as compared with about 1 to 2 percent in mammals, including man.

The carbon dioxide content of most natural waters is low compared to air, often almost nil. In contrast to oxygen, carbon dioxide is extremely soluble in water and diffuses rapidly. Most of the carbon dioxide entering water combines either with the water (to form carbonic acid) or with other substances (to form carbonates or bicarbonates). This buffering capacity maintains a low level of free carbon dioxide and facilitates the maintenance of a favourable diffusion gradient for carbon dioxide exchange by water breathers. In general, oxygen exchange, which is strongly dependent on the oxygen content of the water, is more critically limiting for aquatic forms than is the exchange of carbon dioxide.

Temperature exerts a profound effect on the solubility of gases in water. A change from 5° to 35° C (41" to 95" F) reduces the oxygen content of freshwater by nearly half. At the same time, a rise in body temperature produces an increase in oxygen consumption among animals that do not closely regulate their body temperatures (so-called cold-blooded animals). A fish experiencing both rising water and body temperatures is under a double handicap: more water must be pumped across its gill surfaces to extract the same amount of oxygen as was needed at the lower temperature; and the increased metabolism requires greater quantities of oxygen.

The amount of oxygen available in natural waters is also limited by the amount of dissolved salts. This factor is a determinant of oxygen availability in transitional zones between sea and freshwater. Pure water, when equilibrated with oxygen at 0° C, for example, contains about 50 millilitres of oxygen per litre; under the same conditions, a solution containing 2.9 percent of sodium chloride contains only 40 millilitres of oxygen per litre. Bodies of water may have oxygen-poor zones. Such zones are especially evident in swamps and at the lower levels of deep lakes. Many animals are excluded from such zones; others have become remarkably adapted to living in them.

The Earth's atmosphere extends to a height of many miles. It is composed of a mixture of gases held in an envelope around the globe by gravitational attraction. The atmosphere exerts a pressure proportional to the weight of a column of air above the surface of the Earth extending to the limit of the atmosphere: atmospheric pressure at sea level is sufficient to support a column of mercury 760 millimetres in height (abbreviated as 760 mm Hg—the latter being the chemical symbol for mercury). Dry air is composed chiefly of nitrogen (79.02 percent), oxygen (20.94 percent), and carbon dioxide (0.03 percent), each contributing porportionately to the total pressure. These percentages are relatively constant to about 50 miles in altitude. At sea level and a barometric pressure of 760 millimetres of mercury, the pressure of nitrogen is 79.02 percent of 760 millimetres of mercury, or 600.55 millimetres of mercury; that of oxygen is 159.16 millimetres of mercury; and that of carbon dioxide is 0.20 millimetres of mercury.

The existence of water vapour in a gas mixture reduces the partial pressures of the other component gases but does not alter the total pressure of the mixture. The importance of water-vapour pressure on gas composition can be appreciated from the fact that at the body temperature of man (37° C, or 98.6" F), the atmospheric air drawn into the lungs becomes saturated with water vapour. The water-vapour pressure at 37° C is 47 millimetres of mercury. To calculate the partial pressures of the respiratory gases, this value must be subtracted from the atmospheric pressure. For oxygen, 760 (the atmospheric pressure) —47 = 713 millimetres of mercury, and 713 × 0.209 (the percentage of oxygen in the atmosphere) = 149 millimetres of mercury; this amounts to some ten millimetres of mercury lower than the partial pressure of oxygen in dry air at 760 millimetres of mercury total pressure.

Atmospheric pressures fall at higher altitudes, but the composition of the atmosphere remains unchanged, At 25,000 feet the atmospheric pressure is 282 millimetres of mercury and the partial pressure of oxygen about 59 millimetres of mercury. Oxygen continues to be only 20.94 percent of the total gas present. The rarefaction of the air at high altitudes not only limits the availability of oxygen for the air breather, it also limits its availability for aquatic forms, since the amount of dissolved gas in water decreases in parallel with the decline in atmospheric pressure. Lake Titicaca in Peru is at an altitude of about 12,500 feet (3,810 metres); one litre of lake water at this altitude (and at 20° C, or 68° F) holds 4 millilitres of oxygen in solution; at sea level, it would hold 6.4.

These variations in the characteristics of air and water suggest the many problems with which the respiratory systems of animals must cope in procuring enough oxygen to sustain life.

BASIC TYPES OF RESPIRATORY STRUCTURES

Absence of respiratory organs in some small animals

The first step in respiration takes place at surfaces or interfaces between the organism and an environmental medium that contains oxygen. Small forms, such as protozoans, utilize their entire surface. Multicellular animals of considerably greater bulk than the protozoans may also utilize their general body surface for gas exchange. Flatworms, for example, have no specialized respiratory structures; because of their flatness, the distance between the surface and the interior fall well within the range in which diffusion alone is sufficient to meet oxygen requirements so that no specialized internal transport system is required for delivery. Among other animals the respiratory exchange sites may take the form of gills, lungs, specialized areas of the intestine or pharynx (in certain fishes), or tracheae (air tubes penetrating the body wall, as in insects).

Respiratory structures typically have an attenuated shape with a semipermeable surface that is large in relation to the volume of the structure. Within them there is usually a circulation of body fluids (blood through the lungs, for example). Two sorts of pumping mechanisms are frequently encountered: one to renew the external oxygen-containing medium, the other to insure circulation of the body fluids through the respiratory structure. In air-breathing vertebrates, alternately contracting sets of muscles create the pressure differences needed to expand or deflate the lungs, while the heart pumps blood through the respiratory surfaces within the lungs.

Respiratory organs of invertebrates. ***Tracheal** respiration.* The principal device for external respiration in insects is a system of tubes (tracheae) opening by pores (spiracles) to the outside and branching into smaller units (tracheoles) as they penetrate deeply into the tissues (see Figure 2). There are typically two pairs of spiracles in the thorax and eight in the abdomen. The spiracles may open and close periodically, and it is thought that spiracular closure is a means of preventing excess water loss through evaporation. The rate of opening and closing and the number of spiracles utilized appear to depend on the oxygen demands of the animal. Ventilation of the tracheal system, especially in large insects, may be augmented by muscular pumping motions of the abdomen.

A number of insects spend their larval existence or a good portion of their adult life submerged in water. Some of the adaptations, involving the use of the tracheal system for underwater exchange, are shown in Figure 2. Of special interest are those that might be termed bubble breathers, which, as in the case of the water boatman, take on a gas supply before submerging. The boatman traps an air bubble under its wing surfaces next to its spiracles. The beetles swim down from the surface and cling to algal filaments or other material on the bottom. As oxygen is consumed from the bubble, the partial pressure of oxygen within the bubble falls below that in the water; consequently oxygen diffuses from the water into the bubble to replace that consumed. The carbon dioxide produced by the insect diffuses through the tracheal system into the bubble and on into the water. The bubble thus behaves like a gill. There is one major limitation to this adaptation: as oxygen is removed from the bubble, the partial pressure of the nitrogen rises, and this gas then diffuses outward into the water, The consequence of outward nitrogen diffusion is that the bubble shrinks and must be replaced by another trip to the surface. A partial solution to the problem of bubble renewal has been achieved by small aquatic beetles of the family Elmidae (*e.g., Elmis,* Riolus) that capture bubbles containing oxygen produced by algae and incorporate this gas into the bubble gill. A number of aquatic beetles also augment gas exchange by stirring the surrounding water with their posterior legs.

This figure first appeared in New *Scientist,* the weekly review of science and technology, 128 Long Acre London W C 2

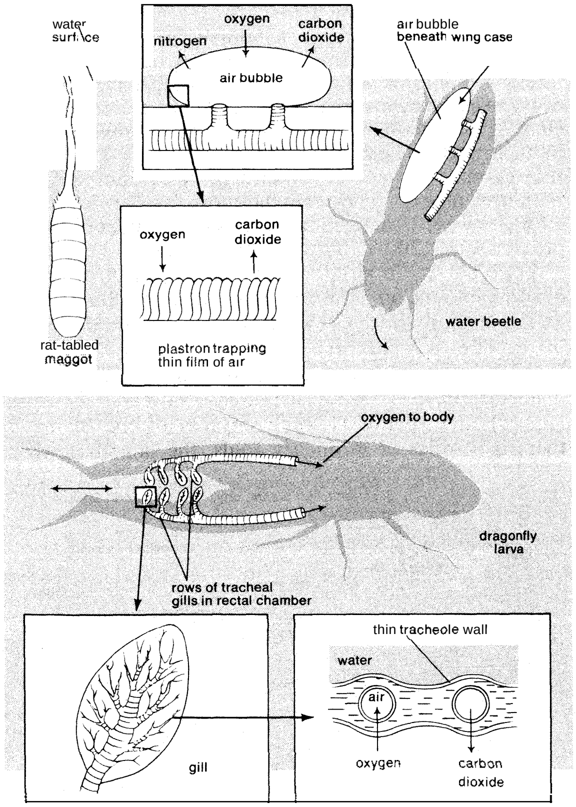


Figure 2: Ways by which aquatic insects obtain oxygen.

An elegant solution to the problem of bubble exhaustion during submergence has been achieved by certain beetles that have a high density of cuticular hair over much of the surface of the abdomen and thorax. The hair pile is so dense that it resists wetting, and an air space is formed below it, creating a plastron, or air shell, into which the tracheae open. As respiration proceeds, the outward diffusion of nitrogen and consequent shrinkage of the gas space are prevented by the surface tension—a condition manifested by properties that resemble those of an elastic skin under tension—between the closely packed hairs and the water. The plastron becomes "permanent" in the sense that further bubble trapping at the surface is no longer necessary, and the beetles may remain submerged indefinitely. Since the plastron hairs tend to resist deformation, the beetles can live at considerable depths without compression of the plastron gas.

The "book lungs" of spiders

The respiratory structures of spiders consist of peculiar "book lungs," leaflike plates over which air circulates through slits on the abdomen. The book lungs contain blood vessels that bring the blood into close contact with the surface exposed to the air and where gas exchange between blood and air occurs. In addition to these structures, there may also be abdominal spiracles and a tracheal system like that of insects.

Since spiders are air breathers, they are mostly restricted to terrestrial situations, although some of them regularly hunt aquatic creatures at stream or pond edges and may actually travel about on the surface film as easily as on land. The water spider Argyroneta aquatica, the frogman of the spider world, utilizes the water-beetle method of capturing air bubbles at the surface. The bubble is pressed against the respiratory openings on the abdomen, but, because there is no permanent plastron, trips must be made to the surface for bubble renewal. Most of the life cycle, however, including courtship and breeding, prey capture and feeding, and development of eggs and embryos, occurs below the water surface. Many of these activities take place in a kind of diving bell formed by silk. The spider weaves an inverted basket-like web that is anchored to underwater plants or other objects. Bubbles captured at the surface are ejected into the interior, inflating the underwater house with air. The combination of bubble trapping and web building has given Argyroneta aquatica access to an environment denied most of its relatives.

Among many immature insects are special adaptations for an aquatic existence. Thin-walled protrusions of the integument, containing tracheal networks, form a series of gills (tracheal gills) serving to bring water into close contact with the closed tracheal tubes (see Figure 2). The nymphs of mayflies and dragonflies have external tracheal gills attached to their abdominal segments, and certain of the gill plates may move in a way that sets up water currents over the exchange surfaces. Dragonfly nymphs possess a series of tracheal gills enclosed within the rectum. Periodic pumping of the rectal chamber serves to renew water flow over the gills. Removing the gills or plugging the rectum results in lower oxygen consumption. A considerable gas exchange also occurs across the general body surface in immature aquatic insects.

The insect tracheal system has inherent limitations. Gases diffuse slowly in long narrow tubes, and effective gas transport can occur only if the tubes do not exceed a certain length. It is generally thought that this has imposed a size limit upon insects.

Varieties of gill-like structures

Gills of invertebrates. Numerous groups of animals, including invertebrates, utilize gills as a major avenue of gas exchange. Almost any thin-walled extension of the body surface that comes in contact with the environmental medium and across which gas exchange occurs can be viewed as a gill. Gills usually have a large surface area in relation to their mass; pumping devices are often employed to renew the external medium.

The marine polychaete worms utilize the general body surface for gas exchange, but they also have a variety of gill-like structures: segmental flaplike parapodia (in *Nereis*) or elaborate branchial tufts (among the families Terebellidae and Sabellidae). The tufts, used to create both feeding and respiratory currents, offer a large surface area for gas exchange.

Although most of the respiratory exchange in echinoderms (starfish, sea urchins, brittle stars) occurs across their tube feet (a series of suction-cup extensions used for locomotion), this exchange is supplemented by extensions of the coelomic, or body-fluid, cavity into thin-walled "gills" or dermal branchiae that bring the coelomic fluid into close contact with seawater. Sea cucumbers (Holothuroidea), soft-bodied, sausage-shaped echinoderms that carry on some respiration through their tube feet, also have an elaborate "respiratory tree" consisting of branched hollow outpouchings off the cloaca (hindgut). Water is pumped in and out of this system by the action of the muscular cloaca, and it is probable that a large fraction of the animals' respiratory gas is exchanged across this system.

Among mollusks gills have a relatively elaborate blood supply, although respiration across the mantle, or general epidermis, also occurs. Clams possess gills across which water circulates, impelled by the movements of millions of microscopic whips called cilia. In the few forms studied, the extraction of oxygen from the water has been found to be low, on the order of 2 to 10 percent. The currents produced, which constitute ventilation, are also utilized for bringing in and extracting food. At low tide or during a dry period, clams and mussels close their shells and thus prevent dehydration. Metabolism then shifts from oxygen-consuming (aerobic) pathways to others in which an accumulation of acid products occurs; when normal conditions are restored, the animals increase their ventilation and oxygen extraction in order to rid themselves of the acid products. Snails have a feeding mechanism that is independent of the respiratory surface. A portion of the mantle cavity in the form of a gill or "lung" serves as a gas-exchange site. In air-breathing snails, the "lung" may be protected from drying out through contact with the air by having only a pore in the mantle as an opening to the outside. Cephalopod mollusks, such as squid and octopus, actively ventilate a protected chamber lined with feathery gills that contain small blood vessels (capillaries); their gills are quite effective, extracting 60 to 80 percent of the oxygen passing through the chamber. In oxygen-poor water, the octopus may increase its ventilation tenfold, indicating a more active control of respiration than appears to be present in other classes of mollusks.

Gills of crustaceans

Many crustaceans (crabs, shrimps, crayfish) are very dependent on their gills; these are often enclosed in protected chambers, and ventilation is provided by specialized appendages that create the respiratory current. As in cephalopod mollusks, oxygen utilization is relatively high —up to 70 percent in the European crayfish Astacus. When the partial pressure of oxygen in the water falls, there is a marked increase in ventilation (the volume of water passing over the gills); at the same time, the rate of oxygen utilization declines somewhat. Although more oxygen is extracted per unit of time, the increased ventilation is not without metabolic cost, since the work of operating the crustacean's respiratory apparatus itself requires oxygen; together with the decrease in extraction per unit of volume, it probably limits aquatic forms of crustaceans to levels of oxidative metabolism lower than those found in many air-breathing forms. Not all crustaceans meet a reduction in oxygen with increased ventilation and metabolism. The square-backed crabs (*Sesarma)* become less active, reducing their oxidative metabolism until more favourable conditions prevail.

Respiratory organs of vertebrates. The organs of external respiration of most vertebrates are usually thin-walled structures well supplied with blood vessels. Such structures bring blood into close association with the external medium so that the exchange of gases takes place across relatively small distances. There are three major types of respiratory structures in the vertebrates: gills, integumentary exchange areas, and lungs. The gills are totally external in a few forms (as in *Necturus*, a neotenic salamander), but in most they are composed of filamentous leaflets protected by bony plates (as in fish). Some fishes and numerous amphibians also use the body integument, or skin, as a gas-exchange structure. Both gills and lungs are formed during the animal's development from outpouchings of the gut wall. Such structures have the advantage of a protected internal location, but this requires some sort of pumping mechanism to move the external gas-containing medium in and out.

The quantity of air or water passing through the lungs or gills each minute is known as the ventilation volume. The rate or depth of respiration may be altered to bring about adjustments in ventilation volume. The ventilation volume of man at rest is approximately six litres per minute. This may increase to more than 100 litres per minute with increases in the rate of respiration and the quantity of air breathed in during each respiratory cycle (tidal volume). Certain portions of the airways (trachea, bronchi, bronchioles) do not participate in respiratory exchange, and the gas that fills these structures occupies an anatomical dead space of about 150 millilitres in volume. Of a tidal volume of 500 millilitres, only 350 millilitres ventilate the gas-exchange sites.

The maximum capacity of human lungs is about six litres. During normal quiet respiration, some 500 millilitres, the tidal volume, is inspired and expired during every respiratory cycle. The lungs are not collapsed at the close of expiration; a certain volume of gas remains within them. At the close of the expiratory act, a normal subject may, by additional effort, expel another 1,200 millilitres of gas. Even after the most forceful expiratory effort, however, there remains a residual volume of approximately 1,200 millilitres. By the same token, at the end of a normal inspiration, further effort may succeed in drawing into the lungs an additional 3,000 millilitres.

Fish gills

Gills of vertebmtes. The gills of fishes are supported by a series of gill arches encased within a chamber formed by bony plates (the operculum). A pair of gill filaments projects from each arch; between the dorsal (upper) and ventral (lower) surfaces of the filaments, there is a series of secondary folds, the lamellae, where the gas exchange takes place (Figure 3). The blood vessels passing through the gill arches branch into the filaments and then into still smaller vessels (capillaries) in the lamellae. Deoxygenated blood from the heart flows in the lamellae in a direction counter to that of the water flow across the exchange surfaces. Microscopic observation has shown that in a number of fishes the water-to-blood distance across which gases must diffuse is from 0.0003 to 0.003 millimetres, or about the same distance as the air-to-blood pathway in the mamm'alian lung.

This diagram first appeared in *New Scientist*, the weekly review of science and technology, 128 Long Acre, London W.C. 2

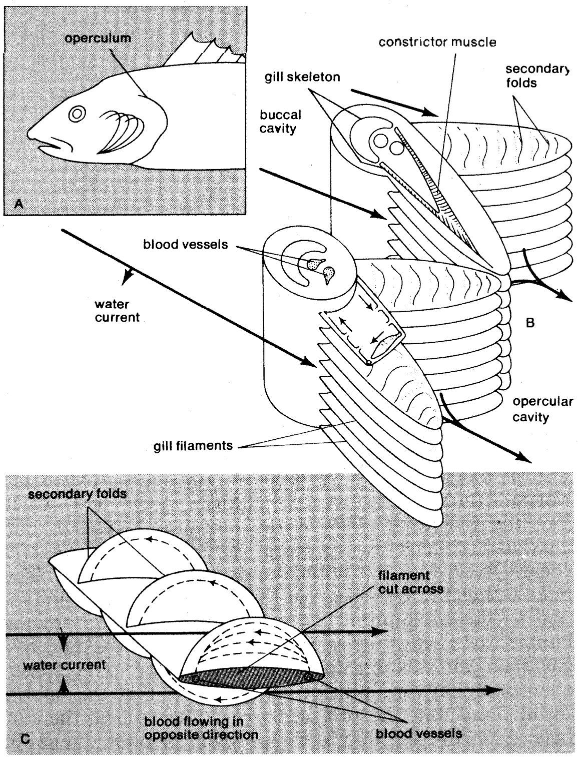


Figure 3: (A) Position of four gill arches beneath operculum on left side of fish. (B) Part of two gill arches with filaments of adjacent rows touching at their tips and the blood vessels that carry the blood before and after its passage over the gills. (C) Part of a single filament with three secondary folds on each side.

The fact that blood flows through the lamellae in a direction counter to that of water has much to do with the efficiency of gas exchange. Laboratory experiments in which the direction of water flow across fish gills was reversed showed that about 80 percent of the oxygen was extracted in the normal situation, while only 10 percent was extracted when water flow was reversed. The uptake of oxygen from water to blood is thus facilitated by countercurrent flow; in this way, greater efficiency of oxygen uptake is achieved by an anatomical arrangement that is free of energy expenditure by the organism. Countercurrent flow is a feature of elasmobranchs (sharks, skates) and cyclostomes (hagfishes, lampreys) as well as bony fishes.

A number of vertebrates use externalized gill structures. Some larval fishes have external gills that are lost with the appearance of the adult structures. A curious example of external gills is found in the male lungfish (Lepidosiren). At the time the male begins to care for the nest, a mass of vascular filaments (a system of blood vessels) develops as an outgrowth of the pelvic fins. The fish meets its own needs by refilling its lungs with air during periodic excursions to the water surface. When it returns to the nest, its pelvic-gill filaments are perfused with well-oxygenated blood, providing an oxygen supply for the eggs, which are more or less enveloped by the gill filaments.

It is theoretically possible for a skin that is well supplied with blood vessels to serve as a major or even the only respiratory surface. This requires a thin, moist, and heavily vascular skin that increases the animal's vulnerability to enemies. In terrestrial animals a moist integument also provides a major avenue of water loss. A number of fishes and amphibians rely on the skin for much of their respiratory ge hibernating frogs utilize the ski for practically all their gas exchanges.

Characteristics of vertebrate lungs

***The** lung.* The lungs of vertebrates range from simple sacklike structures found in the Dipnoi (lungfishes) to the complexly subdivided organs of mammals and birds. An increasing subdivision of the airways and the development of greater surface area at the exchange surfaces appear to be the general evolutionary trend among the higher vertebrates.

In the embryo, lungs develop as an outgrowth of the forward portion of the gut. The lung proper is connected to the outside through a series of tubes; the main tube, known as the trachea (windpipe), exits in the throat through a controllable orifice, the glottis. At the other end the trachea subdivides into secondary tubes (bronchi), in varying degree among different vertebrate groups.

The trachea of amphibians is not divided into secondary tubes but ends abruptly at the lungs. The relatively simple lungs of frogs are subdivided by incomplete walls (septa), and between the larger septa are secondary septa that surround the air spaces where gas exchange occurs. The diameter of these air spaces (alveoli) in lower vertebrates is larger than that in mammals. The alveolus in the frog is about ten times the diameter of that found in humans. The smaller alveoli in mammals are associated with a greater surface for gas exchange: although the respiratory surface of the frog (Rana) is about 20 square centimetres for each cubic centimetre of air, that of man is about 300 square centimetres (46 square inches).

An important characteristic of lungs is their elasticity. An elastic material is one that tends to return to its initial state after the removal of a deforming force. Elastic tissues behave like springs. As the lungs are inflated, there is an accompanying increase in the energy stored within the elastic tissues of the lungs, just as energy is stored in a stretched rubber band. The conversion of this stored, or potential, energy into kinetic, or active, energy during the deflation process supplies part of the force needed for the expulsion of gases. A portion of the energy put into expansion is thus recovered during deflation. The elastic properties of the lungs have been studied by inflating them with air or liquid and measuring the resulting pressures. Muscular effort supplies the motive power for expanding the lungs, and this is translated into the pressure required to produce lung inflation. It must be great enough to overcome (1) the elasticity of the lung and its surface lining; *(2)* the frictional resistance of the lungs; (3) the elasticity of the thorax or thoraco-abdominal cavity; (4) frictional resistance in the body-wall structures; (5) resistance inherent in the contracting muscles; and (6) the airway resistance. The laboured breathing of the asthmatic serves as an example of the added muscular effort necessary to achieve adequate lung inflation when the airway resistance is high, owing to narrowing of the tubes of the airways.

Studies of the pressure-volume relationship of lungs filled with salt solution or air have shown that the pressure required to inflate the lungs to a given volume is less when the lungs are filled with liquid than when they are filled with air. The differences in the two circumstances have been thought to result from the nature of the environment-alveolar interface, that interface being liquid–liquid in the fluid-filled lung and gas–liquid in the air-filled lung. In the case of the latter, the pressure–volume relationship represents the combined effects of the elastic properties of the lung wall plus the surface tension—the property, resulting from molecular forces, that exists in the surface film of all liquids and tends to contract the volume into a form with the least surface area; the particles in the surface are inwardly attracted, thus resulting in tension—of the film, or surface coating, lining the lungs (surface tension being nearly zero in the fluid-filled lung.).

The alveoli of the lungs are elastic bodies of nonuniform size. If their surfaces had a uniform surface tension, small alveoli would tend to collapse into large ones. The result in the lungs would be an unstable condition in which some populations of alveoli would collapse and others would overexpand. This does not normally occur in the lung because of the properties of its surface coating (surfactant), a complex substance composed of lipid and protein. This material causes the surface tension to change in a nonlinear way with surface area. As a result, when the lungs fill with air, the surface tensions of the inflated alveoli are greater than those of the relatively undistended alveoli. This results in a stabilization of alveoli of differing sizes and prevents the emptying of small alveoli into larger ones. It has been suggested that compression wrinkles of the surface coating and attractive forces between adjacent wrinkles inhibit expansion. Surfactants have been reported to be present in the lungs of birds, reptiles, and amphibians.

II. Dynamics of respiratory mechanisms

RESPIRATION IN FISHES

Among the most primitive of present-day vertebrates are the cyclostomes (lampreys and hagfishes), the gill structures of which are in the form of pouches that connect internally with the pharynx (throat) and open outward through slits, either by a fusion of the excurrent gill ducts into a single tube (in Myxine) or individually by separate gill slits (in *Petromyzon*). The gill lamellae of cyclostomes form a ring around the margins of the gill sac, and the series of sacs is supported in a flexible branchial skeleton. The number of paired pouches varies in different forms from six to 14. The pharynx of lampreys divides into an esophagus above and a blind tube below, from which the gill pouches arise. The upper pharynx of hagfishes communicates to the exterior through a nostril, a structure absent in lampreys. When the parasitic lampreys are embedded in the flesh of fish, upon which they live, they maintain a flow of water through the gills by alternate contractions of the gill pouches. When the gill-pouch muscles relax, the pouches expand, and water is sucked in. The water is forced out through the gills by muscular contraction; the branchial musculature apparently prevents reflux of the water into the pharynx while the head of the lamprey is embedded in the flesh of its prey.

In the hagfish Myxine *glutinosa,* the major oxygen supply is derived from water drawn in through the nostril that opens into the pharynx. The velum, a peculiar respiratory structure, lies just behind the nostril opening, where it is suspended from the upper midline of the pharynx by a membrane (the velar frenulum) from which a horizontal bar extends; the structure resembles an inverted T. Membranous scrolls attached to this horizontal bar can extend downward and then roll upward like window shades. During the upward motion, they scoop up water and direct it upward toward the frenulum and backward toward the gill pouches. Muscular sphincters on the inflow and outflow orifices of the gill ducts help to control the directional flow of water through the gill pouches. When the fish exhales, the inflow sphincters contract and the gill-pouch volume decreases; contraction of muscle elements within the gill pouches contributes to the expulsion of water. When the fish inhales, contraction of the outflow and relaxation of the inflow sphincters coupled with the elastic rebound of the gill pouches probably create a favourable pressure gradient for filling. The degree to which velar contractions are coordinated with these gill-pouch contractions is not known. When suspended particles are introduced into the nostril, they produce a violent "sneeze," and the foreign material is expelled from both the mouth and the nostril. This reaction is probably important to an animal having common respiratory and alimentary ducts, in preventing the smothering of respiratory surfaces. Blood flow in the gills of cyclostomes, as in those of bony fishes, is in a direction counter to that of water flow—an arrangement that increases the efficiency of gas exchange across the respiratory surface (Figure 4).

Sharks, rays, and bony fishes

Cartilaginous fishes (sharks and rays) and bony fishes employ a double-pumping mechanism to maintain a relatively constant flow of water over the gill exchange surfaces. The mouth, or buccal, cavity is employed as a positive-pressure pump and the gill cavity as a suction pump. Depression of the buccal cavity produces a pressure below that of the surrounding water, thus promoting flow into the cavity. The gill flaps are closed during this phase of respiration. The gill cavity expands as the opercular muscles relax, and the pressure on the back side of the gill filaments is reduced below the pressure on the buccal cavity. This pressure differential produces flow across the gill surfaces. Following this initial inspiratory phase, the buccal floor is actively elevated, thus creating a positive pressure within the mouth cavity. This produces accelerated flow to the gill cavity, in which the pressure is still negative. As pressure in the gill cavity rises above that of the external medium, the gill flaps are forced open, and water is discharged. This sequence of events has been worked out from studies of the pressures in the buccal (mouth) and opercular (gill) cavities during various phases of the respiratory cycle. In sharks and rays, a small forward gill slit, the spiracle, also provides a channel of water flow into the gill chamber. The importance of the spiracle seems to be correlated with habitat: bottom-dwelling forms (rays and skates) have relatively larger spiracles, and the major portion of the water flow passes through them rather than through the downward-oriented mouth.

Reprinted with permission of the MacMillan Company from *Animal Function*, by M.S. Gordon. Copyright © 1968 by Malcolm S. Gordon

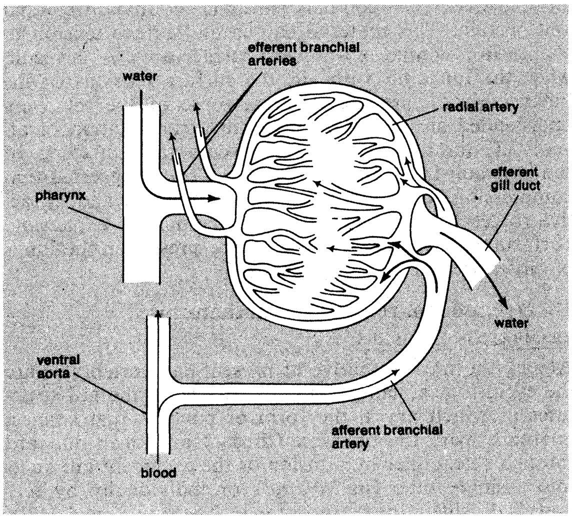


Figure 4: Blood and water flow through gill body of hagfish. Note that water flow is counter to blood flow, an arrangement facilitating gas exchange.

The pumping mechanism is not the only method of ventilation; sharks have been observed to keep both mouth and gill flaps open while swimming, assuring a constant water flow across the gill surfaces. When they slow down or settle to the bottom, the pumping activity is resumed. Tunas and mackerel cannot stop swimming: they have no active respiratory mechanism and are dependent for their gill ventilation on the current that results from their forward motion through the water.

A number of fishes depend in varying degree on aerial respiration. The ability to breathe air enables them to live in places where the oxygen content of water may be low or nil. Two general means of acquiring oxygen are employed. Some fishes stay near the surface of the water where the oxygen pressure resulting from surface diffusion is highest. Others have developed ancillary respiratory structures in the pharynx or the stomach; the gulping of air at the surface is a means of charging these respiratory surfaces (such as the pharyngeal epithelium in Electrophorus or the stomach in Plecostomus); the frequency with which these fishes rise to the surface to gulp air corresponds to their current need for oxygen.

The swamp-dwelling yarrow of Guyana (Erythrinus) uses both aquatic and aerial respiration, varying them according to the gaseous composition of the water. When the oxygen content is low, respiration through the gills ceases; when the oxygen content of the water is high, the fish relies primarily upon its gills except when the carbon dioxide content is also high—when, again, aerial respiration predominates. In other conditions it uses both modes of respiration. This apparently extends the range of conditions in which Erythrinus can survive.

Eels and lungfishes

Eels (*Anguilla*) use their skin as a major respiratory surface in addition to their gills. In water, about 15 percent of their oxygen uptake is across the skin, and this rises to around 50 percent when in air. They are capable of making extensive overland migrations during which, in the first few hours, they draw upon oxygen in the swim bladder. Like most fishes, eels when out of water exhibit a reduced heart rate and less oxygen consumption. When they return to water, their heart rate rises, and both oxygen consumption and blood lactic-acid levels rise. Lactic-acid production results from metabolism without oxygen, and such acid products must themselves be metabolized through higher oxygen consumption. Such patterns have been observed in grunions of the California coast that come ashore to breed and even in flying fishes during their brief aerial excursions.

Lungfishes, which have fossil relatives as far back as Devonian times, are considered closely related to the early amphibians. They extract oxygen from outpouchings of the gut that are biologically similar to the lungs of the higher vertebrates. Lepidosiren, the South American lungfish, has a poorly developed gill circulation and suffocates if denied access to air. But the African form Protopterus is less dependent upon aerial respiration, and the Australian lungfish Neoceratodus has well-developed gills and cannot survive out of water for long periods. Lungfishes inflate their lungs by rising to the surface and gulping a large bubble of air that they force into the lung by compressing the mouth and throat region. The mechanism is thought to be similar to that in many amphibians. Each breath is accompanied by increased heart activity and by a regional shift in blood flow from the body circuit to the lungs. The capacity to maintain a relatively efficient double circulation is an important evolutionary step toward a terrestrial life.

During long periods of drought, both Protopterus and Lepidosiren enter into a state of aestivation, during which metabolism and respiration fall to a low level. *Protopterus* secretes a protective mucous tube around itself with an opening entering the oral cavity. After it builds this cocoon, its oxygen consumption may fall to 10 percent of normal, its respiratory rate to one or two cycles per minute, and its heartbeat to around three per minute. It can survive in this state up to five years.

The bowfin, Amia *calva*, possesses both gills and an air bladder that may be used for respiration. It is almost exclusively a water breather at 10° C (50° F), a temperature at which it shows low physical activity. Its air-breathing rate increases with temperature and activity, and, at around 30° C (86" F), it draws about three times as much oxygen from air as from water. As in lungfishes, carbon dioxide elimination is predominantly across the gills. The bowfin's air-breathing frequency varies inversely with the oxygen content of the water; when oxygen tensions in water decline below 40 or 50 millimetres of mercury at 20° C (68" F), air breathing largely replaces water breathing. When an exchange surface (gill or air bladder) is not being utilized as the primary oxygen-exchange site, there is a tendency for blood to bypass it.

The so-called electric eel of South America (*Electrophorus electricus)* inhabits muddy streams that may become severely oxygen deficient. It is an obligatory air breather that depends upon the exchange of oxygen across the membranes of its mouth, expelling the air through its gill slits. Its blood has a high percentage of red corpuscles, is high in hemoglobin, and has an oxygen-absorbing capacity similar to that of mammals: Carbon dioxide elimination is primarily across the skin and, to a lesser extent, through the vestigial gills.

RESPIRATION IN AMPHIBIANS AND TERRESTRIAL VERTEBRATES

Amphibians. Frogs, toads, salamanders, and caecilians show a varying degree of dependence upon the aquatic habitat. Respiratory surfaces include the skin, the mouth cavity, and the lungs, the dependence on each varying within different groups and with the habitat and the season. The lungless salamanders (Plethodontidae) respire across the skin and to a smaller degree across the membranes of the mouth. Simultaneous use of mouth, skin, and lung respiration occurs in frogs. There is a well-developed blood supply to the lungs in all lunged amphibians; among frogs and toads (Anura) the pulmonary artery supplies not only the lungs but portions of the skin. Blood returns from the lungs directly to the left atrium of the heart; the drainage from the skin combines with the general circulation of blood in the veins.

At some time in their life cycle, frogs use all of the respiratory surfaces seen in other amphibians. In their tadpole stage, they breathe with the skin and the gills. During metamorphosis, they lose their gills and develop lungs. The large tail fins of tadpoles contain blood vessels and are important respiratory structures, but, like the external gills, they are lost during metamorphosis.

The dominant mode of lung inflation in adult frogs is the mouth–throat pumping mechanism seen in lungfishes. The seauence of respiratory movements is shown in Figure 5. The gas sampling probe leading to the respiratory mass spectrometer (a measuring device) is supported by a plastic collar in the rubber mask. Just preceding an expiration, the floor of the mouth is depressed (A), and air is drawn in through the nostrils to the floor of the mouth, where it remains during expiration. The expiratory act (B) is driven by muscles of the body wall and the elastic recoil of the lungs; during this phase the glottis is open and a high-velocity jet of air passes out through the nostrils, without mixing with that inhaled into the floor of the mouth. In the next phase (C), the floor of the mouth is elevated, creating a positive pressure in the mouth cavity and making air flow into the lungs through the open glottis. During this phase the nostrils are closed. After inflation a sequence of pumping actions follows with the glottis closed and the nostrils open (D). This serves to wash out residual expired gas and recharges the buccopharyngeal (mouth–throat) cavity with air.

From C. Gans et *al., Science,* vol. 163, pp. 1123–1225 (March 14, 1969); Copyright 1969 by the American Association for the Advancement oi Science

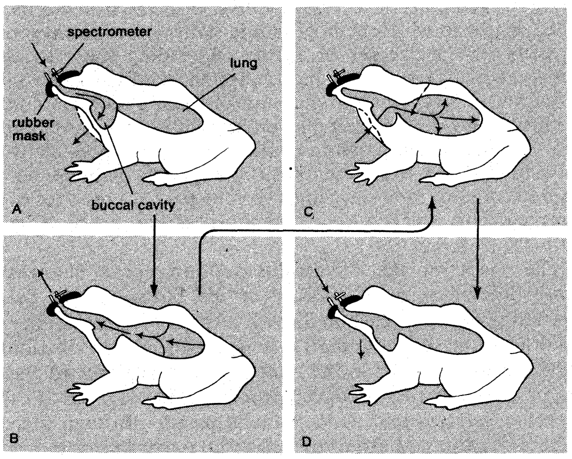


Figure 5: Successive stages of flow during the ventilatory cycle in the American bullfrog, Rana catesbeiana, The arrows indicate the approxirnate sequence of gas flow into and out of buccal cavity and lung (see text).

In frogs, the skin of the back and thighs (the areas exposed to air) contains a richer capillary network than the skin of the underparts. The capillary surface area of the mouth membranes is small compared with that of other respiratory surfaces. The aquatic newt ***Triton*** utilizes both lung and skin respiration, the skin containing about 75 percent of the respiratory capillaries. At the other extreme, the tree frog ***Hyla arborea*** is much less aquatic, and its lungs contain over 75 percent of the respiratory capillary surface area. similar-differences are found even in closely related forms: in the relatively more terrestrial frog, ***Rana** temporaria,* uptake of oxygen across the lung is about three times greater than across the skin; in ***Rana** esculenta,* which is more restricted to water, the lungs are about equal with the skin in the uptake of oxygen. Carbon dioxide is excreted mainly through the skin in both these forms; in fact, the skin appears to be a major avenue for carbon dioxide exchange in amphibians generally.

Reptiles. In becoming land animals, the reptiles had to develop a skin relatively impermeable to water and hence not well suited for respiration. This meant an almost complete dependence upon the lungs. The principal mechanism for lung inflation among reptiles is a suction pump. The power is derived from a muscular expansion of the rib cage and body wall, creating a subatmospheric pressure within the lungs and causing air to flow in.

The respiratory frequency of most reptiles is not regular, the pattern usually consisting of a series of active inspirations and expirations followed by relatively long pauses. As the inspiratory phase ends, the glottis closes, and there is a pause during which the lungs remain inflated. The respiratory muscles relax during this period. As a consequence of their relaxation and of the potential energy stored within the body wall and the stretched lung, the pressures within the abdomen and lungs rise. Expiration may then be largely passive, following upon the opening of the glottis.

Specialized muscles in turtles

The adoption of a rigid shell by turtles necessitated the development of highly specialized muscles to inflate the lungs, which are situated in a frame located between the internal organs and the domed upper shell. In the tortoise ***Testudo graeca,*** lung ventilation is accomplished by changing the volume of the body cavity. Expiration is brought about by the activity of muscles that draw the shoulder girdle back into the shell, compressing the abdominal viscera. The increased pressure in the body cavity is transmitted to the lungs. Inspiration involves opposite muscular actions that produce an increase in the volume of the body cavity and a reduction of lung pressure below that of the atmosphere. Because of the rigidity of the shell, it is not possible to use the potential energy of abdominal wall structures as in other reptiles, and hence for the tortoise both expiration and inspiration are active energy-consuming events.

Birds. Birds must be capable of high rates of gas exchange because their oxygen consumption at rest is relatively high, and it increases severalfold during flight. The gas volume of the bird lung is small compared with that of mammals, but the lung is connected to voluminous air sacs by a series of tubes, making the total volume of the respiratory system about twice that of mammals of comparable size (see Figure 6). The trachea divides into primary bronchi, each of which passes through a lung and onward to the paired abdominal air sacs; they also give rise to secondary bronchi supplying the other air sacs. Tertiary bronchi penetrate the lung mass, and, from

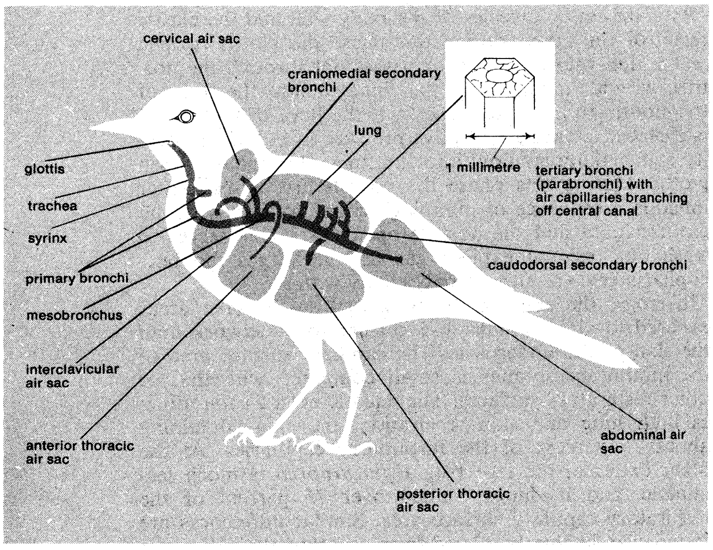

cervical air sac
craniomedial secondary bronchi
lung
1 millimetre
tertiary bronchi (parabronchi) with air capillaries branching off central canal
glottis
trachea
syrinx
primary bronchi
caudodorsal secondary bronchi
mesobronchus
interclavicular air sac
abdominal air sac
anterior thoracic air sac
posterior thoracic air sac

Figure 6: Respiratory system of a bird.

the walls of the tertiary bronchi, rather fine air capillaries arise. These have a large surface area; their walls contain blood capillaries connected with the heart. Gas exchange takes place between the air capillaries and blood capillaries. This surface is analogous to the alveolar surface in mammals.

Complex breathing of birds

The lungs are inflated by enlargement of the chest and abdominal cavity. The sternum (breastbone) swings forward and downward, while the ribs and chest wall move laterally. The pathway of airflow within the lungs and air sacs is not clear. A probable airflow scheme has been developed by inserting small devices into various parts of the airway. Airflow through the lung occurs during both inspiration and expiration, but in expiration it comes from the posterior air sacs. Apparently the anterior air sacs receive the gas after it has traversed the lung during the inflation phase. Fresh air comes into the lung by way of the primary and caudal secondary bronchi. The manner in which gas flow is controlled is not yet clear.

The lungs of birds undergo more complete ventilation than do those of mammals. They show a relatively low partial pressure of carbon dioxide. The ventilation of pigeons increases around 20-fold during flight, brought about by more rapid breathing and not by taking in more air at a breath. There is a precise synchrony between breathing and wing motion: the peak of expiration occurs at the downstroke of the wingbeat. The pigeon's in-flight ventilation is about two and one-half times that needed to support metabolism; around 17 percent of the heat production during flight is lost through evaporative cooling, suggesting that the excess ventilation is for regulating body heat. Studies of evening grosbeaks and ring-billed gulls show that their ventilation, in contrast to that of pigeons, increases in proportion to oxygen consumption. The increased ventilation in these birds is brought about by deeper as well as by more rapid breathing.

The respiratory system of birds is also used for communication through song. The "voice box" is the syrinx, a membranous structure at the lower end of the trachea. Sound is produced only when air flows outward across the syrinx. In canaries, notes or pulses are synchronous with chest movements; the trills, however, are made with a series of shallow breaths. The song of many small birds is of long duration relative to their breathing frequencies.

Mammals. Breathing in mammals is powered by a negative-pressure pump. Expansion of the chest lowers the pressure between the lungs and the chest wall, as well as the pressure within the lungs. This causes atmospheric air to flow into the lungs. The chief muscles of inspiration are the diaphragm and the external intercostal muscles. The diaphragm is a domelike sheet of muscle separating the abdominal and chest cavities that moves downward as it contracts. The downward motion enlarges the chest cavity and depresses the organs below. As the external intercostal muscles contract, the ribs rotate upward and laterally, increasing the chest circumference. During severe exercise other muscles may also be used. Inspiration ends with the closing of the glottis.

In expiration, the glottis opens, and the inspiratory muscles relax; the stored energy of the chest wall and lungs generates the motive power for expiration. During exercise or when respiration is laboured, the internal intercostal muscles and the abdominal muscles are activated. The internal intercostals produce a depression of the rib cage and a decrease in chest circumference.

The membranes surrounding the lungs (visceral pleura) and the membranes of the chest wall (parietal pleura) tend to be held together by the adhesive forces of the fluids between the two membranes. When lung pressure falls below atmospheric pressure, air enters the lungs. A still greater subatmospheric pressure develops in the space between the parietal and visceral pleura, of the order of —6 millimetres of mercury during quiet inspiration. During expiration this pressure rises to about —4 millimetres of mercury. Lung pressure at the beginning of inspiration is subatmospheric (—1 to —2 millimetres of mercury), rising to atmospheric pressure at the close of inspiration; during quiet expiration, lung pressures reach about three millimetres of mercury, returning to atmospheric pressure toward the close. Pressure fluctuations of great magnitude may occur with laboured respiration: intrapleural inspiratory pressure may be as low as —30 millimetres of mercury, while lung pressure may approach 100 millimetres of mercury when active expiratory efforts are made against a closed glottis. The pressure fluctuations within the chest cavity have an important influence on the flow of blood in the large veins that return blood to the heart. Negative pressures tend to create a more favourable gradient for blood flow into the chest cavity, whereas positive pressures tend to impede return of blood to the heart.

GAS TRANSPORT

Respiratory gases move between the air and the blood across the respiratory exchange surfaces in the lungs. The gases are transported in the blood mainly by the red blood cells, which make up about 40 percent of the volume of the blood in most mammals.

Hemoglobin

Transport of oxygen. Most of the oxygen is carried in association with the pigment hemoglobin, which is found in many unrelated groups of animals, including all classes of vertebrates. Other respiratory pigments found in animals include the green chlorocruorins (found in a number of polychaete worms), the violet hemerythrins (found in corpuscles of certain polychaete worms and brachiopods), and the copper-containing hemocyanins (in some arthropods and mollusks).

Hemoglobins have been the most thoroughly studied of the respiratory pigments. The hemoglobins of vertebrates consist of four iron-containing molecules (hemes) chemically bonded to a large "carrier" protein called the globin. Oxygen joins in reversible union with the iron atoms. The degree to which hemoglobin is saturated with oxygen depends on a number of factors, including the partial pressure of oxygen in the blood plasma, the hydrogen-ion concentration, the partial pressure of carbon dioxide, and temperature. By keeping the other factors constant and varying the partial pressure of the oxygen, it is possible to construct a curve representing the capacity of hemoglobin to combine with oxygen at various partial pressures of oxygen. This is called the oxygen-dissociation curve (Figure 7).

The shape of the oxygen-dissociation curve changes with differences in the acidity of the blood and in the partial pressure of carbon dioxide. An increase in acidity and an increase in the partial pressure of carbon dioxide both produce a shift in the dissociation curve toward the right. Active tissues produce carbon dioxide as a result of metabolism, and the capillary blood passing through such tissues is exposed both to this elevated partial pressure of carbon dioxide and to a greater acidity. The shift of the oxygen-dissociation curve to the right shows how this facilitates the unloading of oxygen in the metabolizing

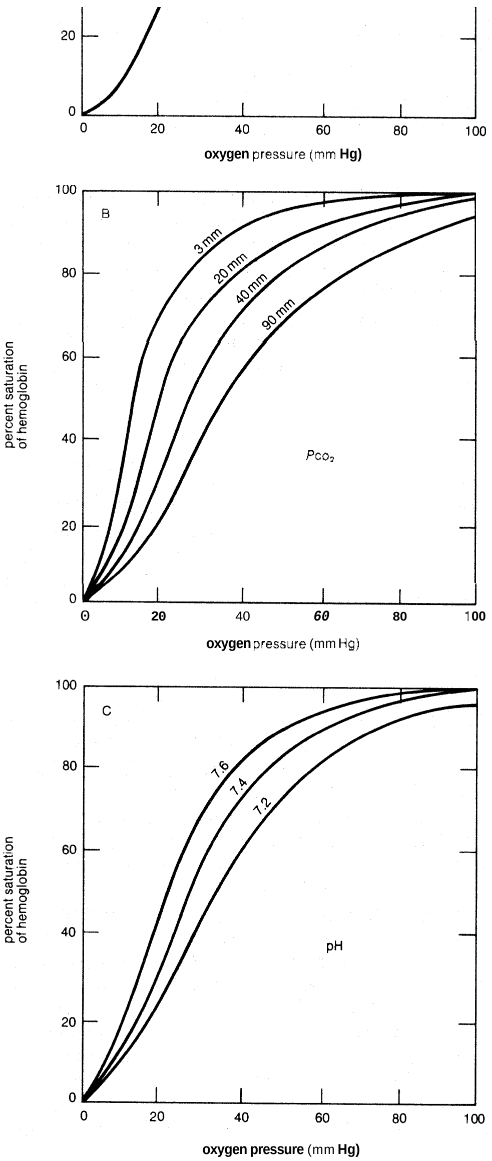


Figure 7: The combination of hemoglobin with oxygen. (A)Oxygen dissociation curve. (B,C) Effects of variation in P_{CO_2} and pH on the dissociation curve (see text).

Reprinted with permission of the Macmillan Company from Animal Function, by M.S. Gordon. Copyright © 1968 by Malcolm S. Gordon

tissues. Since the tissue cells consume oxygen, there is a favourable gradient for diffusion of oxygen from the capillary blood to the tissue cells. In short, the combination of high partial pressure of carbon dioxide and low oxygen pressure in the tissues facilitates the unloading and delivery of oxygen at the site where it is consumed. The steep part of the dissociation curve in the range of 40 millimetres of mercury for the partial pressure of oxygen, close to that found in many tissues, is of great physiological interest: it means that a relatively small decline in the partial pressure of oxygen is associated with a relatively large release of bound oxygen.

Temperature differences also affect the **oxygen-dissociation** curve. In so-called cold-blooded animals, the loading and unloading of oxygen can be carried on at adequate levels only between certain temperature limits. For example, even though octopus blood becomes fully saturated with oxygen at temperatures near 0° C (32" F), the animal will be starved of oxygen because significant unloading at the tissues cannot occur. At 20° *C* (68" F) human hemoglobin is saturated with oxygen, but little oxygen is delivered to the tissues—setting a natural limit to the possibility of lowering body temperature during certain surgical procedures (such as open-heart surgery). The lack of dissociation of oxyhemoglobin at low temperatures also accounts for the red colour of very cold ears.

The hemoglobin associated with many skeletal muscles (myoglobin) has a much higher oxygen affinity at low partial pressures of oxygen than does the circulating hemoglobin. Myoglobin contains only a single heme group. When a comparison of the dissociation curves of mammalian myoglobin and hemoglobin is made, it is found that at the normal partial pressure of oxygen found in venous blood (about 40 millimetres of mercury) the hemoglobin is about 60 percent saturated, whereas myoglobin is over 90 percent saturated. The dissociation curve of myoglobin is far to the left of that for hemoglobin and has a different shape. Since myoglobin unloads oxygen at low oxygen partial pressures, it has been suggested that this molecule facilitates transport of oxygen from the hemoglobin to the intracellular site, where those partial pressures are normally much lower than that associated with the unloading of hemoglobin.

Buffering of hydrogen ions by hemoglobin

Transport of carbon dioxide. Metabolizing tissues produce carbon dioxide, which must be transported to the lungs, gills, or skin where it is exchanged for oxygen. Most of the carbon dioxide (CO_2) transported by the blood is found in reversible chemical combinations in plasma or red blood cells. As carbon dioxide enters the plasma from tissues, it combines with water to yield carbonic acid (CO_2 .+ H_2O e H_2CO_3). This reaction is followed by the dissociation of carbonic acid into hydrogen and bicarbonate ions (H_2CO_3 + H^+ + HCO_3^-). This would result in greater acidification of the plasma (free H+) were it not for the presence of buffering agents such as protein (A–) that combine with the free hydrogen ion (H+ + *A*- + HA). A buffer solution resists change in acidity by combining with added hydrogen ions and, essentially, inactivating them. The formation of carbonic acid from carbon dioxide and water is much too slow to account for the actual rate at which carbon dioxide is taken up by the plasma. The actual rate is explained by the buffering effect of the hemoglobin: as hemoglobin gives up its oxygen to the tissues, it becomes less acid and can therefore take up most of the hydrogen ions. About 80 percent of the increased carbon dioxide content of the blood returning from the tissues is accounted for by the buffering capacity of hemoglobin. The role of hemoglobin buffering in this set of reactions may be summarized as follows: (1) carbon dioxide enters the red blood cell from the plasma by diffusion; (2) once in the red cell, it combines with water to form carbonic acid in the same way as in the plasma. In the red cell, this reaction is extraordinarily rapid because it is speeded up by an enzyme, carbonic anhydrase; (3) carbonic acid dissociates into hydrogen ions and bicarbonate ions; (4) hemoglobin, after giving up oxygen, enters into a complex with the hydrogen ions. The last reaction is of great importance in stabilizing the acidity of the red blood cell.

As carbon dioxide enters the red blood cell, it is accompanied by a shift of the chloride ion, C1-, from the plasma to the interior of the cell. Bicarbonate ions are formed at a very high rate because of the presence of carbonic acid, and they diffuse into the plasma. This inward diffusion of chloride in exchange for bicarbonate maintains the electrical balance across the red-cell membrane, a phenomenon known as the chloride shift.

Hemoglobin also acts in another way to facilitate the transport of carbon dioxide. Amino groups of the hemoglobin molecule react reversibly with carbon dioxide in solution to yield a carbamino compound. About 20 percent of the carbon dioxide is transported in the form of carbamino hemoglobin, about 5 percent in physical solution, and the remaining 75 percent in the form of bicarbonate ions. A reverse sequence of reactions occurs during oxygenation in the lungs, during which the carbon dioxide is unloaded into the alveolar gas to be expired.

Blood pigments and environment. The oxygen-dissociation curves vary widely among different kinds of animals. These variations appear to be related to differences in the chemical structures of the blood pigments. Curves to the extreme right of the range indicate pigments that require relatively high oxygen tensions for loading. Animals possessing such pigments would not flourish in zones of low oxygen tension. Vertebrate animals shows rather striking correlations between their distribution and their oxygen-dissociation curves.

Hemocyanin in mollusks

Among invertebrates, the cephalopod mollusks employ hemocyanin as an oxygen carrier. In the squid *Loligo* and in *Octopus,* the oxygen content of arterial blood is about 25 percent of that found in man. The extraction of oxygen by the tissues, however, is approximately 92 percent, or more than three times the percentage in man. Although such a system transfers a large amount of oxygen to the tissues, it fails to provide a reserve against oxygen scarcity. The oxygen-dissociation curves are affected by carbon dioxide, increased acidity, and increased temperature in much the same manner as in animals with hemoglobin. The cuttlefish of the genus *Sepia* and *Octopus* exhibit oxygen-dissociation curves to the left of that found in squid—a characteristic that correlates with their capacity to survive in waters of lower oxygen concentration.

Comparison of the dissociation curves of fishes reveals that those forms that are inhabitants of water containing a high partial pressure of oxygen (such as trout) exhibit curves to the right of forms living in stagnant waters (carp, catfish). The affinity of hemoglobin for oxygen is thus an important determinant of environmental distributions. A comparison of the behaviour of hemoglobins of trout and carp at 15° C (59" F) and a partial pressure of carbon dioxide of two millimetres of mercury reveals that it requires a partial pressure of oxygen of about five millimetres of mercury to bring about 50 percent saturation of the hemoglobin of the carp, as compared to 18 millimetres of mercury for the trout. Carp can range into waters of low partial pressure of oxygen denied to trout because of the loading characteristics of their hemoglobin. Furthermore, the effect of elevated carbon dioxide pressure is less on the blood of stagnant-water forms than on that of those inhabiting highly oxygenated waters. An elevation in the partial pressure of carbon dioxide from one to ten millimetres of mercury induces a shift to the right in the dissociation curve of the trout, such that only partial saturation with oxygen may be attained, and the fish may suffocate even in an abundance of oxygen.

Comparisons of a variety of animals inhabiting zones of varying oxygen availability indicate that aquatic forms saturate their hemoglobin with oxygen at lower oxygen pressures than do those that live on land. The greater oxygen affinity of their hemoglobins allows extraction of oxygen from environments poor in oxygen relative to air. The more terrestrial amphibians and reptiles require much higher environmental oxygen pressure than do the aquatic forms.

Mammals native to high altitudes (such as the llama and the vicuña) have hemoglobins of greater oxygen affinity than do low-altitude forms. Their hematocrits (percentages of red cells in the blood) and red-blood-cell counts are similar to those of their sea-level relatives. When low-altitude animals ascend to high altitudes, there are increases in their hematocrits, red-cell counts, and total hemoglobin. The net effect is an increase in oxygen capacity, although the dissociation curve does not shift appreciably. Similar responses have been observed in fishes; goldfish exposed to reduced oxygen pressures increase their hemoglobin levels. When mammals are exposed to low-oxygen conditions, they release a hormone, erythropoietin, that accelerates the production of red blood cells and hemoglobin.

Fetal hemoglobin

Similar differences appear in the fetal stage of vertebrates. The oxygen-dissociation curve of the fetus in both sheep and man is considerably to the left of that seen in the adult. This allows for a more efficient transferral of oxygen from the maternal to the fetal blood across the placenta. After birth the fetal hemoglobin is replaced by the adult pigment. Viviparous fishes such as the dogfish shark exhibit fetal hemoglobins having greater oxygen affinity than the adults, and similar pigments have been discovered in viviparous snakes. The dissociation curve of the frog tadpole is to the left of that of the adult, a relationship that seems appropriate in the transition from an aquatic to **a** terrestrial life. The hemoglobin of birds also exhibits a high oxygen affinity during the fetal period.

III. The control of respiration

Because respiration involves the acquisition of oxygen and the elimination of carbon dioxide, it is reasonable to suppose that respiratory activity must be controlled by mechanisms that adjust external respiration to changes in the concentrations of these gases. This was suspected as long ago as 1796, when it was speculated that the increase in respiratory activity at high altitudes was a response to the low partial pressure of oxygen. By the latter part of the 19th century, it had been demonstrated that low oxygen and elevated carbon dioxide pressures were respiratory stimulants. Such observations formed the framework for most of the subsequent work on respiratory control.

NEURALREFLEXES

The ancient Greek physician Galen knew that cutting the spinal cord at a level near the brain brought about a cessation of respiratory activity. A French physiologist, J.J.C. Legallois, located the "respiration centres" in 1812 by demonstrating that cutting the brainstem just behind a mass of nerve fibres called the pons did not abolish respiration but that, as the transections were moved backward through the medulla a few rnillimetres at a time, respiration finally ceased. These observations indicated that the medulla is a major central-nervous-system area in the control of respiration. An English physiologist, Marshall Hall, demonstrated in 1850 that the vagus nerves were involved in reflexes that modify respiration. The observation that cutting off the blood supply to the brain induced respiratory stimulation led other workers to suggest that the chemical composition of the blood circulating in the brain had an important influence on respiration. Taken together, these early observations suggested that respiration adjusts to changes in the environmental-gas composition or to chemical alterations of the body fluids; that the central nervous system is essential for the control of respiration; and that chemical alterations may either directly or by reflex modify the movements of respiration.

Two types of respiratory neurons have been located in the lower portion of the brain (the medulla oblongata) in all vertebrates. One set of neurons activates inspiratory activity; the other activates expiratory activity. Because electrical stimulation of either the expiratory or the inspiratory neurons produces an appropriate respiratory act, these neurons are sometimes thought of as "centres." It is not likely that discrete centres exist, and in fact there is considerable intermingling and overlap of these two neuronal types. When the sensory nerves coming to the respiratory areas are eliminated, it is still possible to produce alternating bursts of expiratory and inspiratory activity in the medulla. The system behaves as if mutual inhibitory and excitatory nerve links exist between the two sets of neurons. As the inspiratory neurons discharge, there is an increase in neural activity not only to the muscles of inspiration but also to the expiratory neurons; their discharge excites the expiratory muscles and also activates inhibitory pathways to the inspiratory centre; inhibition of the inspiratory neurons then decreases the excitation going to the expiratory neurons, leading to a diminution in expiratory-discharge rate and the removal of inhibition of the inspiratory areas. Although speculative, this system is in accord with observed facts. Whatever the case, respiratory activity is not solely under the control of the medulla, since the medullary site is also influenced by sensory input from the body as well as from connections in the higher centres of the brain. Sensory stimuli such as pain, heat, and cold applied to various body areas modify respiratory activity in higher vertebrates; visual or olfactory stimuli alter the respiratory rhythm of fishes. In mammals, at least two additional centres in the area of the pons have important respiratory connections. Experiments in which the brainstem is sectioned at various levels show that, although the medullary centre is capable of producing sequences of respiratory inspiration and expiration, these are of an abnormal character when the medulla is disconnected from higher levels of the brain. In the lower pons, an area known as the apneustic centre governs cessation of respiration during inspiration. When disconnected from the medulla, prolonged inspiratory spasm may occur. In the higher pons, the pneumotaxic centre functions to restrain periodically the apneustic centre; furthermore, fibres in the vagus nerve converge on the apneustic centre to produce inhibition.

MUSCULAR FEEDBACK

Stretch receptors

The behaviour of the respiratory centres in the brain is also modified by neural feedback from various parts of the body. The lungs send sensory information to the brain that alters the activity of the respiratory centres. This type of relationship was studied experimentally in 1868 in Vienna by two physiologists, Ewald Hering and Josef Breuer; they found that, if the airways were blocked following inspiration, the activity of the expiratory muscles (acting against the blocked airways) was prolonged. Moreover, subsequent inspiratory activity was prolonged far beyond the normal length of time. Clamping of the airways after expiration was followed by a prolongation of inspiratory activity. Taken together, their observations indicated that expiratory and inspiratory neural activity was linked to the degree of chest expansion. When they severed the vagus nerve to the brain, they obtained a deeper and slower ventilation pattern; blockade of the airways at the end of inspiration was then followed by expiratory effort of a duration similar to that seen prior to cutting the vagus nerve. It was concluded from these observations that, during inspiration, mechanical receptors were stimulated, which, through their linkage via the vagus nerve to the respiratory centres, augmented the expiratory act. This reflex is known as the Hering-Breuer reflex. The mechanical receptors are in the walls of the bronchioles and their branches. During inspiration these receptors are stretched as the lung is filled; the electrical activity of the vagus nerve has been shown to increase as the lung volume is increased. The receptors differ in their sensitivity to stretch: at the end of a normal inspiration, many of them are stimulated, but others remain quiescent unless further inflation occurs. Other vagal pathways are activated at the end of an expiration. Their activation terminates expiration at a point earlier than normal, as would occur during the deep breathing associated with exercise. During quiet respiration these receptors are apparently not brought into play, but, as breathing volume increases, they institute, bv their action on the respiratory centres, strong inspiratory activity. It requires greater muscular effort to increase the inspiratory volume above a certain level, this effort increasing disproportionately with changes in volume. By setting a limit on inspiratory volume during quiet breathing, respiration is maintained at a more efficient level. Although the inflation reflexes described above were discovered in mammals, they have also been found in birds. In the dogfish shark, intiation of the pharynx inhibits inspiratory activity, an effect that disappears after cutting the spiracular branches of nerves VII, IX, and X. The receptors involved in this reflex are located at the junction of the pharynx and the gill pouches.

Other stretch receptors exist in the skeletal muscles that provide the motive power for respiration, including the diaphragm. (Figure 8 shows the organization of these receptors and their relationship to the main muscle mass.) These are similar to the receptors that produce the knee-jerk reflex. Contraction of the extrafusal muscles powers inspiration. These muscles may be activated as a result of annulospiral receptor activity; the receptors discharge when stretched by intrafusal fibres. This produces a train of impulses over the nerves that run to the spinal cord and interconnect with motor nerves to the muscle; motor-nerve activity is set off that causes the muscle to contract and shorten; secondarily this produces shortening of the annulospiral stretch receptors in the muscle and reduces the frequency of the impulses they send to the spinal cord. Recent evidence suggests that these stretch receptors have muscle fibres (intrafusal fibres) at each end in series with the receptor. The complex of annulospiral endings and associated intrafusal fibres is in parallel with the major muscle fibres (extrafusal fibres). Two types of motor neurons have been differentiated: gamma neurons, which innervate the intrafusal fibres, and alpha neurons, which innervate the major muscle mass. Activation of gamma neurons, producing contraction of the intrafusal fibres, has the effect of stretching the annulospiral endings and thus producing impulses that reach the spinal cord and there synapse with the alpha motor neurons.

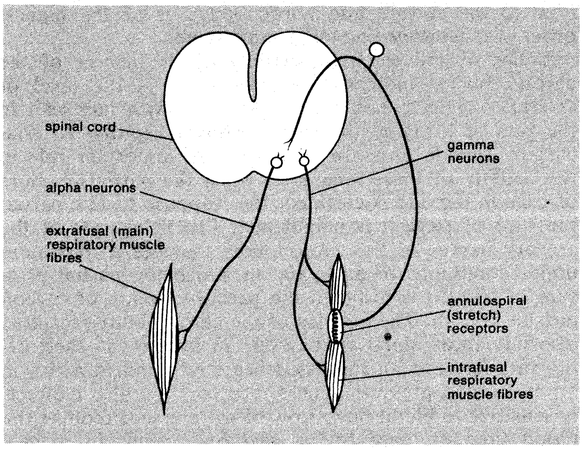

spinal cord
gamma neurons
alpha neurons
extrafusal (main) respiratory muscle fibres
annulospiral (stretch) receptors
intrafusal respiratory muscle fibres

Figure 8: Control of respiratory muscle shortening by stretch receptor mechanism (see text).

The gamma neurons in the respiratory muscles are thought to receive impulses arising in the respiratory centres. The magnitude of gamma activation thus depends upon the intensity of discharge of the respiratory centres in the brain, and the intensity of respiratory-centre discharge is dependent upon the type of sensory information brought in from chemoreceptors and stretch receptors. In order for the annulospiral sensory endings to reduce their discharge rate, the major muscle mass must now shorten, to the extent that the stretch induced by intrafusal contraction is now reduced. This system would appear to give the muscles of respiration a means of matching their activity with the sensory information being received; shortening of the intrafusal fibres would govern the extent of extrafusal-fibre activity necessary to attain a desired breathing volume.

CHEMICALLY SENSITIVE CONTROLS

Breathing is also controlled by specialized nerve cells that are sensitive to chemical changes in the blood. The carotid arteries, which convey the major blood supply to the head and brain, branch near the skull into internal and external carotid arteries. Small vascular spherules, the carotid bodies, are located at this point. The Belgian physiologist Comeille Heymans received the Nobel Prize in 1938 for his demonstration that the carotid bodies and similar structures on the aorta (the artery arising from the heart) are chemosensitive. When he perfused these bodies with blood of low partial pressure of oxygen, he observed increases in the volume and frequency of respiration. In mammals, the nerve fibres from the carotid bodies (cranial nerve IX) and aortic bodies (nerve X) form the afferent limb of a chemically induced reflex. There is some variation in other vertebrates. The blood

The carotid and aortic bodies

flow to the carotid and aortic bodies is of the highest order of magnitude known for any tissue.

Studies of the electrical activity of the nerves of the carotid bodies indicate that reductions in the level of arterial oxygen pressure are associated with increases in the rate of firing of the carotid-body receptors. Arterial carbon dioxide pressure also plays an important role in the activity of these receptors; when the partial pressure of carbon dioxide is elevated, the response to low partial pressure of oxygen is heightened. The firing rate of the sensory nerves of the carotid body reaches a maximum under conditions of asphyxia, in which the partial pressure of oxygen is low and the partial pressure of carbon dioxide is high. The activity of the carotid-body receptors also influences blood circulation; an increase in their firing rate leads to an increased heart rate and to a rise in arterial blood pressure. These receptors are also indirectly sensitive to blood flow. Low blood pressure reduces the blood flow to these bodies and may result in oxygen deficiency to them even in the presence of a normal arterial partial pressure of oxygen—apparently because of the high oxygen extraction associated with the high rate of metabolism of these bodies. A combination of low arterial partial pressure of oxygen and reduced blood flow is a more effective stimulus for the receptors than either is alone.

It had long been supposed that the respiratory centres of the brain were directly sensitive to small changes in partial pressure of carbon dioxide (P_{CO_2}), since perfusion of the brain with blood of very moderately elevated arterial partial pressure of carbon dioxide elicits increased ventilation. More recently it has been found that this response is brought about indirectly through the cerebrospinal fluid. Areas on the surface of the medulla that are normally bathed with cerebrospinal fluid have been shown to be sensitive to solutions of high partial pressure of carbon dioxide or elevated hydrogen-ion concentration; these receptors apparently have connections with the respiratory centres. Cerebrospinal fluid is characterized by a low protein content compared with that of blood and thus a low capacity for absorbing hydrogen ions. Carbon dioxide diffuses rapidly from blood and interstitial fluid to the cerebrospinal fluid; the medullary receptors are well situated for detecting slight changes in hydrogen-ion concentration. It is likely that this sensitivity provides for a critical monitoring of blood partial pressure of carbon dioxide and cerebrospinal-fluid acidity. An elevation in arterial partial pressure of carbon dioxide, through its effects on cerebrospinal-fluid acidity, will lead to increased ventilation of the lungs and expiration of carbon dioxide. Representatives of all the major vertebrate groups respond to elevated partial pressure of carbon dioxide or low partial pressure of oxygen with increased ventilation.

IV. Adaptation to special conditions

ADAPTATION TO DIVING

Many air-breathing vertebrates are able to remain submerged in water for a long period of time (Table 2).

Table 2: Duration and Depth of Diving in Some Mammals

species	duration (in minutes)	depth (in metres)
Platypus	10	
Mink, ***Mustela vison***	3	
Harbour seal, ***Phoca vitulina***	20	
Walrus. ***Odobenus rosmarus***	10	***80***
Steller's sea lion, ***Eumetopias jubata***		146
Gray seal, ***Halichoerus grypus***	20	100
Weddell seal, ***Leptonychotes weddelli***	43	600
Bottle-nosed whale, ***Hyperoodon ampullatus***	120	deep
Sperm whale, ***Physeter catodon***	75	900
Blue whale, ***Sibbaldus musculus***	49	100
Harbour porpoise, ***Phocaena phocaena***	12	20
Bottle-nosed porpoise, ***Tursiops truncatus***	5	
Beaver, ***Castor canadensis***	15	
Muskrat, ***Ondatra zibethicus***	12	
Most men	1	
Experienced skin divers	2.5	61

Source: W. O. Fenn and H. Rahn (eds.), ***Handbook of Physiology*** (1964); and ***G.*** L. Kooyman, ***Science*** (1966).

One of the most striking physiological alterations observed in such diving animals is a marked reduction in the heart rate. This has been shown to result from the action of the vagus nerve.

Lactic-acid concentration in the blood may rise slightly during diving; immediately after emerging, it rises precipitously. This has been observed in a wide variety of divers, including ducks, the alligator, and many mammals. Lactic acid is produced by metabolism occurring in the absence of oxygen. During diving the blood flow is greatly reduced or temporarily eliminated in muscle, kidney, and intestinal tissue. Blood vessels going to many of the tissues are constricted, and, as a result, most of the body's oxygen stores are distributed toward the brain and the heart. The lactate produced during the dive is oxidized afterward and excreted.

The constriction of the blood vessels during diving is controlled by the sympathetic nervous system. Diving animals differ from others in that the sympathetic nerves terminate in the muscular walls of the major supply arteries. In nondiving mammals such as the cat, the sympathetic nerves terminate primarily in the arterioles, the blood vessels just prior to the capillaries where metabolic exchange takes place. During sympathetic stimulation, the reduction of blood flow in the cat is associated with an accumulation of local metabolic materials (vasodilator metabolites) that tend to diffuse to the arterioles, at which place these metabolic materials compete with the sympathetic nerves and ultimately cause the vessels to dilate. The diving animals avoid this by utilizing sympathetic nerves, which cause constriction of main supply arteries that are far enough away from the capillaries so as to be unaffected by vasodilator metabolites produced in the tissues.

Lung volumes

The lung volumes of diving animals do not appear to be different from those of nondiving forms, with the exception of whales, which have a smaller relative lung volume than nondivers. It has been suggested that the small lung volume of whales (about 50 percent less than other mammals per unit of weight) enables them to descend to great depths without danger of nitrogen bubbles forming in the blood (caisson disease). The lungs may actually collapse during descent to great depths. It has been calculated that, at a depth of 100 metres (330 feet), all of the lung gas could be contained in the nonexchange respiratory components (dead space), the exchange surfaces being collapsed. The effect of such a redistribution of the lung gases under pressure would be a decrease in nitrogen diffusion into the blood. On ascent, only a relatively small load of dissolved nitrogen would have to be disposed of, thus reducing the danger of bubble formation. Most diving animals actually exhale before or shortly after diving. Observation of trained porpoises with underwater television cameras has shown that the thorax becomes increasingly compressed as the animal reaches greater depth.

Turtles, despite their heavy shells, manoeuvre easily through the water, floating near the surface at one moment or quietly sinking to the bottom the next. In the red-eared turtle ***Pseudemys scripta***, this depends on the animal's ability to change the relative volumes of lung air and stored water. The water storage is done through the cloacal bursae, outpouchings from the terminal portion of the hindgut; the mechanism closely resembles that used in ballasting submarines.

The tidal volume (that is, the volume of air drawn in) of mammalian divers is greater than that of nondivers, and the lungs are also more completely emptied during expiration.

Many diving animals have a larger blood volume than nondivers. The relative blood volumes of the porpoise and seal are approximately twice that found in dog and man, and diving birds may have about a 60 percent greater blood volume than their nondiving relatives. Ducks seem to store considerable oxygen reserves in their venous system. Diving mammals may also store a considerable supply of oxygen in the form of oxymyoglobin. Myoglobin unloads oxygen at much lower tensions than does hemoglobin. It has been calculated that nearly 50 percent of the oxygen reserves of diving seals is in the form of oxymyoglobin.

A set of responses that may be called a "diving reflex" is found in all the vertebrate groups. Fish "dive" when removed from water, exhibiting a slowing of the heart rate. Reduced cardiac output and heart rate are also characteristic of diving frogs. The cardiac response to diving is exhibited even in nondivers such as man, although to an extent much smaller than in specialists such as seals, porpoises, or whales.

ADAPTATION TO HIGH ALTITUDES

Ascent from sea level to a high altitude has well-known effects upon respiration. The decline in barometric pressure is accompanied by a fall in the partial pressure of oxygen, and it is this that imposes the major respiratory challenge to man at higher altitudes. At moderate elevations, the decline in alveolar partial pressure of oxygen is partially offset by increased ventilation. On ascent to higher levels, however, the alveolar partial pressure of oxygen declines even more than does the ambient partial pressure of oxygen. Since the air entering the respiratory system is saturated with water vapour before reaching the alveolar exchange surfaces, the gases are "diluted" by the presence of water vapour (which has a partial pressure of around 47 millimetres of mercury [mm Hg] at body temperature). At 30,000 feet (9,000 metres), for example, the barometric pressure is 226 millimetres of mercury and the gases are partitioned in the alveoli as follows: $P_{CO_2} = 24$ mm Hg; $P_{N_2} = 134$ mm Hg; $P_{H_2O} = 47$ mm Hg; and $P_{O_2} = 21$ mm Hg. The ambient partial pressure of oxygen is around 47 millimetres of mercury at this altitude, and it is apparent that water vapour is responsible for diluting both oxygen and nitrogen in the lungs. Alveolar partial pressure of carbon dioxide remains at 24 millimetres of mercury at altitudes exceeding 20,000 feet (6,000 metres). This is because breathing approaches its upper limit, and therefore maximum ventilation capacity is reached at this altitude. Since carbon dioxide production is relatively constant, a relatively constant alveolar partial pressure of carbon dioxide results.

The decline in the partial pressure of oxygen is offset to some extent by greater ventilation. On first arriving at high altitude, the declining arterial oxygen saturation causes the chemoreceptors to induce greater respiratory activity; ventilation increases to about one and one-half times normal. This is far below the ventilation level associated with exercise at sea level. This limitation on ventilation is associated with lower partial pressure of carbon dioxide in the blood and decreasing acidity, which cause the receptors in the medulla to suppress respiration and limit the level of ventilation. Gradually the blood acidity is corrected toward normal by kidney mechanisms. As the blood becomes less alkalotic, the medullary receptors begin again to augment respiratory activity, and ventilation may increase around sevenfold.

Hemoglobin production at high altitudes

The scarcity of oxygen at high altitude stimulates increased production of hemoglobin and red blood cells. At altitudes of around 17,000 feet (5,000 metres), the red-cell count may ultimately rise from the sea-level value of 5,000,000 per cubic millimetre to 7,000,000, and hemoglobin levels may rise from 15 to 22 grams per 100 millilitres of blood. ·The proportion of red cells in the total blood supply rises from 40 percent to as high as 70 percent. The result is a marked increase in oxygen-carrying capacity.

Upon initial exposure to high altitude, the cardiac output at rest increases by 20 to 50 percent; as adjustments in oxygen-carrying capacity are made, the cardiac output returns to near-normal levels. Exercise produces a smaller increase in cardiac output at high altitude than it does at sea level. The higher blood viscosity because of the increase in red cells imposes an additional work load on the heart. It is thought that the pacemaker tissue of the heart may also be suppressed because less oxygen is available.

The flow of blood to the brain, heart, and skeletal muscles increases during the initial adjustments to high altitude, but the flow to the kidneys and skin is reduced because of constriction of the blood vessels. On longer exposure to low barometric pressure, the tissues develop more blood vessels, and, as capillary density is increased, the length of the diffusion path along which gases must pass is decreased—a factor augmenting gas exchange.

Men climbing to heights above 20,000 feet reach the upper limit of chemoreceptor-driven ventilation. As the oxygen saturation of the arterial blood falls to 50 percent, faintness or unconsciousness follow. The critical height for 50 percent saturation is about 23,000 feet (7,000 metres). Even at 12,000 feet (4,000 metres), hypoxia may be severe enough to induce mental fatigue, drowsiness, headache, or euphoria. The journals of mountain climbers offer interesting testimony to this fact: often the handwriting becomes unintelligible, and phrases are repeated or appear out of context. Night vision is impaired because of lack of oxygen at the retina. Above 23,000 feet, suppression of central nervous system functions leads to convulsions, coma, and death.

Mammals that live at high altitudes, such as the llama and vicuña, have hemoglobin of high oxygen affinity. The oxygen-dissociation curves of these animals are to the left of lower altitude mammals; *i.e.*, full saturation of the blood with oxygen occurs at lower partial pressures of oxygen. Although their hemoglobin levels are relatively high, the proportion of red blood cells to the total volume of blood is not: at a partial pressure of oxygen of 40 millimetres of mercury, the blood of man holds about 14 millilitres of oxygen to each 100 millilitres of blood, while that of the vicuiia holds 18. The hematocrit (proportion of blood represented by red blood cells) in the high-altitude cameloids does not vary much at various altitudes, and moreover, it is rather low (30 percent) by comparison with that of man at sea level (40 percent). These mammals of high altitude are adapted to low oxygen pressures by their possession of hemoglobin of high oxygen affinity.

BIBLIOGRAPHY. JULIUS H. COMROE, *Physiology of Respiration* (1965), covers the basic aspects of respiration in mammals. PIERRE DEJOURS, *Physiologie*, vol. 3, ch. 1 (1963; Eng. trans., *Respiration*, 1966), is a concisely written introduction (for medical students) to mammalian and human respiration. ERNEST FLOREY, *An Introduction to General and Comparative Animal Physiology* (1966), contains sections on gases in the environment and on how various animals, invertebrates as well as vertebrates, extract and transport oxygen and carbon dioxide. M.S. GORDON *et al.*, *Animal Function: Principles and Adaptations* (1968), presents the mechanisms by which various groups of vertebrates deal with gas acquisition and transport. GEORGE M. HUGHES, *Comparative Physiology of Vertebrate Respiration* (1963; reprinted with corrections, 1965), is a brief introductory text. L. IRVING, "Comparative Anatomy and Physiology of Gas Transport Mechanisms," in W.O. FENN and H. RAHN (eds.), *Respiration*, sect. 3, vol. 1 of the *Handbook of Physiology* (1964), reviews invertebrate and vertebrate respiratory mechanisms including those of diving vertebrates. AUGUST KROGH, *The Comparative Physiology of Respiratory Mechanisms*, new ed. (1959), is a classic for those interested in comparative aspects of respiration. F.H. MCCUTCHEON, "Organ Systems in Adaptation: The Respiratory System," in D.B. DILL, E.F. ADOLPH, and C.G. WILBER (eds.), *Adaptation to the Eitvironment*, sect. 4 of the *Handbook of Physiology* (1964), discusses respiratory mechanisms in relation to the environment, with sections on chemical regulation, gas transport, and evolutionary patterns of respiratory adaptations. J.F. PERKINS, "Historical Development of Respiratory Physiology," in *Respiration (op. cit.)*, traces the chronology of the discoveries leading to modern respiratory physiology. C. LADD PROSSER and FRANK A. BROWN, JR., *Comparative Animal Physiology*, 2nd ed. (1961), has numerous examples of blood pigment behaviour as related to external and internal environments. H. RAHN, "Aquatic Gas Exchange Theory," in N. BALFOUR SLONIM and J.L. CHAPIN (eds.), *Respiratory Physiology* (1967), is a technical discussion of the application of gas laws to problems of gas transport in aqueous media. F.J.W. ROUGHTON, "Transport of Oxygen and Carbon Dioxide," in *Respiration (op. cit.)*, treats gas transport in the body fluids from a physicochemical point of view. P.F. SCHOLANDER, "Animals in Aquatic Environments: Diving Mammals and Birds," in *Adaptation fothe Environment (op. cit.)*, is a fundamental article for those interested in diving physiology.

(F.N.W.)

Respiratory System, Human

The respiratory system in man consists of the nasal cavity, the throat (or pharynx), the voice box (or larynx), the windpipe (or trachea), the bronchi, and the lungs, all of which are involved in the act of breathing.

Other bodily systems involved in breathing

Other parts of the body sharing in the process of moving air into, and out of, the lungs are ordinarily given membership in other body systems. The diaphragm—the muscular structure that separates the chest from the abdominal cavity—and the ribs and muscles of the chest walls play a necessary role in expanding and contracting the lungs so as to bring air in and push it out, but the ribs are viewed as part of the skeleton rather than the respiratory system and the muscles as part of the body's system of voluntary muscles.

If breathing is thought of as the process of instilling oxygen into the blood and bringing it to the tissues and of carrying carbon dioxide from the tissues out of the body, the elements of the respiratory system listed in the first paragraph above consist essentially of the organs in which the exchange of oxygen for carbon dioxide takes place—the lungs, and more particularly the pulmonary alveoli, or tiny air sacs, and the minute blood vessels around them—and the elaborate system of passageways through which air travels on its way to the alveoli and from them back into the external world. If breathing is the exchange of oxygen for carbon dioxide, it is clear that other parts of the body are involved: the blood itself, which transports the gases between the tissues and the lungs, and the cardiovascular system, consisting of the vessels through which the blood travels, and the heart, which provides the power by which the blood is driven.

The present article is focussed on the structure of the elements listed in the first paragraph: the nasal cavity, the pharynx, the larynx, the trachea, the bronchi, and the lungs (Figure 1). The functioning of the respiratory system is dealt with in RESPIRATION, HUMAN; its disorders are dealt with in RESPIRATORY SYSTEM DISEASES.

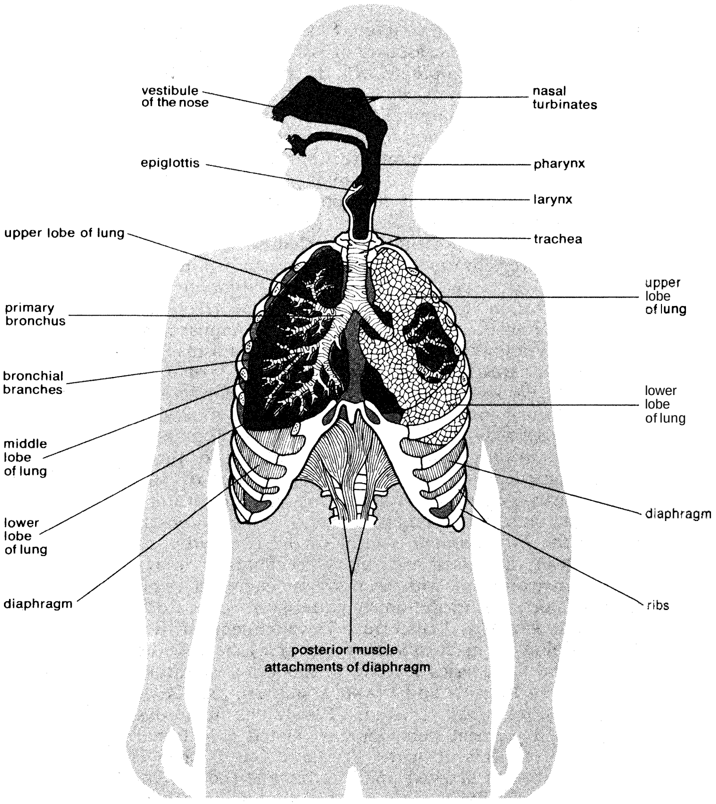


Figure 1: Respiratory system and its location.

THE UPPER PORTION OF THE RESPIRATORY TRACT

The nasal cavity. The nasal cavity is a space of complex shape, lined with mucous membrane and bounded below by the mouth and above by two air-filled cavities, the frontal and sphenoidal sinuses, and by a third cavity, the cranial cavity, which encloses the brain. **A** bony and cartilaginous partition, the nasal septum, divides the nasal cavity into right and left sides. Into the cavity's outer walls three ridges project, formed by the lower, middle, and superior turbinate bones. The air passageways with open inner side thus formed below each ridge are called the inferior, middle, and superior meatuses. (The configuration of the nasal cavity is described in SPEECH, PHYSIOLOGY OF.)

Divisions and passageways

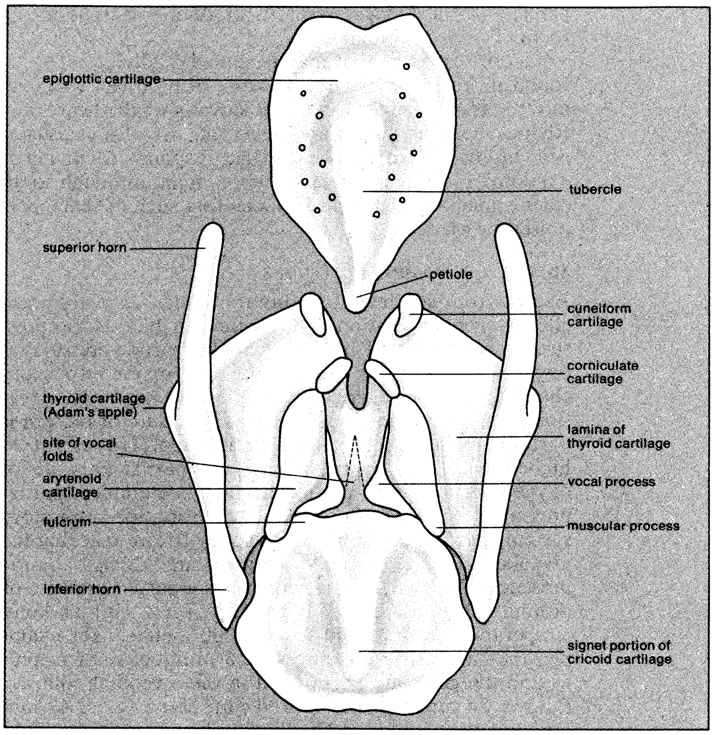


Figure 2: Cartilages of larynx seen from behind.
From H. Morris, Human Anatomy (1953); reproduced by permission of Blakiston Division, McGraw-Hill Book Company, Inc.

The uppermost portion of each nasal cavity contains an area highly specialized for olfactory sensation. Here are located the sensory cells that are the receptors for the sense of smell (see SENSORY RECEPTION, HUMAN).

The pharynx. The upper portion of the pharynx, the nasopharynx, is separated from the rest of the pharynx by the soft palate, which roofs the back of the mouth and terminates in the uvula. The nasopharynx is traversed exclusively by air and by nasal secretions. The nasopharynx and the mouth open into the oral pharynx, a common passageway, like the mouth, for air and food. By means of the epiglottis, which serves as a cap for the voice box, the latter is ordinarily closed during the process of swallowing, so that food is routed from the oral pharynx into the laryngeal pharynx, the portion of the pharynx behind the epiglottis and the voice box, and into the esophagus, the passage to the stomach.

The larynx. The larynx, a structure at the upper end of the trachea, or windpipe, serves a dual purpose, as one section of the airway to the lungs and as an important part of the apparatus by which voiced sounds are made. The outer framework of the larynx is made up of cartilage joined together by ligaments and membranes.

The thyroid cartilage, the largest of the cartilages in the larynx, consists of two plates joined at the front midline. Toward the top of this juncture is a notch, the thyroid notch, and just below the notch is the forward projection of the larynx, known as the Adam's apple. To the notch, on its inner side, is attached the epiglottis, the leaf-shaped plate of cartilage that serves as the lid for the larynx.

At the bottom of the larynx and forming a joint with the thyroid cartilage at one point on either side is a ringlike cartilage, the cricoid, which has a wider section, like the signet of a signet ring, in back (Figure 2).

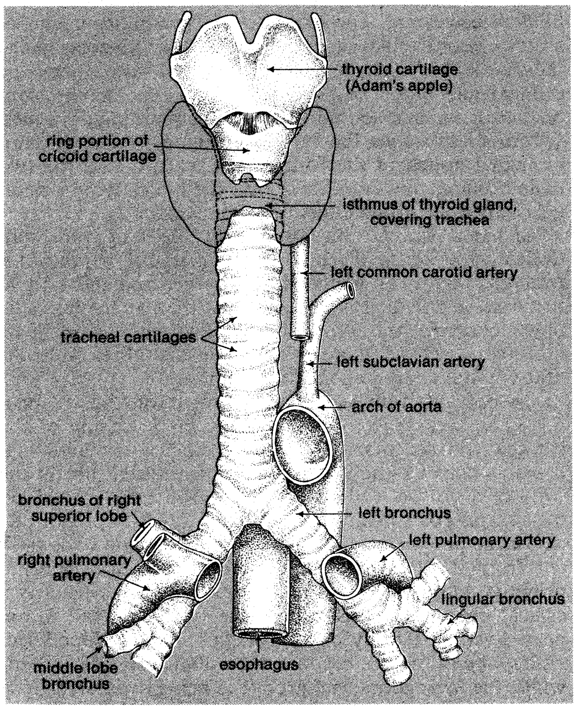

thyroid cartilage (Adam's apple)
ring portion of cricoid cartilage
isthmus of thyroid gland, covering trachea
left common carotid artery
tracheal cartilages
left subclavian artery
arch of aorta
bronchus of right superior lobe
left bronchus
left pulmonary artery
right pulmonary artery
lingular bronchus
middle lobe bronchus
esophagus

Figure 3: Larynx, trachea, and primary bronchi.

Adapted from Cunningham's Textbook of Anatomy edited by G.J. Romanes and Published bv Oxford Universitv Press as an Oxford Medical Publication

Dividing the larynx into an upper and a lower compartment are the vocal folds, white folds of mucous membrane that, at their inner edges, enclose the vocal ligaments, or vocal cords. A slit between the vocal cords, called the rima glottidis, is closed or opened wider by the swinging inward or outward of the arytenoid cartilages. These cartilages, attached to the vocal cords near the rear end of the rima glottidis, are pyramid-like in form and rest upright on the cricoid cartilage.

Vocal cords

Above the vocal folds are the vestibular folds, or false vocal cords, formed of mucous membrane that is a continuation of the membrane lining the larynx. The gap between the false vocal cords is wider than that between the true ones; thus, the true cords may be seen if the inside of the larynx is inspected with an instrument called a laryngoscope. The recess between the true vocal cords and the false ones is called the ventricle of the larynx.

The muscles of the larynx are classified as either extrinsic or intrinsic. The extrinsic muscles are those that extend between the larynx and adjacent structures, such as the hyoid bone and the breastbone, and move the larynx upward, as in swallowing, and downward. The intrinsic muscles have two functions, to bring the vocal cords together and to tense them.

The trachea. The trachea, or windpipe, is a tube consisting of a number of cartilages shaped like horseshoes or like **Y's** that are bent in a curve and lie on one side; these cartilages are enclosed in an inner and an outer layer of elastic fibrous membrane. In back, where the trachea lies next to the esophagus, there is a layer of smooth muscle that can draw the open ends of the cartilages together and so diminish the calibre of the tube. The trachea, like the larynx, is lined with mucous membrane. Its length, from the cricoid cartilage of the larynx to the fork at the juncture with the rieht and left main bronchi, the air passages directly into the lungs, is about four to 4.5 inches (ten to 11 centimetres) in the adult, and its diameter is about one inch (about 2.5 centimetres). The fork at the bottom of the trachea, which has a ridge, or keel, on its inner side, has developed into a vocal organ, the syrinx, in birds.

THE LUNGS AND BRONCHI

Structure of the bronchi. The right and left main bronchi, the stretches of bronchus between the lower end of the trachea and the lungs, have the same construction as the trachea, with cartilages shaped like broken rings or bent Y's and with tough covering membranes. The larger bronchi within the lungs have, instead of broken rings, irregularly shaped plates of cartilage completely around the wall, so that the intrapulmonary bronchi are cylindrical, rather than D-shaped like the trachea and the extrapulmonary bronchi.

Cartilages in bronchi

The layer of muscle, instead of being concentrated in the rear wall, as in the trachea and the extrapulmonary bronchi, completely encircles the intrapulmonary bronchi. There are blood vessels outside this smooth-muscle tissue and penetrating into it. On the outside of the bronchi and penetrating into the surrounding lung tissue is a tough, dense layer of connective tissue.

As the bronchi divide and subdivide and become increasingly smaller, the walls become thinner. Bronchioles (small bronchi), of a diameter of only a few millimetres, no longer have plates of cartilage in their walls. The smooth muscle is present, however, through the whole range of bronchial sizes, even down to the extremely fine alveolar ducts (see below Respiratory substance).

Shape, size, and location of the lungs. The two lungs rest against the outer walls of the chest cavity. Between them is the mediastinum, a thick group of structures that includes the heart, the major blood vessels, the trachea, and the esophagus. The lungs are roughly conical in shape, with pointed apexes jutting slightly above the collarbone, outer sides convex because of the rib cage, against which they rest, obliquely concave below because of the arching diaphragm beneath them, and deeply indented on their medial surface—the surface toward the centre of the body—by the heart and the other structures between them. Within the indentation in the medial surface of each lung is the hilus, the entrance point for the bronchi, blood vessels, and nerves.

The right lung is ordinarily somewhat larger than the left; together their average weight is about 600 grams (about 1.3 pounds). Each lung is enveloped in pulmonary (visceral) pleura, a tough membrane that exudes a small amount of lubricating fluid, as does the parietal pleura—the membrane lining the chest cavity—of which it is a continuation. Deep fissures in the lungs mark the outlines of their main divisions, the lobes. The right lung is divided by a horizontal fissure and an oblique one into three lobes; the left lung is divided by an oblique fissure into two.

The lobes

Bronchopulmonary segments, or lobules. Each lobe of the lungs is served by a lobar bronchus, one of the first branchings of the main bronchi. The three lobar bronchi in the right lung and the two in the left divide into segmental bronchi, ten in the right lung and nine in the left (Figure 4). The section of lung tissue served by a segmental bronchus and its many subdivisions is known as a bronchopulmonary segment, or lobule. The segments are separated by veins and thin membranes of connective tissue.

From E. Boyden in J. Myers, Diseases of the Chest and Heart. Courtesy of Charles C. Thomas, Publsher. Springfield, Illinois

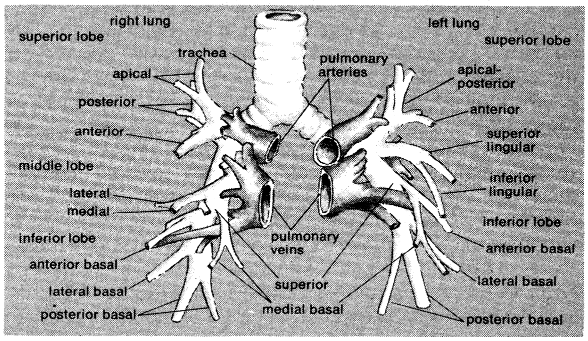


Figure 4: Segmental bronchi and pulmonary vessels (see text).

The bronchopulmonary segments are significant divisions of the lungs because such a segment or a group of such segments may represent the limit of involvement by a disease process, such as bronchiectasis. Thus, when surgical removal of diseased lung tissue proves necessary, it

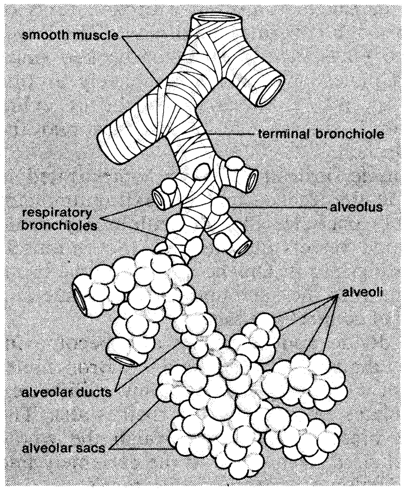


Figure 5: Terminal respiratory units of bronchial tree.

may be possible to confine the surgery to one or more segments and save most of the lobe.

Subdivisions of segmental bronchi. Each of the segmental bronchi divides and subdivides until there are 20 or more generations of branchings before the tiny bronchi called terminal bronchioles are reached. The first ten generations of bronchi are widely spaced divisions and occupy two-thirds of the distance from the hilus to the periphery of the bronchial apparatus. The remaining ten to 15 generations of branches are bronchioles, distinguished from the larger bronchi by not having cartilages. These bronchioles are more closely spaced and have a fairly uniform diameter, from about three millimetres down to about one (from about 0.12 to 0.04 inch).

Alveoli and alveolar sacs

Respiratory substance. The last bronchioles to be lined by cuboidal epithelium are called terminal bronchioles. There are an estimated 20,000–80,000 of these. Each gives rise to an acinus, a "berry"-shaped terminal group of air-filled respiratory bronchioles, alveolar ducts, and terminal sacs (Figure 5). It has been estimated that there are 300,000,000–400,000,000 alveoli in the lung. They are the functional units of respiration. In the walls of an alveolus is a dense network of capillaries, the smallest of the blood vessels. A network of fibres forms a scaffold on which the capillaries rest. Through the capillary walls oxygen from the air in the alveolus is exchanged for carbon dioxide in the blood. The respiratory substance of the lungs consists of the respiratory bronchioles, the alveolar ducts, alveolar sacs, and alveoli, together with the accompanying blood vessels, lymph channels, nerves, and connective tissue.

Blood vessels, lymph channels, and nerves. Blood laden with carbon dioxide and low in oxygen is carried from the right side of the heart through the pulmonary arteries to the lungs. Shortly after entering the lung at the hilus, the pulmonary artery divides into central arteries for the individual bronchopulmonary segments. These branches further divide and subdivide; a vessel to each alveolar duct breaks up into a capillary network for all the alveoli that open directly or indirectly into the duct. The bronchopulmonary segments are compartmentalized with respect to arteries but not veins; an intersegmental vein, lying in the membrane that separates adjacent segments, drains both. The oxygenated blood from the capillary networks is gathered into veins running through the bronchopulmonary segments and in the connective tissue between them. Near the hilus, the veins merge into about ten large veins that follow the course of the bronchi. The blood leaves the lungs in the pulmonary veins and is delivered to the atrium of the left side of the heart.

There are two sets of lymphatic vessels in the lungs. The lymphatic capillaries of one set lie under the pulmonary, or visceral, pleura, the membrane enveloping the lung, which they drain. These vessels converge at the borders of the lungs and at the edges of the fissures between lobes and end in the bronchopulmonary lymph nodes, at the hilus of each lung. The other set of lymphatics begins as a network of capillaries in the walls of the alveolar ducts; the capillaries converge into vessels that accompany the bronchi and blood vessels. They drain, as do the lymphatics of the first set, into the bronchopulmonary nodes. From these nodes the lymph drains into the tracheobronchial nodes, at the fork of the trachea and the two main bronchi, and from there into the major lymph channels, called the bronchomediastinal trunks.

Sympathetic and parasympathetic nerve fibres

The blood vessels and bronchi of the lungs are controlled by sympathetic and parasympathetic nerve fibres from the vagus, or tenth cranial nerve, and the sympathetic nerve trunk. It is believed that the parasympathetic fibres constrict the bronchi and that the sympathetic fibres constfict the blood vessels. Sensory fibres from the lung are thought to join the vagus nerve.

Development of lungs. The rudiment of the tracheobronchial tree—the treelike formation consisting of the trachea and the many branchings of the bronchi—appears as early as the 25th day of intrauterine life in the form of a groove, called the laryngeotracheal groove, on the ventral (or forward) side of the tubular foregut just below the pharynx. A day or so later, two lung buds push out from the lower end of this groove. These become primitive lung sacs, or primary bronchi. This period is a critical one, and, occasionally, one lung may fail to appear. By the 31st day each bronchus has developed secondary buds, which become the lobar bronchi. Meanwhile, the laryngotracheal groove is being separated from the esophageal portion of the gut by a wall that moves progressively upward from the level of the bronchial buds to the larynx. As each lobar bronchus continues to grow, it divides. By the 36th day all ten segmental buds have appeared, giving the exterior of the lung a scalloped appearance. By the 40th day several additional generations of branchings have appeared, so that a model of the bronchial tree resembles great clusters of grapes.

This period of branching (pseudoglandular period) continues through the fourth month. About 16–26 generations of branching are then established. Between the fourth and sixth fetal months (canalicular period) the surface epithelium is flattened out from below by invading capillaries. Two adherent cellular membranes for exchange of gases thus form. Between the sixth month and birth (terminal sac period) this double membrane invades the surrounding connective tissue, forming clusters of sacs in the walls of which primitive alveoli develop just before birth. Cells secrete a lipoprotein substance (surfactant) that facilitates expansion of the air–fluid interface at birth. From this time on the premature infant is viable.

The first respirations after birth expand only a limited area. Full expansion may require weeks. Portions of the lung may never dilate (a condition called congenital atelectasis). In premature infants and those delivered by cesarean section, a proteinaceous membrane, probably derived from leaking adjacent capillaries, sometimes lines the alveoli, preventing respiration.

It has been estimated that from birth to maturity the lung volume increases from 0.2 to 5.5 litres, that the number of alveoli increase from 24,000,000 to 296,000,000, and that the air-tissue interface (for exchange of gases) increases from 2.8 to 75 square metres (30 to 810 square feet). At birth the only alveoli present are located in the very end branches of the tree, in tiny clusters of alveolar sacs one millimetre in diameter. Within two months, new alveoli appear along the bronchioles in a centripetal direction so that what were clusters of saccules become alveolar ducts, and what were smooth respiratory bronchioles bulge with new alveoli.

BIBLIOGRAPHY. Further information may be found in E.A. BOYDEN, *Segmental Anatomy of the Lungs* (1955) and "The Structure of the Human Pulmonary Acinus," *Am. J. Anat.*, 132:275–300 (1971); R.C. BROCK, *The Anatomy of the Bronchial Tree*, 2nd ed. (1954); W.S. MILLER, *The Lung*, 2nd ed. (1947); V.E. NEGUS, *The Comparative Anatomy and Physiology of the Larynx* (1949, reprinted 1962), and *Physiology of the Nose and Paranasal Sinuses* (1958).

(E.A.Bo./Ed.)

Respiratory System Diseases

The respiratory system includes the structures and organs that have to do with breathing—the nasal cavities, the throat or pharynx, the larynx, the trachea, the bronchi, and the lungs. The lungs are the site of an exceptionally large range of disorders. The main reasons for this are three: first, they are exposed to the environment in which man lives and are therefore affected by dust or other substances in the air that he breathes; second, they possess a large capillary network through which the entire output of the heart has to flow every minute, and this means that diseases with a general effect on small blood vessels are likely to affect the lung; and third, they may be the site of "sensitivity," or allergic phenomena that may profoundly affect their function.

General considerations

SIGNS AND SYMPTOMS OF RESPIRATORY DISEASES

By contrast, the symptoms of lung disease are relatively few. Diseases of the major airways, of which chronic bronchitis is an example, cause a cough productive of sputum as their most important manifestation. In more severe degrees of involvement of the bronchial tree, in which there may be considerable enlargement of the mucous glands and excessive mucus production, as much as half a pint of sputum may be produced in a day. Much more commonly, however, an ounce or two of sputum is produced in a 24-hour period, and particularly during the first two hours of the morning. The presence of blood in the sputum (hemoptysis) is an important sign and one that can never be disregarded. Although it may have been caused only by an exacerbation of an existing chronic infection, it may indicate the presence of a serious pathological process such as malignant disease. Bloodstaining of the sputum is a classical symptom of pulmonary tuberculosis but may also occur when there is congestion of the lungs as a result of heart disease. Cough is a particularly important sign of all diseases that affect any part of the bronchial tree, and an irritative cough may be caused by involvement of the bronchial tree by extension of malignant disease or cancer from nearby organs.

The second most important indication of lung disease is shortness of breath. This sensation, known as dyspnea, may be insidious in onset and slowly progressive in some lung diseases, or it may occur acutely, as in the case of spasmodic asthma, in which the airway obstruction may be sudden in onset and equally swift in remission. Diseases such as pulmonary emphysema, in which there may be extensive lung destruction, are associated with unremitting shortness of breath, which may become so severe as more or less to immobilize the victim. Such dyspnea leads to a life of invalidism and may be one of the most distressing features of chronic lung disease. The slow development of diffuse interstitial fibrosis of the lung—widespread growth of fibrous tissue in the walls of the air sacs or alveoli—leads to slowly worsening dyspnea, and it is commonly this symptom which first causes the patient to seek advice from a physician. The absence of shortness of breath, however, may mislead someone into thinking that a given lung condition is necesssarily not a serious one. Actually, a small lung cancer, if it is not obstructing a major airway, will not produce shortness of breath. Though chest pain may be an early symptom of chest disease, it is most often associated with an attack of acute pneumonia or an embolus to the lung and in both of these situations is part of the symptom complex produced by these conditions. When an acute inflammatory process involves the external covering of the lung (the pleura), the pain is characteristically felt when a deep breath is taken, and it is often severe enough to limit the deep breathing of the patient. In some episodes of pneumonia, the chest pain may be very severe and disturb the afflicted by preventing sleep. Usually fluid soon collects in the intrapleural space and the pain of the pleurisy disappears. Severe chest pain may be occasioned by the spread of malignant disease outward from the lung to involve the pleura and even to extend to the chest cage. The intractable pain caused by this condition may require surgery to cut the nerves supplying the affected segment and give permanent relief from what is otherwise a severe and incapacitating symptom. Fortunately, however, this occurrence is the rare exception rather than the general rule.

Lesser signs and symptoms

To these major symptoms may be added a few more that are of less importance. The ailing person may observe a wheeziness in the chest, or this may be noticed by others. The cause of the wheeziness is airway obstruction. Some diseases of the lung are associated with the development of swelling of the tips of the fingers (clubbing of the fingers), and this may be a feature of bronchiectasis (chronic dilation of the bronchial tubes), diffuse interstitial fibrosis, and lung cancer. In the latter case, this unusual sign has been observed to disappear after removal of the tumour. In some diseases of the lung, the first symptom may be swelling of the lymph nodes that drain the affected area. Most particularly these may be small nodes situated above the collarbone in the neck, and enlargement of lymph nodes in these regions should always lead to a suspicion that the primary disease may be in the lungs. The biopsy of nodes in this region constitutes an important diagnostic technique in some types of lung disease, particularly sarcoidosis (a chronic disease characterized by the formation of many nodules or small lumps) and lung cancer. As noted in a later section, lung cancer may produce a wide variety of symptoms that may be mistaken for primary disease in other organs.

Debilitation from chronic lung disease

The generally debilitating effect of chronic lung disease is well recognized. The person with primary lung tuberculosis or with lung cancer, for example, may be conscious only of a general feeling of malaise, of unusual fatigue, and of apparently rather minor symptoms as the major indicators of his disease. Loss of appetite and loss of weight, a disinclination for physical activity, some general psychological depression, and some symptoms apparently unrelated to the lung such as mild indigestion or occasional headaches may be diverse indicators of these diseases. Not infrequently, he may feel as if he were still convalescent from an attack of influenza, and he may not realize that the primary cause of his malaise is the development of a serious lung condition. It is for this reason that the examination of the chest and the taking of a chest X-ray are regarded as basic to any general physical examination of a patient.

The development of X-rays and pulmonary function tests has diminished the importance of the stethoscope in making a definitive diagnosis, but in spite of this its convenience and usefulness in diagnosing airway obstruction and pulmonary congestion ensure that it will continue to be used. Skilled clinical diagnosis of chest disease depends upon the proper deployment of many methods of investigation, and the skilled integration of the resulting information. Great advances have been made in the treatment of lung disease. The one condition that was synonymous with lung disease for many years, pulmonary tuberculosis, is much less of a problem in civilized countries than it used to be. However, the alarming increase in the incidence of cancer, chronic bronchitis, and emphysema, and the increasing recognition of industrial lung disease, have been prominent features of the last 20 years.

Lung disease may be caused by physical factors such as dust, by bacteria and viruses, by sensitivity phenomena such as occur in asthma, or by disorders in the circulating blood, and the lungs may be involved in primary malignant disease and are commonly the site of secondary deposits from malignant disease elsewhere.

In the article, a general description of most of the important diseases affecting the respiratory system will be given. Details of medical treatment and specific drug therapy have in general not been included, but particular attention has been directed to the wider aspects of lung disease as these affect the community. The bibliography has been carefully selected to provide the reader with the answers to particular technical questions, and in them will be found elaboration of most aspects of lung disease.

MORPHOLOGICAL CLASSIFICATION OF RESPIRATORY DISEASE

Structure of lungs

It is helpful to recall the main divisions of the lung as a basis for the morphological distribution of respiratory system disease. Broadly speaking, the lungs may be visualized as consisting of the upper airway, the trachea or windpipe, and of progressive divisions of that airway system into progressively smaller tubes, until finally the terminal bronchioles, which are about one millimetre (mm) in diameter, are reached. Approximately 16 generations of division occur between these bronchioles and the trachea. Although there is one airway at the beginning, namely the trachea, there are approximately 650,000 terminal bronchioles. The cross-sectional area of the bronchial tree in fact increases with increasing subdivision, although the diameter of each individual bronchus is decreasing. The end of the terminal bronchioles opens into the "acinus," so called because the structure resembles a cluster of grapes, and from this point onward the gas-exchange portion of the lung is reached. The alveoli or air sacs, which are divided into groups or "lobules" by fibrous partitions or septae, are small hexagonal structures forming a blind end to the acinus. The wall of the alveolus consists almost entirely of blood capillaries, and the remaining structures are extremely thin, only providing supporting tissue for the rich blood capillary bed that constitutes the parenchyma, or the essential tissue of the lung itself. The blood is distributed to the lung through the pulmonary artery, which subdivides with the bronchial tree and which accompanies the smaller bronchioles into the region of the acinus to supply the capillaries of the alveolar wall. The pulmonary veins drain the oxygenated blood from the alveoli and run at some little distance from the bronchioles. An interstitial space exists around the alveoli and around the bronchioles and blood vessels, and this connects with lymph nodes (the small masses of lymphatic tissue that occur along the course of the lymph vessels) situated in the midline of the lung and extending up in a chain into the neck and down into the abdomen. The lung is covered by a protective membrane, the pleura, and the inner lining of the chest wall is comprised of a similar membrane. The "intrapleural space," which exists between these two fibrous coverings, normally contains no air, and only a few millilitres of liquid for lubrication purposes. The pleurae may, however, become involved with inflammatory disease, in which case an effusion of fluid may occur between the two layers of the pleura.

Morphologic categories of respiratory tract disease

From this general description, the diseases of the respiratory tract may be broadly grouped into the following categories.

Diseases of major bronchi. These become the seat usually of chronic inflammation as in chronic bronchitis or bronchiectasis, or they may be the common site of primary malignant disease or lung cancer.

Diseases of smaller bronchi and bronchioles. It is in the smaller bronchi that major obstruction commonly occurs in asthma, since these bronchioles are supplied with smooth muscle and this may contract to obstruct the airway in this disease. The small radicles of the bronchial tree, the bronchioles, are commonly involved in infective processes and are the primary site of dust deposition in particular occupations.

The alveolar ducts and alveoli. These are the site of primary involvement in many infections, including pneumonia, and it is on the parenchyma of the lung that the main effects occur when the primary artery is obstructed by a blood clot. The capillary bed surrounding the alveoli is subject to damage, and it is through the alveolar capillaries that leakage of fluid leading to its accumulation in the lungs (pulmonary edema) may occur. The alveolar walls themselves may be involved with diffuse interstitial thickening, and this is a prominent part of the diseases grouped under the general heading of "diffuse interstitial fibrosis."

The lymphatic system. The system of channels draining the lung may be involved in primary disease of the lungs, and some diseases may first show themselves by enlargement of the lymph nodes in this area at a time when the essential tissue of the lung is still minimally or inconspicuously involved.

The pleura. The pleura is commonly involved in acute inflammatory processes; this gives rise to the symptom of pleurisy. The intrapleural space may also be the site of the collection of fluid forming a "pleural effusion." This most commonly occurs either as a result of acute or chronic lung infection, or as a result of lung cancer.

Although these divisions give a general outline of the ways in which diseases may affect the lung, they are by no means rigid. It is common for more than one part of the system to be involved in any particular disease process, and disease in one area not infrequently leads to involvement of other parts of the system. Thus, obstruction or blockage of a branch of the pulmonary artery leads to extensive congestion, hemorrhage, and pathological change in the capillaries and alveoli beyond the obstruction. The process of pneumonia may simultaneously involve the pleura and the pulmonary parenchyma. And the disease complex of chronic bronchitis and emphysema most often simultaneously involves chronic inflammatory change in the major bronchi, obliterative chronic inflammatory change in bronchioles, and, finally, destruction of pulmonary alveoli, commonly the alveoli situated in the middle of the lobule. Involvement therefore of different parts of the lung may be synchronous, and a rigid classification is consequently not possible.

THE EFFECT OF RESPIRATORY DISEASE ON PULMONARY FUNCTION

Normal function and symptoms of abnormality. The primary purpose and function of the lung is to fully oxygenate the blood and enable carbon dioxide (CO_2) to be eliminated. In many conditions, the lungs are unable to accomplish these tasks and there is therefore a fall in the oxygen content and oxygen pressure of arterial blood, and there may be an increase in carbon dioxide as a manifestation of lung disease. These disorders most commonly occur when there is inefficient ventilation and perfusion within the lung, or when there is a gross decrease in effective ventilation. There may, however, be severe physiological abnormalities detectable long before the gases dissolved in the blood have changed. In some diseases, the primary defect is in an increase in the resistance to airflow through the bronchial tree. This is the primary manifestation of the airway narrowing that occurs in bronchial asthma, and it is also a prominent feature of chronic bronchitis. The abnormality may be detected by the retardation in airflow that occurs when a subject makes a maximal expiration. Extensive and diffuse involvement of the alveolar tissue of the lung may lead to a severe increase in the "stiffness" of the lung tissue, which may be measured as a reduction in pulmonary compliance. In very serious examples of this disorder there may be a severe defect of gas transfer across the alveolar wall of the lung. In many respiratory diseases, the lesions are not uniform and, in many situations, complex patterns of pulmonary dysfunction may be observed.

Causes of shortness of breath

The sensation of shortness of breath is one of the common symptoms of pulmonary abnormality. During normal exercise, there is a considerable increase in pulmonary ventilation, and the lungs are so constructed that this can normally take place without any consciousness of major effort. In moderately severe exercise, the lung ventilation may increase from a resting value of about 10 litres (1) a minute to as much as 50 or 60 litres a minute. Any obstruction to airflow, however, will lead immediately to a consciousness that more effort is needed to shift these volumes of gas in and out of the lungs. Thus, although a person may be comfortable sitting at rest, as soon as his ventilation demand increases on exercise he may become aware of difficulty in breathing. In some chronic conditions such as chronic bronchitis and pulmonary emphysema, there may be severe airway obstruction, and this is the prime cause of the shortness of breath that the ill person notices and the main reason why his exercise may become severely limited. Many factors contribute to this airway obstruction, including changes in the major bronchi themselves, inflammatory changes in small airways, and the development of pulmonary emphysema, which involves destruction of pulmonary alveoli with a consequent loss of the normal elastic recoil of the lung. The loss of this recoil means that the airways are more likely to collapse and obstruct airflow when expiration is occurring.

The normal lung is capable of a more or less normal distribution of ventilation within it, which is to say that the gas being inspired is distributed more or less uniformly to all parts of the lung. The presence of lung disease, or the presence of obstruction in one airway, means that this will no longer occur and the distribution of ventilation may become nonuniform. Gross disturbance of the lung architecture will also cause abnormalities of ventilation distribution and may contribute to the dysfunction of the lung.

The transfer of oxygen across the lung is dependent on the maintenance of good ventilation of the alveoli and hence a high pressure of oxygen in the alveoli at all times. Transfer of oxygen from gas to blood occurs by diffusion from alveolar gas into the blood. Oxygen, being a relatively insoluble gas, does not move quickly through biological membranes and thus takes a finite time to saturate the blood in the alveolus. Normally, the red cells of the blood are only in contact with gas for about 0.7 second in a normal subject at rest, and on moderately severe exercise this time may be reduced to 0.3 second. In normal people at sea level, full saturation of the blood occurs in all circumstances within these time limits. Ascent to altitude, where the alveolar tension is reduced, may limit the attainment of full equilibrium, and the same thing may occasionally occur if the alveolar wall is extensively involved by disease even when the subject is at sea level.

An important cause of failure of lung function is diminution in ventilation of the lung. Occasionally, this may be the result of some defect in the regulatory system and may be caused by lesions in the brain. More commonly, there is general underventilation of large parts of the lung as a result of some disease process, and this may cause failure of gas exchange. The normal maintenance of arterial oxygen tension and elimination of carbon dioxide is very much dependent on the uniform distribution of ventilation and perfusion within the lung, and if these are mismatched or if there is a decrease in effective ventilation, the gas pressures are changed, leading to deficient oxygenation of arterial blood and deficient elimination of carbon dioxide. It is important, however, to realize that there may be considerable airway obstruction with shortness of breath and troublesome symptoms of lung disease before any change may be noted in the arterial blood gases, and it is for this reason that simple measurement of the arterial blood is not the sole pulmonary function test of importance in the understanding of chest disease.

Tests useful in diagnosis of lung disease

The measurement of capability with respect to pulmonary ventilation, the precise ascertainment of the oxygen and carbon dioxide pressures of the arterial blood, measurements of the efficiency of distribution of ventilation and perfusion in the lung, measurements of gas transfer across the alveolar bed, and finally measurements of the lung stiffness and airway resistance by simultaneous recording of esophageal pressure and airflow have contributed immeasurably to the understanding of lung-disease processes. Many of these techniques are in routine use in hospitals, and some of them have proved important in survey work in relation to chronic respiratory disease within a given population. In particular, for the study of chronic bronchitis, which is one of the most important contemporary disease problems, particularly in a population with heavy cigarette consumption, the surveys making use of measurements of ventilatory capability have proved of exceptional value in indicating, in a quantitative manner, the ventilatory capability of the subjects. This may be very hard to gauge from clinical examination or questioning, and since the plain chest film in this condition is commonly normal, this kind of survey has become very dependent on field laboratory measurements.

Not uncommonly, it is necessary to be able to measure the function of the two lungs separately. This is usually done by passing into the trachea a rubber tube that divides the air coming from the two lungs. More recently, radioactive gases and particularly xenon-133 have proved invaluable in measuring relative efficiency, both in terms of ventilation and blood flow, of different parts of the lung.

The respiratory system presents a number of effective protective mechanisms. The nasal mucosa acts as an air-conditioner and traps large dust particles. The major bronchi are equipped with ciliated cells and lined by a layer of mucus to which particles and bacteria may adhere, to be moved continuously upward by the action of the cilia. The act of coughing is an important protective mechanism, and this powerful reflex is trigged by irritation in the major bronchial tree. The forcible airflow that results is extremely effective in moving large pieces of material upward. The lung is equipped with major healing capability particularly after infections, but the damage caused by some disorders such as emphysema seems to be generally beyond repair. At the alveolar level there are wandering macrophages, cells with the ability to ingest particles, and it is likely that the surface-active lining of the alveoli may have an important protective function. Particles that have been breathed in and that reach the alveolar wall may be removed to lymphatic channels and stored in lymph nodes to which these drain.

Characteristics and course of respiratory failure

Respiratory failure. In many of the conditions described below, but particularly in chronic conditions such as pulmonary emphysema and conditions involving the pulmonary musculature, respiratory failure may occur. In this condition there is a failure to oxygenate the blood normally and to eliminate sufficient carbon dioxide to keep the tension of this gas in the blood at its normal value of 40 millimetres of mercury (mm Hg). In some conditions, particularly acute asthma, there may be a dangerous drop in arterial oxygen tension before much retention of carbon dioxide has occurred, and this may also be seen as a result of injury to the chest or in certain postoperative situations. The commonest cause of this reduction in arterial oxygen tension is the continued perfusion of parts of the lung with blood, although the ventilation to these regions has been greatly reduced. Reduction of arterial oxygen tension may in turn produce a fall in systemic blood pressure, which aggravates the oxygen lack existing at the tissue level. Although easily reversed, this stiuation may be difficult to diagnose clinically in its early stages. Later, often when the patient is fatigued or exhausted, the arterial carbon dioxide tension begins to rise. At levels of 80 to 90 millimetres of mercury, carbon dioxide produces a narcotic effect; this state is called carbon dioxide narcosis. The rise in carbon dioxide is accompanied by increasing acidity of the blood. If untreated, these changes combined with hypoxemia lead to failure of the heart and circulation, and death. The introduction of modern methods of analyzing the arterial blood for its gas content and tension have revolutionized the management of this condition in hospitals. The provision of medical staff and nurses specially trained in the technique of suction of secretions from the bronchial tree, the administration of enriched percentages of oxygen in a controlled manner, and, if necessary, the protracted artificial ventilation of the ill person either through a tube passed through the nose into the trachea, or through an opening made in the trachea, have resulted in the successful treatment of many persons with acute respiratory failure who would previously have certainly died. The maintenance of respiration when muscle disease has led to respiratory weakness may be necessary for several weeks until the disease spontaneously improves. The person with some degree of chronic lung disease may be thrown into respiratory failure by an acute infection, and if he can be brought through this successfully, he may have useful activity for years afterward. Sufferers from spasmodic asthma may, and commonly do, recover completely normal pulmonary function although an acute episode of status asthmaticus may be life threatening (see above). The prompt and careful management of these patients is required for successful therapy.

RELATIONSHIP OF RESPIRATORY DISEASE TO AIR POLLUTION

Although the influence of cigarette smoking and of certain industrial occupations on the prevalence of different kinds of lung disease has been established, the general influence of air pollution is much more difficult to define. Air pollutants commonly measured consist of sulfur dioxide and general particulate pollution. It has been clearly shown that in most cities where the levels of these pollutants are high, the general incidence of chronic bronchitis and emphysema and acute respiratory infections bears some relationship to the level of the pollutants. It has also been demonstrated that the incidence of respiratory infections in children under the age of two years is related to the general level of air pollution as indicated by these two measurements. In large industrial areas, it has been shown that there is a general relationship between morbidity from respiratory disease in people who already have some respiratory condition and peaks of worsening air pollution as measured by smoke and sulfur dioxide concentration. However, it is far from proved that all these effects should be attributed directly to sulfur dioxide. It is not clear what other aspects of city living, such as overcrowding, exposure to other people with respiratory infections, or even potentiation of the growth of respiratory viruses in the atmosphere as a consequence of the presence of air pollutants, may play in producing these results. The influence of other pollutants such as oxides of nitrogen, which are produced by automobile exhausts, and ozone, which may be produced by the action of sunlight on automobile exhaust, is not well defined. One of the reasons for the difficulty in clarifying the influence of air pollution on respiratory disease has been the concomitant increase of cigarette smoking and of the diseases which have been shown to be closely related to this. Indeed, only if the noncigarette-smoking population is studied can reliable statistics be obtained. If different populations are to be compared, it is essential that the cigarette-smoking habits of each group be most carefully analyzed. It may be said, however, that the evidence linking air pollution with exacerbations of respiratory disease is solid enough to form a basis for legislative action to limit the deteriorating environment of most industrial regions.

Diseases of the respiratory system

DISEASES CAUSED BY VIRUSES AND SIMILAR ORGANISMS

A wide variety of viruses is responsible for respiratory disease. It is a common experience that the common cold may extend to involve the trachea and to cause laryngitis, and there is little doubt that these processes may extend to involve the lower bronchial tree. After such episodes the ciliary lining of the bronchial tree may be damaged, but the repair processes are usually rapid.

Infections with the rhinovirus and adenovirus groups of viruses are particularly important in children, in which they may cause an acute febrile illness sometimes with a severe degree of bronchial tree involvement. Recovery from all these processes is usually rapid. In exceptionally severe infections there may be involvement of the pulmonary parenchyma, or an effusion into the intrapleural space with pleurisy, and in rare instances concomitant involvement of the pericardium—the protective membrane enclosing the heart—and even of the heart muscle itself. Isolated severe examples of these diseases may occur during an epidemic.

By contrast, influenza and parainfluenza may commonly cause more severe respiratory illness. The influenza virus attacks many systems of the body simultaneously, but it appears likely that its primary site of reproduction is in the alveolar cells of the lung. Tremendous multiplication of the virus occurs in these cells within a 24-hour period, and the pulmonary involvement may begin at the parenchyma and cause considerable consolidation and inflammation of lung tissue. It is common for there to be severe bronchitis and tracheitis at the same time. A less severe example of the disorder is that described as viral pneumonia, in which a distinguishing feature is the presence of patchy areas of atelectasis, or partial collapse of lung tissue, without very widespread bronchial-tree involvement. These conditions may be particularly dangerous in elderly people and in small children, and the lung that has been the seat of a severe attack of influenza may quickly become secondarily infected with streptococci or staphylococci.

Influenza epidemic of 1918–19

The severe epidemic of influenza of 1918 and 1919 was one of the worst human catastrophes on record. It has been estimated that more than 20,000,000 people died around the world as a result of this epidemic, and of the 20,000,000 persons who suffered from the illness in the United States, approximately 850,000 died. It was a characteristic of this epidemic that young people were commonly severely affected, and in them the virus illness was complicated by addition of a bacterial pneumonia to the original virus infection. At that time there was no effective chemotherapy for the secondary bacterial invasion. There are three immunologically distinct types of influenza virus, designated A, B, and C, and there is also a group designated by the letter D, known as the parainfluenza viruses. However, within type A there are now known to be at least four strains. The "Asian" strain of type A was responsible for the 1957 epidemic. Epidemic influenza tends to go in two- or three-year cycles, and the careful study of epidemics has enabled some predictions to be made concerning their future occurrence. Although there is lasting immunity to a particular strain following an attack of influenza, the immunity is highly specific as to type and no protection is afforded even against closely related strains. Artificial immunization with high-potency vaccines is of value in protecting against previous strains, and the vaccines have been shown to ameliorate the infection in the general population. They are particularly indicated for use in elderly people or those with pre-existing lung disease.

Mycoplasma, the organism identified in 1944 as responsible for a group of pneumonias, is an organism having many forms and capable of passing through filters, and for this reason it was originally regarded erroneously as a virus. It is a member of a group of organisms known as the pleuropneumonia-like organisms (PPLO) and has also been called the "Eaton agent" after the scientist who first described it. It is commonly responsible for what used to be known as "viral pneumonia." It produces characteristic soft shadows on the X-ray and relatively few signs that are observed in clinical examination. Usually a nonproductive cough and a fever persist for a few days. The disease is common in children and particularly occurs in epidemic form in young people brought together as into an army recruiting centre. Normally, it is a self-limiting disease of short duration.

Psittacosis and ornithosis, primarily infections of birds and particularly common among parakeets and parrots, are transmitted to human beings by inhalation as dust from the droppings of infective birds. The onset of psittacosis may be quite severe with headache, insomnia, and even delirium. Gastrointestinal symptoms such as vomiting and abdominal pain are frequent, and a cough with production of clear sputum usually develops after a few days. The areas of pulmonary consolidation ("solidification" with white blood cells, which shows in X-rays) may change over the course of the illness, and mild attacks of the condition are probably often unrecognized or dismissed as influenza. Recovery is usually complete, but convalescence may be slow. The diagnosis may be made by concurrent X-ray evidence of consolidation of the lung, and identification of the micro-organism in the sputum or in the blood. The greatest epidemic of this disease occurred in 1929 and 1930, when a pandemic was caused by the shipment of 5,000 parrots into Argentina from Brazil for auction. Many of the birds died, and there was a considerable human mortality from this particular episode. This disease is to be distinguished from the sensitivity diseases that may result from contact with birds.

Q fever

Q fever is an infection with *Coxiella **burnetii.*** The disease gets its name from the fact that it was first described in Queensland, Australia, though it is now known to be widely prevalent in North America and Europe. It is probably a common disease of cattle, and transmission is believed to occur between mammals through ticks and mice. Slaughterhouse workers are particularly susceptible to the disease, but the method by which humans contract it is often not clear. It may be transmitted through milk, particularly if this has not been pasteurized, and in the past this has probably been an important method of transmission. Q fever is usually a self-limited disease and calls only for symptomatic treatment; the pulmonary aspect of the condition is usually confined to temporary inflammation and accumulation of white blood cells with no later complications.

The disease chicken pox (varicella), particularly when it occurs in adults, may affect the lung. Acute lesions may occur at the alveolar level, leading to a transient, but significant, fall in arterial oxygen tension and occasionally necessitating oxygen therapy. Recovery from this condition seems to be slow but complete, but radiological shadows may be left as a result of it.

It seems clear from recent information that the reparative processes occurring in the lung after viral attack may be quite slow. Apparent clinical recovery may occur relatively quickly and the X-ray may show no remaining shadows, yet at the level of the alveolar wall the process of repair and restitution may take several further weeks.

Whooping cough occurs in epidemic form, particularly among children, and is important not because of its immediate effects but because there is evidence linking it with the later development of the chronic infective process known as bronchiectasis.

BACTERIAL DISEASES

Bacterial pneumonia. Until the development of effective antibiotics, pneumonia was the respiratory disease responsible for the greatest mortality and was one of the most feared of all diseases. The commonest form of this disease was caused by the "pneurnococcus" (Diplococcus pneumoniae). This organism characteristically caused an acute illness of sudden onset with high fever, consolidation of a whole lobe or more than one lobe of the lung, followed either by such complications as a lung abscess or pleurisy or heart failure, but naturally terminating after about seven days in a high peak of fever and then a sudden crisis followed by a sharp fall in temperature and slow resolution. This dramatic illness, so often fatal in older people, came to be known as the "old man's friend." The classical form of the illness is now rarely seen by the medical student, since prompt antibiotic therapy controls the acute process within 24 hours. The streptococcus causes a diffuse type of bronchopneumonia and it is this condition that is most likely to occur as a consequence of forced recumbency in old people—sluggish venous circulation results in accumulation of fluid in the lungs—or as a complication of a severe attack of influenzal pneumonia. This disease has also been largely controlled by antibiotics. Staphylococcal pneumonia now occurs as an acute illness of small children and of adults whose resistance has been lowered. In small children it may lead to rapid destruction of lung tissue with abscess formation, but from which the lung fully recovers if the acute state is survived (as it usually is with modern therapy). The development of antibiotic-resistant staphylococci has meant that this form of pneumonia has become a problem in the hospital environment and may complicate other lung disease or lung damage and may particularly be a problem in the postoperative period in those who are debilitated and weak. Pneumonia due to infection with *Klebsiella* pneumoniae may be particularly difficult to treat and characteristically may occur as a repetitive series of episodes of pneumonia each running a rather long course with slow resolution.

Resistant staphylococci

In all these bacterial pneumonias the diagnosis may be made by characteristic X-ray patterns combined with the isolation of the bacteria primarily responsible from the sputum. The organism *Haemophilus influenzae* is commonly isolated from the sputum of patients with chronic bronchitis (see below) and may be associated with acute exacerbations of that condition, though its exact role as a cause of disease is not well understood.

Pulmonary tuberculosis. Of all the lung diseases caused by bacteria, historically, pulmonary tuberculosis is by far the most important. Particular features of this condition, to which many writers and novelists of the last hundred years have turned for dramatic material, include the severe general debilitation and weakness that it may cause; the insidious nature of the onset of its initial symptoms, which may not be primarily pulmonary in nature; the fact that certain families are particularly prone to it; the long, drawn-out nature of the course of the disease and the distressing nature of many of its manifestations—particularly, severe hemorrhage of the lung, involvement of other organs, especially the brain in fatal meningitis, and the hopelessness of all medical attempts to control it until the development of effective antibiotic treatment. These compounds have largely displaced traditional methods of treatment such as collapsing the lung by pneumothorax (the injection of gas into the chest cavity), and they have greatly reduced the mortality from this disease in all civilized countries. In countries in which medical programs have not yet been applied to any major extent to rural populations, and particularly tropical countries in which this disease may take a heavy toll, pulmonary tuberculosis remains a major public-health problem. Diagnosis within a population may be greatly assisted by a survey of the population by taking miniature X-rays of large numbers of people; this method has contributed much to the control of the disease. The much more effective treatment and control of this condition has had the beneficial effect of reducing people's fear of it and has also resulted in the necessary reorientation of large numbers of sanatoriums and pulmonary-tuberculosis hospitals that are no longer required for long-term care.

In its classic form, pulmonary tuberculosis occurs principally at the apexes (upper portions) of the lung and may slowly progress to form a chronic cavity within the lobe affected. The cavity develops a fibrous wall, and once the disease has reached this stage complete healing may be difficult even with antibiotic therapy. Secondary infection of the cavity may occur and be difficult to eradicate. Pulmonary tuberculosis, when still active, is a constant threat to the patient because blood-borne spread to many organs may occur at any time and the disease may become more active or extend to other parts of the lung. The diffuse spread of tuberculosis may occur as its first manifestation, and this condition, known as miliary tuberculosis, usually begins as a generalized illness of unknown cause. The diagnosis may be suspected when the chest X-ray reveals diffuse small shadows, and when other causes of fever have been excluded. The exact sequence of events that leads to this disseminated type of tuberculosis is not understood, but prompt treatment is essential if spread to the brain is to be avoided. Tuberculous meningitis, in which spread to the brain has occurred, is one of the most dreaded complications of pulmonary tuberculosis and, before antibiotic therapy, was almost uniformly fatal. Tuberculous involvement of the adrenal glands leads to the syndrome of adrenal insufficiency called Addison's disease, a condition that comvounds the general debilitating effects of the pulmonary disease alone.

Treatment of pulmonary tuberculosis

The treatment of pulmonary tuberculosis has been revolutionized by the introduction of effective chemotherapy. Streptomycin was the first clinically successful antituberculosis drug and has remained important. It is now used mostly in conjunction with other antituberculosis drugs, isonicotinic acid hydrazide (isoniazid) or para-aminosalicylic acid (PAS), and, although originally toxic effects of streptomycin were commonly seen as a consequence of heavy dosage, it is now used in a dose well below that causing toxic symptoms. The symptoms of toxicity are selective on the nervous system and produce unsteadiness and attacks of giddiness with abnormalities and deficiency in hearing. Isoniazid has considerable potency against tuberculosis but has the disadvantage that the tubercle bacilli may rapidly acquire resistance to the drug. This drug is consequently never given alone but in combination with PAS or streptomycin. The drug diffuses well into the cerebrospinal fluid, and it is particularly valuable in treating tuberculous meningitis or miliary tuberculosis. It has few serious toxic effects, but it has been suspected of occasionally causing rheumatic symptoms in patients under long-term treatment for tuberculosis. PAS is usually given in combination with isoniazid. Although it is valuable in the treatment of tuberculosis, it is a difficult drug to administer because it produces side reactions. These are mainly gastrointestinal irritation, with impairment of appetite and some dyspepsia and diarrhea. Although these are not indications of serious toxicity, the symptoms are troublesome and make patients reluctant to take the drug if they are being affected by them. Viomycin sulfate is also used in the treatment of tuberculosis; it is now most commonly prescribed in doses lower than those usually responsible for its neurotoxic side effects. Pyrazinamide is not used alone in the treatment of tuberculosis but may be valuable in conjunction with other drugs. It should not be used in patients with a history of gout because it has a secondary effect of causing elevation of the level of uric acid in the blood. Ethionamide and ethambutol hydrochloride are both occasionally used in the treatment of tuberculosis, particularly when resistant strains of the organism have developed.

It is important to stress that there is no one "best" regimen superior in all circumstances to every other, in the treatment of pulmonary tuberculosis. Tubercle bacilli are able to acquire resistance to most of the antituberculous drugs, and skillful treatment consists in combining antibiotics so that resistant bacilli are less likely to be produced. The resistance to antibiotics may be delayed by the concomitant use of two or more drugs, by continuous treatment without significant interruption until bacterial growth ceases, and, thirdly, by the use of bed rest or surgery in selected cases in addition to antibiotic therapy. Clinical improvement of most patients with symptom-producing pulmonary tuberculosis is rapid if the bacilli are sensitive to the specific drug given. The general symptoms of the illness diminish, and radiological regression begins within a few weeks and may be complete within a few months. New shadows almost never appear on X-rays during treatment with antibiotics. One of the problems of treating pulmonary tuberculosis is that of ensuring that therapy is maintained and that the victim is continuously supervised. For this reason the treatment of pulmonary tuberculosis in undeveloped parts of the world or in remote village communities poses severe problems for the public-health authorities. Effective programs of control in the village environment have been initiated in India, but it is recognized that major pools of infection may be hard to eradicate in some of these isolated communities without continuing medical care. Effective treatment of pulmonary tuberculosis in the environment of the Western city may be equally complicated by the difficulty of ensuring that elderly, unemployed, and possibly alcoholic patients are kept on a proper regimen of treatment. It may be difficult to ensure continued supervision and treatment of such patients particularly if they drift from one city to another. Although the death rate from respiratory tuberculosis has diminished greatly in western Europe and in the United States since 1950, it remains an extremely serious problem in some underprivileged communities, in various tropical countries, and in any circumstances in which the medical supervision of the population is difficult or inadequate.

ALLERGIC LUNG DISEASES

There are at least four reasons why the lungs are particularly liable to involvement in "allergic" responses to proteins. They are exposed to the outside environment and hence particles of foreign protein—*e.g.*, in pollen—may be deposited directly in the lung; the walls of the bronchial tree contain a considerable amount of smooth muscle that is very likely to be stimulated to contract if histamine is released by cells affected by the allergic reaction; the lung contains a large vascular bed and therefore, in general responses causing diffuse inflammation of blood vessels, it is likely to be involved; and, finally, the lung has biochemical functions (at present very poorly understood), which mean that it is a store of histamine, has the task of denaturing circulating substances such as serotonin and bradykinin, and may well have important biochemical functions relating to proteins and allergins. It is therefore not surprising to find that sensitivity phenomena are common and represent an important aspect of pulmonary disease as a whole. The commonest and most important of these is asthma. This word is occasionally misused to indicate all kinds of conditions in which there is airflow obstruction in the lungs, but it is better reserved for those conditions in which an allergic component of the bronchial obstruction is likely to be present.

Causes and course of asthma

Asthma. Spasmodic asthma is a disease characterized by contraction of the smooth muscle of the airway, and later by airway obstruction as a result of failure of mucus clearance. This results in a greater or lesser degree of difficulty in breathing, which is often of fairly acute onset and equally rapid disappearance. Asthma generally develops in childhood or at least is present during adolescence if it is going to develop at all, and it poses a threat to life only when it is so severe that ventilation is grossly reduced. The acute asthmatic attack is alarming both for the sufferer and the onlooker. There may be extreme difficulty in breathing out, and the chest assumes a more and more inspiratory position. In spite of the severe respiratory difficulty, the patient remains fully conscious and, because oxygen in the arterial blood is initially not depleted, appears of a normal colour. In some patients an acute asthmatic attack may be precipitated by psychological factors, but it seems clear that an organic mechanism underlies this disorder in all patients, and, although the reactivity of the bronchial tree may be affected by fatigue and anxiety, this disease is not to be regarded as primarily psychogenic in origin. In some patients a sensitivity to particular foodstuffs plays some part in this disorder, and this is probably particuarly true of children. In others, sensitivity to animals or to dust of all kinds may be a feature, but it may be said that the condition is characterized by a generally increased reactivity of the bronchial tree to any stimulus, and in many patients these stimuli need not be highly specific.

Status asthmaticus

The most dangerous form of the condition is known as status asthmaticus. In this situation, the severe bronchial spasm lasts for several days and finally the bronchi become obstructed by plugs of mucus, which completely fills the passages in the bronchial tree. This causes severe ventilatory difficulty and, by the time this stage is reached, the administration of epinephrine, which in the early stages usually relieves the attack, becomes ineffective. At this stage of the disease, the arterial oxygen tension begins to fall to dangerously low levels, and terminally the arterial carbon dioxide tension begins to rise. When this severe state of respiratory insufficiency is reached, artificial assistance may have to be given to ventilation, and close supervision is obligatory. If the condition progresses, the victim becomes unable to make the enormous respiratory effort needed to overcome the resyiratory obstruction., and with increasing failure to eliminate carbon dioxide and with increasing ;depletion of oxygen, the heart stops beating. As a result of recent work, it has become clear that occasionally heart stoppage may occur at a somewhat earlier stage of the condition before the ventilatory obstruction has been established for very long. There has recently been considerable anxiety that deaths from status asthmaticus, particularly in younger people, have been increasing, and it is not at all clear why this should be so. The management of patients with asthma calls for considerable skill and patience on the part of the physicians and may be materially assisted or made much more difficult by the psychological and domestic circumstances of the patient.

Rarely this disease may interrupt normal schooling and interfere with normal activities to such an extent that it casts a shadow over the psychological development of the individual personality. Much more commonly, it tends to diminish in severity with age and assisted by therapy may be quite compatible with an active life. So-called asthma developing in the middle age is much more commonly the occurrence of considerable bronchial obstruction as a consequence of chronic bronchitis, the latter disease usually resulting from cigarette smoking and other environmental factors. In such instances, although there may be some sensitivity to *Haemophilus influenzae*, which complicates chronic bronchitis, a true familial allergic spasmodic asthma is not present. The patient with spasmodic asthma commonly has a normal chest X-ray, and this condition does not lead to the development of lung destruction or emphysema unless it is associated with chronic bronchitis.

Characteristically, in asthmatic persons, even though physical evidence of pulmonary abnormality may not be found, it is possible to demonstrate that there are abnormalities of pulmonary function, particularly an abnormal distribution of inspired gas, indicating that there is a failure of uniform ventilation of different parts of the lung. These findings indicate that the lung of the asthmatic patient is not infrequently abnormal at a time when he feels well and is free of symptoms.

Hay fever

Hay fever and farmer's lung. Hay fever is an allergic condition affecting the nasal mucosa and sometimes associated with asthma or with other allergic conditions. It is commonly familial and occurs in its most severe form as an allergic response to the inhalation of pollens, particularly the pollen of ragweed. It may cause the nose to run with attacks of sneezing and may be associated with uncomfortable conjunctivitis. It commonly has a seasonal variation and is closely associated with the exposure of the patient to the pollen. Allergic inflammation of the nasal passages is a more severe and more chronic form and is mole commonly associated with asthma.

Farmer's lung

In recent years much attention has been given to the important disease farmer's lung. It is now known that this is a sensitivity reaction to the inhalation of spores that germinate in mouldy hay. In most farming areas of the Western world in which careful studies have been made, the disease has been shown to be much more common than hitherto supposed. In Wisconsin and the New England states, and in the west of England, as well as in France, it is not unusual in farming communities. Characteristically, the victim notices shortness of breath after he has been exposed to hay that has been stored over the winter, and that he may be putting out for feed for horses. The dyspnea may become more severe and he may develop a febrile illness with increasing shortness of breath. The X-ray shows a finely mottled appearance. At this stage of the disease, there is not only considerable airway obstruction but in addition interference with gas transport in the lung leading to a reduced arterial oxygen tension. Spontaneous resolution of this condition may occur, and its treatment is facilitated by the administration of adrenocorticosteroid compounds. It seems likely, however, that, if untreated, the disease may assume a more chronic form and may lead to permanent changes in the lung, causing considerable chronic interference with pulmonary function. A somewhat similar group of diseases have recently been described occurring in those with close contact with birds. Variously known as pigeon breeder's lung and bird fancier's lung, these represent different kinds of allergic response to proteins from birds, and particularly, probably, proteins contained in the excreta of pigeons, budgerigars, and canaries. The allergic response may be primarily in the bronchial tree, but it not uncommonly involves the lung parenchyma, and the patient may complain of shortness of breath and wheezing. It seems unlikely that this group of disorders leads to any permanent pulmonary damage, but removal of the ailing person from contact with birds is essential if he has developed a sensitivity to them.

Pigeon breeder's lung

In a number of other diseases an allergic response or sensitivity phenomenon plays an important part in determining the disease. This is true of byssinosis (see below ***Industrial lung disease)*** and it is also probably true of response to infection with some organisms. It is not known to what extent a sensitivity phenomenon may determine the bronchial spasms that result from an acute exacerbation of infection with ***Haemophilus influenzae*** in chronic bronchitis. There may be a sensitivity phenomenon associated with infection with the fungus ***Aspergillus,*** and it is possible that some of the respiratory disorders described in grain handlers may be based on a sensitivity phenomenon occurring in response to inhalation of dust from this material.

BRONCHITIS

Acute bronchitis. Acute bronchitis most commonly occurs as a consequence of virus infection. It may also be precipitated, however, by acute exposure to harmful gases, as for example chlorine and sulfur dioxide, and perhaps most commonly it occurs as an exacerbation of chronic bronchitis. The bronchial tree may be reddened and congested, and minor blood streaking of the sputum may occur in this condition. In most instances resolution is reasonably quick and full recovery occurs. If there is extensive chronic lung disease present, however, the additional acute inflammatory reaction occurring in large and smaller segments of the bronchial tree may be dangerous, and this sequence of events is not uncommonly observed in patients with chronic bronchitis and emphysema. Similarly, persons suffering from bronchiectasis—chronic dilation of the bronchial tubes—may suffer from acute exacerbations of infection, and this chronic disease may well be characterized by a sequence of such acute events.

Bronchiolitis obliterans is an acute disease of small airways. It probably occurs to some extent in acute viral disorders, and acute bronchiolitis is a well-recognized clinical syndrome in children between the ages of one and two years. In these infants, the cause is most commonly a virus, which gives rise to a syndrome of respiratory obstruction with gross enlargement of the chest and considerable difficulty in breathing. In severe instances it may threaten life, but it normally clears spontaneously and leaves the lung apparently undamaged. In adults acute bronchiolitis of this kind is not a well-recognized clinical syndrome, though there is little doubt that in most patients with chronic bronchitis acute exacerbations of infection are associated with further damage to small airways. Acute bronchiolitis obliterans in adults occurs most characteristically as a result of inhaling gas fumes. These may be oxides of nitrogen, which cause specific damage at the level of the small bronchioles. Workers may be exposed to this gas in a number of different situations. Oxides of nitrogen are generated in grain-filled silos, and a farmhand, being unaware of the danger, may be inadvertently exposed to this gas at the top of a silo. Oxides of nitrogen are generated as a result of underground explosions, and cases of acute bronchiolitis obliterans have occurred when workers have moved too quickly into an area where a dynamite charge has been set off below ground. The combustion of some plastic materials, particularly in enclosed spaces, may lead to the generation of oxides of nitrogen to a damaging concentration, and thus this syndrome may be seen in firemen or those who have been exposed to the fumes from fires. The syndrome is characterized by dyspnea, and resolution may be slow. The extent to which inflammation of the terminal bronchioles and its slow resolution occur in virus diseases is not fully understood, since the recognition of this condition in the absence of distinctive radiological change is not easy.

Industrial hazards

The syndrome of spasmodic asthma may in rare instances be difficult to control, and the obstruction it causes may become chronic. Spasmodic asthma is however not, on balance, a common cause of generalized obstructive lung disease, since it is most commonly remittent and periodic.

Chronic bronchitis and emphysema. The diseases chronic bronchitis and pulmonary emphysema, on the other hand, are much less liable to be remittent and much harder to relieve once they have become established. There is much concern about the incidence of these two diseases in modern industrial society because there is evidence that they have increased considerably over the past 20 years. In the United States, chronic bronchitis and emphysema appear to have increased as a cause of mortality and morbidity by a factor of 12 over the decade of the 1960s, and the rate of increase appears still to be unchecked. Similar increases have been noted in western Europe and in Canada, and there is much discussion of the relative importance of cigarette smoking, air pollution, and climate in determining their increasing incidence. There is general agreement that the major increase is most closely related to cigarette smoking, and the general incidence of these conditions in any population appears to correlate most closely with cigarette smoking. Other factors such as climate, industrial environment, air pollution, and even genetic factors may be responsible for some additional increase. Chronic bronchitis is characterized by changes in major bronchi, particularly hypertrophy of the mucous glands in the wall, changes in smaller bronchioles leading to obliteration and obstruction, and, finally, alterations in the distribution of ventilation and perfusion within the lung that may lead to altered tensions of carbon dioxide and oxygen in the blood. Chronic production of sputum is a feature, and this may occasionally be blood-stained. The relatively normal chest X-ray gives little indication of the severity of these changes until a late stage has been reached. Careful radiologic studies, however, using contrast material in the bronchial tree, may reveal the extent of the changes that have occurred at a relatively early stage.

Chronic bronchitis gives rise to a progressive fall in the ventilatory capability of the patient and leads to a progressive shortness of breath. Initially, this may only be noted during winter months or during times when the acute infection is worse than usual, but later the shortness of breath becomes chronic. Chronic bronchitis may continue to be a mild disorder for many years, and there is little doubt that in some victims chronic hypertrophy of the mucous glands may be productive of a blob of morning sputum, but the disease may progress little from that stage. In other cases, however, chronic bronchitis is complicated by more severe changes in the bronchial tree and the gradual destruction of the lung parenchyma of pulmonary emphysema. The disordered blood-gas tensions to which chronic bronchitis of itself may give rise and the development of pulmonary emphysema represent the two most severe complications of this condition. Even if chronic bronchitis does not advance to these more serious stages, the repetitive chest infections and periods away from work with fever and cough, the general ill health that may result from the chronic bronchial infection, and the susceptibility of such individuals to more severe chest infections cause considerable morbidity in the normal working population. The statistical data linking these events and the changes in the bronchial tree to cigarette smoking seem to be beyond dispute. It is not at all clear what determines the severity of the condition in different persons; it is evident that the relationship, like most other biological relationships, exhibits considerable individual variation. It is only in the last 20 years that the general features of chronic bronchitis have come to be recognized for what they are. Previously, a slight cough in the early morning would be dismissed as a smoker's cough, and physicians and the general public had little understanding of the potential severity of this condition if allowed to continue its progress uncontrolled.

Pulmonary emphysema

There has been a striking increase in the mortality rate from pulmonary emphyrema in North America during the decade of the 1960s. This disease exists in two distinct pathological forms. The first is called centrilobular emphysema; in this form a hole appears in the centre of each pulmonary lobule. Alveoli in this situation are apparently destroyed, and there are also inflammatory changes of a chronic nature in many of the terminal bronchioles. There is reason to suppose that this form of the disease is preponderantly found in cigarette smokers. Persons with this generalized type of disorder almost invariably have moderately severe chronic bronchitis, and it is not precisely known how much of the abnormality of arterial blood-gas tensions found in this condition is to be attributed to the bronchitis and how much to the pathological lesions of centrilobular emphysema. The second main type of pulmonary emphysema is known as panlobular. In this condition there appears to be a generalized loss of alveoli throughout the lobule and there is no selective destruction of the alveoli at the centre. This type of emphysema is commonly seen in older people and is the pathological type of the disease that has been shown to be associated with a deficiency in the blood of antitrypsin, which counteracts the activity of the digestive agent trypsin. Other types of emphysema are of less importance than these major varieties, but emphysema may occur along the dividing walls of the lung (septal emphysema) and in association with scars from other lesions. It is an essential feature of emphysema as now defined that it should involve evident destruction of lung parenchyma.

The exact cause of pulmonary emphysema is not yet fully understood. It is clear that multiple factors are involved, but its common association with chronic bronchitis suggests that chronic infection may well be important. Earlier theories that the lung destruction could be produced by blowing wind instruments or as a consequence of spasmodic asthma have now been shown to be incorrect. There is a general loss of lung elasticity as the normal lung ages, but it is not clear that this observation is relevant to the pathological lesions that may be found. Once emphysema of any variety is well established, there is severe ventilatory defect, and constant breathlessness is a feature of this condition. Since there is usually associated chronic bronchitis, there is often a productive cough, but in some instances this may not be a prominent feature of the condition. It is now known that emphysema in relatively young non-cigarette-smoking adults may occur as a consequence of a deficiency of antitrypsin in the serum, but it is not yet understood why the deficiency of this enzyme should lead to the development of panlobular emphysema.

Once the arterial blood–oxygen tension has become abnormally low and the carbon dioxide pressure elevated, an increased pressure in the pulmonary artery results and in this way the pulmonary lesion leads to heart failure. The average age of incapacity from emphysema, however, is in the mid-50s, and it is incorrect to describe this disease as a disease of advanced old age. Differentiating pulmonary emphysema from chronicbronchitis during life presents considerable difficulty. A combination of tests of pulmonary function, detailed radiological examination, together with a carefully taken history of shortness of breath usually enable a reasonably accurate diagnosis to be made. Before these methods were fully understood and developed, older physicians placed much reliance on the shape of the chest, but this can be seriously misleading, and little reliance can be placed on a visual inspection of the individual's chest as indicative of the change within the lung.

Unusual forms of pulmonary emphysema are occasionally reported. In one of these the disease may apparently involve one lung only. It is believed that this form of the disease is most commonly caused by an infection, probably of virus origin, occurring during childhood and preventing normal maturation of the lung on one side. "Congenital lobar emphysema" of infants is misnamed because there is no alveolar destruction. It is caused by overinflation of the lobe of the lung because of maldevelopment of cartilages in the wall of the major bronchus. Accurate descriptions of the structure of the lung with specimens of the organ taken in the inflated state have led to a much more accurate classification of this disorder than was hitherto possible, and many of the advances of understanding of this disease have occurred because of a quantitative approach to its pathology.

Bronchiectasis

Bronchiectasis is a disease believed usually to begin in childhood. As a consequence of whooping cough, some damage to the bronchial tree may occur; it reveals itself in adult life as a localized area of dilatation of major bronchi. These become chronically infected, and production of sputum and episodes of chest infection may become common. Apparently, the disease may also be caused by localized obstruction to the airway, and bronchiectasis may be a feature of congenital conditions or of inherited defects. The most important of these is the condition known as mucoviscidosis (or cystic fibrosis), an inherited genetic disorder characterized by deficiency of pancreatic excretion, abnormality of chloride secretion in the sweat, and pulmonary changes, particularly bronchiectasis and bronchiolar abnormality. This condition does not progress to lead to pulmonary emphysema but rather causes obliteration and fibrosis of small airways, and dilatation and infection of large bronchi. Thick viscid secretions found in the bronchial tree resist treatment. Bronchiectasis may also result from acute inhalation of irritants such as sulfur dioxide, and it may also occur as a result of other infections in the lung, particularly pulmonary tuberculosis of chronic form. In many of these conditions the treatment has been much improved by the use of antibiotics, but when lung damage is severe and recurrent infection cannot be controlled, surgical resection of the region may have to be undertaken.

GENERALIZED NONOBSTRUCTIVE LUNG DISEASE

The lung may be involved in generalized changes particularly involving the alveolar walls and not associated with chronic infection. These diseases primarily involve the alveoli of the lung and the capillary bed; they may progress to pulmonary fibrosis of considerable severity. One form of this group of disorders, known as diffuse interstitial fibrosis, is of unknown cause. A diffuse and progressive thickening of the alveolar walls occurs, associated with marked disturbances of pulmonary function and progressive X-ray change. This condition may occur in association with diseases of the connective tissue, such as scleroderma, and it may run a short or very protracted clinical course. Pulmonary alveolar proteinosis is another generalized lung disorder, in which characteristic protein-rich material accumulates within the alveolar spaces. It is not usually associated with irreversible changes in the lung and is a disease that may run a remittent course and finally resolve completely. It produces characteristic radiological change and is unassociated with any other changes in any other organs of the body. Its cause is unknown. Some diseases characterized by the formation of granulomas—nodules of chronically inflamed tissue—of unknown origin are important in relation to the lung. Boeck's sarcoid, the commonest of these, is a disorder usually involving many systems of the body in which small granulomas may be found. Very commonly the lungs are the primary source of involvement, and the disorder may involve predominantly the lymph nodes in the chest, or the lung parenchyma. When this is affected, small granulomatous lesions are found, particularly around small blood vessels in the lung, and they may be of sufficient number to cause change visible in X-rays. Spontaneous resolution of this condition is the most likely clinical outcome, but some residual lung damage is not unusual. Exceptionally severe pulmonary fibrosis may occur, leading to incapacitating lung disease; in such instances there are usually severe changes in other organ systems. Eosinophilic granuloma (also known as histiocytosis X) is a disease in which the lungs may become infiltrated with eosinophil cells—white blood cells that stain readily with eosin. Progressive fibrosis may occur as a consequence of this, leading eventually to formation of small cysts in the lung and considerable fibrotic change. This disorder tends to be self-terminating but may leave the lung substantially damaged.

Diseases that affect blood vesselsin lungs

In rare instances, the lung may be involved as part of the generalized rheumatoid process, and pulmonary changes in association with rheumatoid arthritis have been described. It is clear, however, that this is a rare event. When it occurs, the changes in the lung appear to be similar to those occurring in diffuse interstitial fibrosis, but on a patchy and uneven basis. Certain diseases that attack the vascular system and which are believed to be related to abnormalities of sensitization and immune response are prominent in the lung. Among these, the acute inflammatory disease periarteritis nodosa is an important cause of varying degrees of blood-vessel inflammation within the lung. Acute hemorrhagic pneumonitis may occur in association with changes in the kidney (Goodpasture's syndrome), and pulmonary hemorrhage occur as part of the clinical condition of pulmonary hemosiderosis, the accumulation of the iron-containing substance hemosiderin in the tissues. In most of these conditions it seems likely that the primary basis for lung involvement is an abnormality of the normal immune mechanism affecting the small vessels and capillaries of the lung. In some of these syndromes the pulmonary involvement may be so severe as to threaten life, but usually it is part of general system involvement, and the outcome may be determined by the extent of involvement in other organs.

This group of disorders is of relatively recent general recognition. Accurate diagnosis of them has been much improved by the development of pulmonary function tests, refinements in X-ray technique, and the improvements in thoracic surgery and anesthesia that have made a biopsy of the lung a much less dangerous procedure than formerly. In the past, many of these diseases were confused with each other, and accurate understanding of some of them has only been possible with the advances in understanding of immunological processes that have occurred in the past ten years.

PULMONARY RESPONSE TO CHEMICAL AND PHYSICAL IRRITANTS

Industrial lung disease. It has been known for many years that silica has a particularly irritant effect on the lung. The dangers of silicosis are generally well recognized, and the problems caused by this dust have been reduced by better engineering techniques in underground mining. Silica produces a distinctive reaction in the lung that leads eventually to masses of fibrous tissue and distinctive nodules of thick fibrosis, which, by contracting, distort and damage the lung. Silica is to be regarded as a hazard of any occupation in which dust containing this compound is free in the atmosphere. Hard-rock drilling underground, various industrial processes involving grinding and molding, and sandblasting are all examples of such potentially hazardous occupations. Silicosis is usually fairly easily detectable on a chest X-ray; it causes considerable impairment of lung function with shortness of breath as its characteristic feature. Coal dust alone, even if its silica content is very low, causes a distinctive pattern of lung change known as coal-worker's pneumoconiosis. Initially the dust is deposited in the terminal bronchioles, and it may lead to a fibrotic reaction around these. Later the disorder progresses to a more generalized form (lung disease caused by dust), and in some instances may progress to a stage known as progressive massive fibrosis. It is not clear whether this stage may be reached by dust alone or by some interaction between the dust and pulmonary tuberculosis. In both silicosis and coal-worker's pneumoconiosis, it is known that the disease may progress after the man has left the working environment in which exposure occurred, probably as a result of progressive fibrosis occurring within the lung.

The dangers of exposure to asbestos consist of the development of a diffuse fibrosis of the lung in response to inhalation of asbestos fibres, and secondarily some increased risk of development of lung cancer as a result of asbestos inhalation. This mineral is being increasingly used in modern industry and has proved valuable in the manufacture of such things as acoustic ceiling tiles and in all kinds of insulating work. The reason why this particular fibre causes a considerable reaction within the lung is not understood, but increasing attkntion is being paid to the proper engineering techniques in its manufacture and processing to avoid any major exposure to airborne fibres.

Organic dust has been shown to be responsible for various syndromes involving the lung. The dust produced when raw cotton is being processed causes a reversible airway obstruction, but continued exposure may lead to some permanent impairment of lung function. This disorder, known as byssinosis, has been shown to be present in cotton workers in several different countries. Workers in sugar cane may be affected by a somewhat similar syndrome (bagassosis) and sisal workers are also exposed to fibres that may cause a sensitivity reaction and considerable airway obstruction. In most of these diseases, the recognition of the cause and the modification of the industrial process or removal of the worker from the environment produces amelioration of symptoms.

There are many chemical irritants that are exceedingly damaging to the lung in high concentration. These include oxides of nitrogen, ammonia, oxides of sulfur, chlorine, ozone, gasoline vapour, and benzene. Acute episodes of poisoning by these gases are well documented, but relatively little is known of the long-term chronic effects of exposure to very low concentrations of them. In industrial processes involving their use, the hazard is well recognized. They mostly produce an inflammatory reaction in the bronchial tree coupled with accumulation of fluid in the lungs. If the patient survives the acute episode, there may be residual damage to the bronchial tree leading to bronchiectasis, there may be obliteration of small radicles of the bronchial tree, and in certain other instances there may be some degree of pulmonary fibrosis.

Radiation damage

The lung may be damaged by excessive radiation. Radiation therapy for malignant disease of the breast may cause transient pneumonitis in the lung lying within the irradiated area, and in the majority of such instances the lung heals completely and without any fibrotic change. In a small minority of patients, the pulmonary reaction may be more severe and may result in extensive pulmonary fibrosis and contraction and shrinkage of the lung with concomitant dyspnea. There is considerable individual variation in reaction to the same radiation dose, and there is still much to be learned as to what determines the radiation sensitivity of a particular organ in a given individual.

The prompt recognition of the dangers of beryllium, which causes severe pulmonary disease, led to the discontinuance of its use in the manufacture of fluorescent light tubes. Although the dangers of many classical materials such as silica and asbestos are well understood, the industrial physician and toxicologist is continuously confronted with new compounds that may have a deleterious effect on the lung and bas to be alert to detect these. One such substance, toluene diisocyanate, used in the manufacture of polystyrene foam (an expanded plastic used for insulation and packing), caused a transient flulike syndrome associated with considerable reversible airway obstruction. Prompt recognition of this syndrome led to modifications in the process and discontinuance of exposure of men to these particular fumes. Although the acute effects of many of these substances in high concentration are well documented, there is very little information, indeed, concerning the importance or unimportance of regular exposure to very low concentrations of gases such as oxides of nitrogen and sulfur dioxide, and of airborne metals such as lead, which occur as part of the polluted urban environment in modern Western nations.

Localized obstructive **lung** disease. Acute laryngeal obstruction and asphyxiation may be caused by the accidental inhalation of food, which sometimes happens when an anesthetic has to be administered for emergency surgery shortly after a meal has been taken, or which may occur accidentally during a meal, particularly if aperson is elderly. More common aspiration is into a bronchus, and classically this occurs in children, often when they are eating peanuts or other food of this kind. Such obstruction may cause overinflation of the lung supplied by the bronchi, or in other circumstances may cause a collapse of the affected region, which may be a quite small segment. It is likely that small regions of atelectasis or collapse precipitate the pneumonia of old people who have to be recumbent for several days. Occasionally the aspiration of a foreign body is followed by acute infection in the lung; in these circumstances a lung abscess may result. In the past, this was a much feared complication of pneumonia because it was difficult to treat and often extended to produce inflammation in the pleural cavity leading to empyema. In the absence of effective chemotherapy, these were dangerous complications. Abscess formation in the lung may occur rapidly in infants with staphylococcal pneumonia, but complete healing normally occurs and the acute infection can be well controlled by antibiotics. Abscesses now occur more commonly in association with cancer of the lung, as a consequence of pulmonary infarction (death of a section of lung tissue), or as a result of long-standing inflammatory processes such as tuberculosis and histoplasmosis, infection with the fungus *Histoplasma capsulatum.*

CONGENITAL AND DEVELOPMENTAL ABNORMALITIES

A wide variety of malformations occur in the lung. These include congenital absence of the whole lung, failure of different parts of the pulmonary artery to develop, anomalies in drainage of pulmonary veins, and failure of the cartilages to develop in a lobar bronchus, giving rise to infantile emphysema. **A** generalized cystic condition of the lung may occur as a congenital defect, but it is a very rare condition and many of the cases previously described under this heading are now considered to be instances of acquired cystic bronchiectasis. A remarkable inherited condition is that of microlithiasis alveolaris in which the alveoli are filled with small concretions. This gives rise to a spectacular chest X-ray that may be mistaken for other conditions. In general, this defect produces little interference with pulmonary function though a later pulmonary fibrosis has been described in association with it. The recent description of a syndrome of panlobular pulmonary emphysema in association with deficiency of serum α_1-antitrypsin has been noted above in the section on emphysema.

Respiratory-distress syndrome of newborn

Respiratory-distress syndrome, an important disorder of the newborn infant, appears to be either a failure to produce the surfactant lining of the lung at the time of birth, or a destruction of this factor. Whichever of these processes is fundamental to the condition, the infant born with a deficiency of this substance suffers from progressively severe atelectasis—collapse—of the lung. Failure of the lung to re-expand normally may lead within a few hours to a condition of acute respiratory failure. This condition is an important cause of neonatal mortality; although great advances have been made in understanding its essential nature, methods of preventing its development are still unknown.

OTHER CAUSES OF BRONCHOPULMONARY DISEASE

Circulatory disorders. The lung is commonly involved in disorders that are primarily those of the circulation. An instance of this is the pulmonary disease associated with failure of the left ventricle. The left ventricle is responsible for the pumping of blood into the main circulation. If the action of this ventricle becomes deficient—*e.g.*, as a result of disease of the coronary arteries or as a consequence of high blood pressure, the lungs may be subjected to an elevated pressure, because the right ventricle continues to put blood into the pulmonary artery but the outflow from the lung into the left atrium is impeded by the failing function of the left ventricle. This gives rise to pulmonary congestion. If the congestion becomes sufficiently severe, fluid passes through the alveolar wall, leading to pulmonary edema, and this may be a fatal complication of an acute heart attack. The shortness of breath that may accompany systemic hypertension is believed to be mainly due to pulmonary congestion, causing an increased stiffness of the lung and, perhaps in addition, obstructing airflow through the collection of edema fluid around small airways.

Acute pulmonary embolism

The lung is by its nature particularly susceptible to blockage of its blood vessels by circulating blood clots. These originate either in the veins of the legs or in the veins of the pelvis; such emboli classically may become detached a week or ten days after a major surgical operation or injury, and be carried to the lung, where they acutely block some part of the pulmonary vasculature. This condition of acute pulmonary embolism may be immediately fatal, or it may give rise to infarction of part of the lung, which may be accompanied by acute pleuritic pain and spitting of blood from the lungs. In most instances full recovery of lung function occurs after such an event. Similarly, fat emboli to the lung may occur after fractures to major bones; these may cause a deficiency of arterial blood oxygenation for a few days but usually full recovery occurs. In some uersons, either a process of progressive thrombosis or unresolved embolization leads to a chronic state of severe increased blood vressure in the pulmonary circuit. This may be so severe as to eventually cause heart failure. A variant of this condition occurs in schistosomiasis, a tropical disease common in Egypt and parts of West Africa in which the pulmonary arterial tree becomes blocked by ova of the parasite.

In some patients with chronic lung disease, both of the fibrotic and the destructive type, there may be a considerable rise of pressure in the pulmonary artery. The pulmonary vessels are reactive to deficiency of oxygen, and it seems likely that it is the abnormal gas concentrations in the alveoli that play the major part in determining this response. The presence of pulmonary hypertension in normal subjects occurs at altitude when there is alveolar hypoxia, and the development of cor pulmonale — heart failure consequent upon lung disease—appears to be an extension of this naturally occurring phenomenon in many cases.

Lung cancer. Cancer of the lung 40 years ago was a relatively rare condition. The increase in its incidence noted after World War II was at first ascribed to better diagnostic technique, but by 1952 it had become clear that the rate of increase of the condition was too great to be accounted for in this way. At that time the first evidence was obtained that indicated that a heavy cigarette smoker was more likely to develop lung cancer than a nonsmoker. In recent years there has been a tremendous volume of epidemiological work covering many different populations in many different countries that has clearly shown that a heavy cigarette smoker is more likely to develop lung cancer than a nonsmoker, and that the risk of developing this disease depends on the duration and intensity of cigarette smoking. By contrast, cigar smoking and pipe smoking seem to carry very little additional risk of lung cancer.

Cigarette smoking and cancer

By 1965, cancer of the lung and bronchus accounted for 43 percent of all cancers in United States men, an incidence nearly three times greater than that of the second commonest cancer (cancer of the prostate) in men, which accounted for 16.7 percent of total incidence.

The ***Report of the Advisory Committee to the Surgeon General of the Public*** *Health* ***Service*** (1964), stated:

> (1) Cigarette smoking is causally related to lung cancer in men; the magnitude of the effect of cigarette smoking far outweighs all other factors. The data for women, though less extensive, point in the same direction. (2) The risk of developing lung cancer increases with duration of smoking and the number of cigarettes smoked per day, and is diminished by discontinuing smoking. (3) The risk of developing cancer of the lung for the combined group of pipe smokers, cigar smokers, and pipe and cigar smokers is greater than for nonsmokers, but much less than for cigarette smokers.

The data in the third section are insufficient to warrant a conclusion for each group individually.

The reason for the carcinogenicity of tobacco smoke from cigarettes is not known. Tobacco and tobacco smoke contain a large number of carcinogenic materials and, although it is presumed that the "tars" in cigarette smoke probably contain a substantial fraction of the cancer-causing condensate, it is far from being established which of these is responsible. Although the tobacco industry has concentrated its efforts on removing the tar content as much as possible from cigarette smoke, it is too early to know whether or not this has had any effect on the carcinogenicity of cigarettes. There is a presumption that cigarette smoking may be additive to other factors, but the possible role of air pollution or of exposure to industrial irritants remains in general unclear. In spite of some advances in diagnostic technique and general advances in pulmonary surgery, the outlook for a person found to have this condition is still severely limited. It is a characteristic of lung cancer that, by virtue of its site, it may disseminate generally in the body at a time when the primary tumour is still too small to cause local symptoms. Thus the first manifestation of a lung cancer may be the development of a secondary tumour in the brain, in the bones, or in the liver. Usually the lymph nodes draining the affected area of lung are involved relatively early, and the diagnosis may first be suggested by the presence of enlarged nodes in the neck or mediastinum, the partition of membrane that divides the chest into right and left compartments. Lung cancer not infrequently occurs in patients with chronic bronchitis who already have a productive cough. If any cigarette smoker suddenly develops a cough and this does not clear up within a few days, an investigation should be made; in some patients it is the first appearance of hemoptysis added to chronic production of sputum that indicates the presence of lung cancer. The condition may also begin by causing a pneumonia that proves difficult to treat, or, if the tumour is situated peripherally in the lung, the first evidence of it may be a pleural effusion. The diagnosis essentially depends on securing tissue for pathological examination. Routine chest X-rays in the cigarette-smoking population have been advised for early detection of lung cancer, and increasing use has been made of a careful study of the sputum for the presence of malignant cells. Examination of the bronchial tree by bronchoscopy may make the diagnosis possible, but in some instances a lesion visible on the chest X-ray must be removed from the lung directly before the diagnosis can be established.

It is largely because of the demonstrated increase in lung cancer that many countries have begun to limit cigarette advertising and have insisted that cigarette packages be labelled with a warning to the smoker. In spite of advances in radiation therapy and a much lower generalized risk from operations on the lung, the mortality from the condition remains high even if it is diagnosed early. Long-term cure is unfortunately the exception rather than the rule, and, if all cases of lung cancer are considered, the five-year survival rate from the first time the disease is diagnosed is, in most reported series, less than 20 percent of the total. Although in the vast majority of cases of lung cancer the tumours consist of scalelike cells such as those in squamous epithelium, there are other types of lung cancer probably not related to the cigarette in their etiology. Alveolar cell cancer is a slowly spreading condition that seems to affect men and women in equal proportion and does not seem to be related to cigarette smoking. Pulmonary adenocarcinoma (a type of cancer with gland-like structures) of the lung also has a more equal sex incidence, and its relationship to cigarette smoking is much less well established than in the case of the squamous-cell type of cancer.

DISEASES OF THE NONPULMONARY STRUCTURES

Diseases of the pleura. The lungs move in an airtight space, the thorax lined on its inner side with a layer of smooth fibrous tissue. The external surface of the lung itself is covered by another layer of pleura and the space between these two membranes is occupied by only a small volume of fluid for lubrication purposes under normal circumstances. Rupture of the pleura covering the lung means that this space communicates with the airway and the lung, by virtue of its intrinsic recoil, consequently collapses. In some individuals the lung, particularly at its apex, contains some small bullae, or blisters, and it is not uncommon for one of these to rupture causing a spontaneous pneumothorax, a disease particularly of people in their second and third decade of life. Most commonly, this condition resolves spontaneously, but it may cause acute onset of chest pain and some shortness of breath and necessitate several weeks of rest while the lung re-expands. It usually does not indicate the presence of any underlying lung disease, but if recurrences occur, it may be necessary to seal the pleura or to remove the few blebs, or small liquid-filled blisters, at the top of the lung, which are responsible for the trouble. The primary diseases of the pleura are almost all inflammatory in nature. A pleurisy with an effusion may be the presenting symptom of pulmonary tuberculosis and pleurisy may accompany any kind of pneumonia. Acute inflammation of the-pleura normally produces pain when the patient attempts to take a deep breath, but once the effusion has developed, the pain disappears. Pleural effusions may also accompany pulmonary embolization and pulmonary infarction. The pleura may be attacked by primary or secondary malignant disease, and rarely it may be invaded by direct extension of processes from the abdominal cavity. When a pleural effusion in pneumonia becomes secondarily infected, a condition of empyema results with pus in the pleural cavity. This complication, dreaded before the antibiotic era, gave rise to drainage of affected material from a sinus in the chest for months or even years after the acute event. Such episodes are now rarely seen as a result of acute infections, but draining sinuses still may occur as a consequence of tuberculosis or of fungus disease. Infection of the lung and later the pleural cavity by the fungus ***Actinornyces*** was particularly prone to produce a draining empyema and sinus.

Diseases of the mediastinum and diaphragm. Important structures in the thorax run through the mediastinum, and disease in them may produce changes within the mediastinum. Perforation of the esophagus by a foreign body such as a fish bone may give rise to a mediastinitis, and this may be a life-threatening condition. Cancer of the esophagus not uncommonly spreads to involve the mediastinum, and enlargement of lymph glands in the mediastinum commonly occurs in a wide variety of conditions involving the lung. A mediastinitis may also be caused by radiation therapy aimed at the chain of lymph glands in the mediastinum and this may lead to some residual fibrosis with involvement of structures within.

Diseases affecting diaphragm

There may be congenital abnormalities of the diaphragm, but the commonest disease that involves it is paralysis caused by interruption of the motor nerve of the diaphragm, the phrenic nerve, at some point in the neck or in the mediastinum. Accidental injury to this nerve during operations on the neck, or involvement of the nerve in disease in the mediastinum, may cause paralysis of one side of the diaphragm. This does not cause severe effects and may be demonstrated only by X-ray. A bilateral paralysis of the diaphragm causes some impairment of pulmonary function but is a relatively rare condition. Herniation of the diaphragm is an important development defect that may require reparative surgery. In extreme cases the stomach may be mostly within the thorax, and there are many variations of abnormality that may occur, but they all lead to considerable impairment of function of the affected side.

BIBLIOGRAPHY. H.C. HINSHAW, *Diseases of the Chest,* 3rd ed. (1969), a general textbook covering all types of chest diseases, including their diagnosis and treatment; D.V. BATES, P.T. MACKLEM, and R.V. CHRISTIE, *Respiratory Function in Disease,* 2nd ed. (1971), a detailed text on impairment of lung function caused by disease; A.C. STERN (ed.), *Air Pollution,* 2nd ed., 3 vol. (1968), an authoritative and encyclopaedic analysis of every aspect of air pollution, including its effects on the respiratory system; R.G. FRASER and J.A.P. PARE, *Diagnosis of Diseases of the Chest* (1970), an up-to-date text on the diagnosis of different forms of lung disease.

(D.V.B.)

Restaurant

The public dining room that came ultimately to be known as the restaurant originated in France, and the French have continued to make major contributions to the restaurant's development. The first restaurant proprietor is believed to have been one A. Boulanger, a soup vendor, who opened his business in Paris in 1765. The sign above his door advertised restoratives, or *restaurants*, referring to the soups and broths available within. The institution took its name from that sign, and "restaurant" now denotes a public eating place in the English, French, Dutch, Danish, Norwegian, Romanian, and many other languages with some variations. For example, in Spanish and Portuguese the word becomes *restaurante*; in Italian it is *ristorante;* in Swedish, *restaurang*; and Russian, *restoran*; in Polish, *restauracia.*

Although inns and hostelries often served paying guests meals from the host's table, or table d'hôte, and beverages were sold in cafés, Boulanger's restaurant was probably the first public place where any diner might order a meal from a menu offering a choice of dishes.

The first luxury restaurant

Development of the restaurant. Boulanger operated a modest establishment; it was not until 1782 that La Grande Taverne de Londres, the first luxury restaurant, was founded in Paris. The owner Antoine Beauvilliers, a leading culinary writer and gastronomic authority, later wrote *L'Art du cuisirzier* (1814) a cookbook that became a standard work on French culinary art. Beauvilliers achieved a reputation as an accomplished restaurateur and host, and the French aphorist and gastronomic chronicler, Jean-Athelme Brillat-Savarin, a frequent guest, credited Beauvilliers with being

> the first to combine the four essentials of an elegant loom, smart waiters, a choice cellar, and superior cooking.

Brillat-Savarin also noted that Beauvilliers would

> point out here a dish to be avoided, there one to be ordered instantly . . . ; and send, at the same time, for wine from the cellar, the key of which he produced from his own pocket; in a word, he assumed so gracious and engaging a tone, that all these *extra* articles seemed so many favours conferred by him.

Before the French Revolution, aristocratic French households maintained elaborate culinary establishments, but when the revolution reduced the number of private households offering employment, many chefs and cooks found employment in restaurant kitchens or opened their own eating establishments. By 1804 Paris had over 500 restaurants, producing most of the great chefs of history and creating many famous dishes.

French restaurants of the 19th century. During the Napoleonic era the Palais-Royal, the colonnaded, tree-lined area adjacent to the Louvre, became the location of many of the finest restaurants in Paris. The Véry, a leading restaurant of the era, was lavishly decorated and frequented by beautiful women escorted by dashing officers. The menu listed a dozen soups, two dozen fish dishes, 15 beef entrees, 20 mutton entrees, and scores of side dishes. The novelist Honor6 de Balzac often dined at the Véry, consuming prodigious quantities of oysters, fish, meat dishes, fruits, wines, and liqueurs. It was a favourite haunt of gourmet-author, Grimod de la Reynière, who considered it the finest restaurant in France.

The Véry was absorbed in 1869 by the neighbouring Le Grand Véfour. This restaurant was still in business in the early 1970s and is regarded as one of the finest eating places in France. Another outstanding Paris establishment of the 19th century was the Café Foy, later called Chez Bignon, a favourite dining place of the English novelist William Makepeace Thackeray and of the Italian composer Gioacchino Rossini, who lived in the same building. The Café de Paris, on the Boulevard des Italiens, was the first of many restaurants in Paris and elsewhere that have operated under this name. Other favourite eating places were the Rocher de Concale, on the rue Montorgueil, famous for its oysters and fish, and the Restaurant Durand, at the comer of the Place de la Madeleine and the rue Royale, a favourite gathering place of politicians, artists, and writers, including the authors Anatole France and Émile Zola.

The most illustrious of all 19th-century Paris restaurants was the Café Anglais, on the Boulevard des Italiens at the comer of the rue Marivaux, where the chef, Adolphe Dugléré, created classic dishes such as *sole Dugléré* (filets poached with tomatoes and served with a cream sauce having a fish stock base) and the famous soup *potage Germiny.* On June 7, 1867, the Café Anglais served the now-famous "Three Emperors Dinner" for three royal guests visiting Paris to attend the Universal Exposition. The diners included Tsar Alexander II of Russia; his son the tsarevich (later the tsar Alexander III); and King William I of Prussia, later the first emperor of Germany. The meal included soufflés with creamed chicken (d *la reine*), fillets of sole, escalloped turbot, chicken *à la portugaise* (cooked with tomatoes, onions, and garlic), lobster *à la parisienne* (round, flat medallions glazed with a gelatin-mayonnaise mixture and elaborately decorated), ducklings *à la rouerznaise* (the carcasses stuffed with liver and pressed, presented on a platter with boned slices of the breast and the grilled legs, and served with a red wine sauce containing pureed liver), ortolans (small game birds) on toast, and eight different wines.

Although the Café Anglais closed in 1913, when the building was demolished, the table setting for this dinner is now displayed at La Tour d'Argent, the oldest surviving restaurant in Paris.

Toward the end of the 19th century, in the gaudy and extravagant era known as *la belle époque,* the luxurious Maxim's, on the rue Royale, became the social and culinary centre of Paris. The restaurant temporarily declined after World War I but recovered under new management, to become an outstanding gastronomic shrine.

Kitchen teams

France produced many of the world's finest chefs, including Georges-Auguste Escoffier, who organized the kitchens for the luxury hotels owned by César Ritz, developing the so-called *brigade de cuisine,* or kitchen team, consisting of highly trained experts each with clearly defined duties. These teams included a chef, or *gros bonnet,* in charge of the kitchen; a sauce chef, or deputy; an *entremettier,* in charge of preparation of soups, vegetables, and sweet courses; a *rôtisseur* to prepare roasts and fried or grilled meats; and the *garde manger,* in charge of all supplies and cold dishes. In Escoffier's time, the duties and responsibilities of each functionary were sharply defined, but in modern times the rising labour costs and the need for faster service have broken down such rigidly defined duties. In the kitchens of even the leading modern restaurants, duties at the peak of the dinner-hour preparations are likely to overlap widely, with efiiciency maintained amid seeming chaos and confusion.

French restaurants in the 20th century. In the 20th century, with the development of the automobile, country dining became popular in France, and a number of fine provincial restaurants were established. The Restaurant de la Pyramide, in Vienne, regarded by many as the world's finest restaurant, was founded by Fernand Point and after his death, in 1955, retained its high standing under the direction of his widow, Madame "Mado" Point. Other leading French provincial restaurants have included the Troisgros in Roanne; the Paul Bocuse Restaurant near Lyons; the Auberge de I'll in Illhaeusern, Alsace; and the hotel Cote d'Or, at Saulieu.

Restaurants throughout France are evaluated annually by the *Guide Michelin,* a publication devoted to surveying eating establishments and hotels in more than 3,400 towns and cities, and awarding one, two, or three stars, based upon quality. One star denotes, "a good restaurant in its class"; two stars mean "excellent cuisine, worth a detour"; three stars, the top rating, signify "one of the best tables in France, well worth the journey."

French restaurants today are usually in one of three categories: the bistro, or brasserie, a simple, informal, and inexpensive establishment; the medium-priced restaurant; and the more elegant grand restaurant, where the most intricate dishes are executed and served in luxurious surroundings.

Restaurants of other countries. Other nations have also made many significant contributions to the development of the restaurant.

European restaurants. In Italy the *botteghe* (coffee shop) of Venice originated in the 16th century, at first serving coffee only, later adding snacks. The modern *trattorie,* or taverns, feature local specialities. The *osterie,* or hostelries, aie informal restaurants offering homelike cooking. In Florence small restaurants below street level, known as the *buca,* serve whatever foods the host may choose to cook on a particular day.

Austrian coffeehouses offer leisurely, complete meals, and the diner may linger to sip coffee, read a newspaper, or even to write an article. Many Austrians frequent their own "steady restaurants," known as *Stammbeissl.*

In Hungary the *csárda,* a country highway restaurant, offers menus usually limited to meat courses and fish stews.

The beer halls of Czechoslovakia, especially in Prague, are similar to coffeehouses elsewhere. Food is served, with beer replacing coffee.

The German *Weinstube* is an informal restaurant featuring a large wine selection, and the *Weinhaus,* a food and wine shop where customers may also dine, offers a selection of foods ranging from delicatessen fare to full restaurant menus. The *Schenke* is an estate-tavern or cottage pub serving wine and food. In the cities a similar establishment is called the *Stadtschenke.*

In Spain the bars and cafes of Madrid offer widely varied appetizers, called tapas, including such items as shrimp cooked in olive oil with garlic, meatballs with gravy and peas, salt cod, eels, squid, mushrooms, and tuna fish. The tapas are taken with sherry, and it is a popular custom to go on a *chateo,* or tour of bars, consuming large quantities of tapas and sherry at each bar. Spain also features the *marisco* bar, or *marisqueria,* a seafood bar; the *asadoro,* a Catalan rotisserie; and the *tasca,* or pub-wineshop.

In Portugal, *cervejarias* are popular beer parlors also offering shellfish. Fado taverns serve grilled sausages and wine, accompanied by the plaintive Portuguese songs called *fados* (meaning "fate").

The smörgåsbord

In Scandinavia sandwich shops offer open-faced, artfully garnished sandwiches called *smørrebrød.* Swedish restaurants feature the *smorgdsbord,* which literally means "bread and butter table," but actually is a lavish, beautifully arranged feast of herring, shrimp, pickles, meatballs, fish, salads, cold cuts, and hot dishes. The food is taken with aquavit or beer.

The Netherlands has sandwich shops, called *broodjeswinkels,* serving open-faced sandwiches, seafoods, hot and cold dishes, and cheeses from a huge table.

English city and country pubs have three kinds of bars: the public bar, the saloon, and the private bar. Everyone is welcome in the public bar or saloon, but the private bar is restricted to habitues of the pub. The food served in pubs varies widely through England, ranging from sandwiches and soups, to pork pies, veal and ham pies, and steak and kidney pies, bangers (sausages) and a pint (beer), bangers and mash (potatoes), toad in the hole (sausage in a Yorkshire pudding crust), and Cornish pasties, or pies filled with meat and vegetables.

Middle Eastern restaurants. In the *tavérnas* of Greece, customers are served such beverages as retsina, a resinated wine, and ouzo, an anise-flavoured aperitif, while they listen to the music of the *bouzouki.* Like other Mediterranean countries, Greece has the grocery-tavtrna where one can buy food or eat.

The Turkish *işkembeci* is a restaurant featuring tripe soup and other tripe dishes; *muhallebici* shops serve boiled chicken and rice in a soup and milk pudding.

Asian restaurants. In Japan, *tempura* bars offer fried shrimp at a counter. Other food bars serve *sushi,* raw fish fillets or other ingredients rolled in vinegared rice, and *sashimi,* raw fish slices. *Yu-dōfa* restaurants build their meals around a variety of bean curds, and the elegant tea houses serve formal Kaiseki table d'hôte meals that amount to a Japanese haute cuisine.

In China, restaurants serving the local cuisine are found, and noodle shops offer a wide variety of noodles and soups. The *dim-sum* shops provide a never-ending supply of assorted steamed, stuffed dumplings and other steamed delicacies.

A common sight in most parts of Asia is a kind of portable restaurant, operated by a single person or family from a wagon or litter set up at a particular street location, where specialties are cooked on the spot. Depending upon the country, the food and cooking utensils vary widely.

The cafeteria

U.S. sontributions to restaurant development. The cafeteria, a U.S. contribution to the restaurant's development, originated in San Francisco during the gold rush. Featuring self-service, it offers a wide variety of foods displayed on counters. The customer makes his selections, paying for each item as he chooses it or paying for the entire meal at the end of the line. Other types of quick-eating places originating in the U.S. are the drugstore counter, serving sandwiches or other snacks, and the lunch counter, where the diner is served a limited quick-order menu at the counter. Coin-operated vending machines also offer snacks and beverages.

The specialty restaurant, serving one or two special kinds of food, such as seafood or steak, is another distinctive U.S. establishment.

The Pullman car diner, serving full-course meals to long distance railroad passengers, and the riverboat steamers, renowned as floating gourmet palaces, were original American conceptions. They belong to an earlier age, when dining out was a principal social diversion, and restaurants tended to become increasingly lavish in food preparation, decor, and service.

In many modern restaurants customers now prefer informal but pleasant atmosphere and fast service. The number of dishes available, and the elaborateness of their preparation, has been increasingly curtailed as labour costs have risen and the availability of skilled labour decreased. The trend is toward such fast and efficient food operations as snack bars, cafeterias, coffee shops, and drive-ins. The trend in elegant and expensive restaurants is toward smaller rooms and intimate atmosphere, with authentic, highly specialized and limited menus.

BIBLIOGRAPHY. RICHARD HERING, *Lexikon der Kiiche,* 11th ed. *(1957;* Eng. trans., *Dictionary of Classical and Mod-*

ern Cookery, 1958), a comprehensive chefs reminder with brief recipes (sometimes several versions of same); A. DUMAS, *Grand dictionnaire de cuisine* (1873), a dictionary that has influenced gastronomic literature and practice for more than a century; GEORGES AUGUSTE and PHILEAS GILBERT ESCOFFIER, and EMILE FETU, *Le Guide Culinaire* (1903), a definitive book of its kind; A. KETTNER, *Kettner's Book of the Table* (1877, reprinted 1968), the finest Victorian document about the complexities of the kitchen; LA VARENNE, *Le Cuisinier François* (1651) and *La Cuisine Methodique* (1662), two books considered to be the foundation of the French Grande Cuisine; CHARLES RANHOFER, *The Epicurean* (1894, 1920), a mammoth undertaking written by the chef of the original Delmonico's, New York City, including the bills of fare of that restaurant from 1862 to 1894; C. HERMAN SENN, *The Art of the Table, Including How to Wait at Table, How to Fold Napkins and How to Carve,* 3rd ed. enl. (1923), another part of gastronomy dealt with on a professional level; CURNONSKY (ed.), *Cuisine et vins de France* (1953), a definitive work discussing gastronomic regions of France; EUGENE BRIFFAULT, *Paris à table* (1846), one of the best descriptions of mid-19th-century Parisian eating patterns; E. RICHARDIN, *La Cuisine française du XIV^e au XX^e siècle* (1907), a sound treatise on the state of French gastronomy from about 1800 to the early 1900s; *Larousse gastronomique,* rev. ed. (1961), an encyclopedia of food and wine; *Funk and Wagnalls Cooks' and Diners' Dictionary: A Lexicon of Food, Wine, and Culinary Terms* (1969). (G.L.)

Reticuloendothelial System, Human

The term reticuloendothelial system (RES) denotes certain tissue cells that occur in widely separated parts of the body and have the specific ability to take up particulate substances. Should these substances be noxious, as are some bacteria, they can be destroyed by the cells of the reticuloendothelial system by enzymatic action. RES cells are thus an important part of the defense system of the body. When the substances are chemically inert, as are carbon particles, they are stored in the cell body and are in this way removed from the circulation or other extracellular spaces.

Names of the cells

Reticuloendothelial cells have a variety of names. Histiocytes are stationary cells located in loose connective tissue; together with leukocytes, or white blood cells, they are an important factor in cellular defense. The cells in the liver that belong to the reticuloendothelial system are called von Kupffer cells or stellate cells of von Kupffer. Those in nervous tissue are called microglial cells, or microglia, while those in the air spaces of the lung are called alveolar macrophages, or dust cells (a name derived from their ability to ingest inhaled dust particles). As part of the framework, or reticulum, of lymph nodes, bone marrow, and spleen, they may be called reticulum cells.

The cellular process by which RES cells take in substances has been termed phagocytosis (eating by cells) or pinocytosis (drinking by cells). Since a sharp distinction cannot always be made, and since fundamentally the same cellular activity may be involved, the term endocytosis is often used to encompass both processes. The designation endocytosis, however, when interpreted in its broadest sense, encompasses cellular internalizations not strictly belonging to the functions of the reticuloendothelial system.

Elie Metchnikoff, a Russian zoologist, first called attention to the importance of phagocytic activity. He called the white blood cells microphages ("small -eaters^w) because of their relatively small size, and he called larger cells that are also phagocytic macrophages ("large eaters").

The pathologists K.A.L. Aschoff and K. Kiyono found that when colloidal dyes are injected into the bloodstream they are taken up by certain types of cells. The cells that take up the greatest amount of dye are the von Kupffer cells; the special endothelial lining cells of the sinuses (channels) of the lymph nodes, spleen, bone marrow, and liver; the cells forming a reticulum in the stroma, or pulp, of these organs; and widely scattered wandering cells (also called histiocytes, though they are not stationary) in the connective tissues. The similarity of phagocytic activity of these cells suggested that they be grouped together as a system, named the reticuloendothelial system by Aschoff and Kiyono. This term, while still the most popular, is somewhat misleading because it appears to include the nonphagocytic endothelia of somatic vessels, as well as the phagocytic endothelium of the liver, bone marrow, spleen, and lymph nodes. From time to time suggestions have been made to return to the earlier term "macrophage system," while the name "mononuclear phagocyte system" has also been proposed. The latter term is, in part, based on the fact that macrophages have common precursor cells that are located in the bone marrow and that are deposited through intravascular transport at various places in the body and then transformed into macrophages.

While typical macrophages and the endothelial cells lining the sinuses of the liver, bone marrow, spleen, and lymph nodes can take up particulate substances, they do so in different ways. Macrophages engulf by extending processes from their cell bodies around the particulates and are capable of ingesting relatively large particles, such as red blood cells, or large aggregates of small particles. The sinus endothelial cells of the liver and bone marrow ingest particles by small indentations ("coated pits") at their surfaces that then pinch off and are found in the cytoplasm as small vesicles. These latter cells are not able to interiorize larger objects or large aggregates of particles.

While the modes of endocytosis of these two cell types are different, the eventual intracellular disposition of the ingested material is, in principle, the same in both cases. The endocytosed material is sequestered in the phagocytic cell in membrane-enclosed vesicles, termed phagosomes. These phagosomes fuse with other vesicles, called lysosomes, that contain hydrolytic enzymes that break down digestible material.

Control of the RES cells

Immune phagocytosis by macrophages may be mediated by the association of extracellular material with two types of receptors on the plasma membrane. The first, the F$_c$ receptor, binds the F$_c$ portion of IgG antibodies (immunoglobulin), with which bacteria may be coated, and thereby causes these antibodies and the bacteria to be adhesive to the phagocytic cell. The second receptor binds a complement cleavage product (C3b), which is bound by antigen-antibody complexes attached to bacteria. In addition to phagocytosis by these immunologically modified materials, macrophages are capable of ingesting materials independently of such factors in a manner that is not yet known.

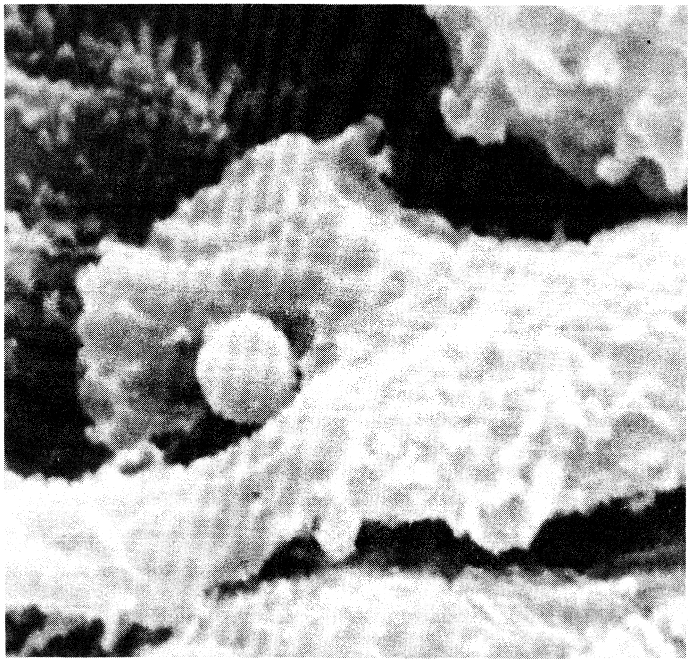
In the present article primary attention is paid to the functions of the reticuloendothelial cells and to the places where they are found in greatest numbers. The microglia are not discussed; they receive attention in the article NERVOUS SYSTEM, HUMAN.

ORGANS AND TISSUES OF THE SYSTEM

General morphology. In the spleen and bone marrow are thin-walled venous blood channels called sinuses or sinusoids; in the lymph nodes, lymphatic sinuses are interposed between the incoming and outgoing lymph streams. The cells forming the walls of both types of sinuses are elongated and resemble endothelial cells. The tissue around the sinuses is composed of spindle-shaped or branching cells that form a network, or reticulum, of reticulum cells and supporting reticulum fibres. The reticulum cells are actively phagocytic. The spaces of the reticulum contain many free, rounded cells, including large mononuclear (single-nucleus) reticulum cells that have become free and mobile and that have retained their phagocytic capacity. These are the macrophages described by Metchnikoff. The endothelial cells lining the sinuses are also phagocytic, a property not shared by the endothelial cells lining blood and lymph vessels elsewhere.

The walls of the sinuses of the liver have endothelial cells similar in function to those of the bone marrow and splenic sinuses. In addition, special macrophages that have a stellate shape occupy the sinus passages to a variable extent. These are the so-called von Kupffer cells.

There are also macrophages in the connective tissue, particularly in the loose connective tissue, throughout the body. They are called histiocytes or, less frequently, clasmatocytes or histoid wandering cells.



A macrophage engulfing a foreign particle.
By courtesy of the National Heart, Lung and Blood Institute

Spleen. The spleen is an important phagocytic organ; its phagocytic capacity per unit weight exceeds that of other organs or tissues of the reticuloendothelial system. The small arteries entering the substance of the spleen are for some distance surrounded by masses of lymphocytes—called splenic lymph follicles or Malpighian corpuscles—that form the so-called white pulp of the spleen. Interspersed between these lymphocytes are reticulum cells that are phagocytic. Within the tissue surrounding the white pulp—called red pulp because it contains a large amount of blood—are thin-walled venous sinuses that contain the cellular elements of peripheral blood and free macrophages. Around the sinuses is a network of reticulum cells and reticulum fibres; in the meshes of this network are blood cells and free macrophages. The lining of the sinuses differs from the endothelium of ordinary blood vessels elsewhere in that it is phagocytic.

Endocytosis in the spleen

Particulate material present in the circulation enters the splenic tissue proper through small capillaries. The material is endocytosed by the reticulum cells in white and red pulp, by the free macrophages outside and inside the venous sinuses, and by the endothelial lining cells of the venous sinuses. In this way the spleen functions as a reticuloendothelial organ inserted into the bloodstream. Sometimes red corpuscles that are old, damaged, or abnormal are segregated in the pulp spaces. The circulation in the pulp is slow, and the contained cells are subjected to influences that cause their destruction. Macrophages ingest the liberated hemoglobin and metabolize it, forming hemosiderin, the presence of increased amounts of which may cause bronzing of the skin, a condition known as hemosiderosis.

(P.P.H. De B.)

There are many conditions in which destruction of blood results in anemia. In some of these the spleen is involved only to the extent that its reticulum cells participate in the removal of cell debris and in the metabolism of iron pigments, An example is malaria, in which the malarial parasite destroys erythrocytes (red cells), resulting in the deposition of blood-derived pigments in thc reticuloendothelial cells,

In other cases the role of the spleen is not as obvious. In congenital spherocytic anemia, also called congenital hemolytic jaundice, for example, some of the red cells are spherical instead of being biconcave disks, as is normal. It has been postulated that their spherical shape prevents them from passing through the narrow openings in the walls of the splenic sinuses and thus entering the venous circulation. These spherical cells are more fragile than normal ones and, as they remain in the pulp spaces in contact with reticulum cells, undergo hemolysis (liberation of hemoglobin) and are thus eliminated. The importance of the spleen in this condition is evident from the effect of its surgical removal (spienectomy), which usually ends the excess blood destruction of the disease, presumably because of the removal of a large source of reticuloendothelial cells. It is notable that if accessory spleens are present, they may enlarge after removal of the primary spleen and resume the excessive destruction of blood cells.

In hemolytic anemias, caused by the presence of antibodies on the red cells, the presence of the antibody so alters the cells that they are treated as objects foreign to the body and are segregated and disposed of in the splenic pulp. In this condition, splenectomy is often of little value.

There are other conditions in which the number of circulating blood elements is low and in which splenectomy may be beneficial. Idiopathic thrombocytopenic purpura is a condition associated with coagulation defects and tendencies toward bleeding in which there is an inadequate number of blood platelets. The number of platelets usually promptly increases upon removal of the spleen, a result suggesting an action similar to that in hemolytic anemia, especially since immune bodies have been found in the blood serum.

(M.N.R.)

Lymph nodes. The substance of the lymph nodes is composed of a network of phagocytic reticulum cells and supporting reticulum fibres. In the meshes of this network are lymphocytes and free macrophages. This reticulum is traversed by lymphatic sinuses that receive lymph from incoming lymphatic vessels and are drained by outgoing lymphatics. The lining of the lymphatic sinuses is continuous and consists of phagocytic endothelial cells. The lumen (interior passage) of a lymphatic sinus is spanned by trabeculae (bundles of fibres) comprised of reticular fibres and adhering reticulum cells to which macrophages are attached. Free macrophages may occur in the lumina of the lymphatic sinuses. The endothelial lymphatic sinus lining cells, reticulum cells of the trabeculae, free intraluminal macrophages, and macrophages adhering to the trabeculae are all phagocytic and form—together with the reticulum cells and free macrophages outside the sinuses—the reticuloendothelial components of the lymph nodes. The lymph nodes thus function as phagocytic organs inserted in the lymph stream. Bacteria, cell fragments, and pigments that originate within and outside of the body (*e.g.*, hemosiderin, melanin, coal dust) all may be found in these reticuloendothelial elements. (See also LYMPHATIC SYSTEM, HUMAN; LYMPHATIC SYSTEM DISEASES.)

Bone Marrow. In addition to having an important blood-forming function, the bone marrow plays a significant role in the function of the reticuloendothelial system.

Structure of the bone marrow

Bone marrow consists of a network of phagocytic reticulum cells and some supporting reticulum fibres. In the meshes of this reticulum are immature and mature red and white blood cells, blood-platelet-forming cells (megakaryocytes), fat cells, and free macrophages. The major blood vascular components in bone marrow are the venous sinuses, into which the mature blood cells are deposited and from which they reach the peripheral blood circulation. The walls of these sinuses are composed of a continuous lining of endothelial cells that remove particulate material from the bloodstream by capturing it in pitlike indentations (bristle-coated pits) in the endothelial cell surface. The pits pinch off from the cell surface and the phagocytosed material is thereby sequestered in small intracellular vesicles. These endothelial cells, as well as the extravascular reticulum cells and macrophages, are the essential components of the reticuloendothelial system in the bone marrow. Since phagocytosis by the endothelia! lining cells is through small pits and small vesicles, the particulate uptake of these cells is restricted to small particles and small aggregates of such particles. The macrophages that engulf by pseudopodial extensions are capable of the uptake of large particles or larger aggregates of small particles. (For the blood-forming functions of the bone marrow, see BLOOD, HUMAN.)

Liver. The liver, while per unit weight less phagocytic than the spleen, is the most phagocytic organ of the human body because of its size. Its phagocytic activity is intercalated in the bloodstream. The substance of the liver is traversed by venous sinuses that—unlike the sinuses of the spleen, lymph nodes, and bone marrow—have no network of reticulum cells or fibres and no free macrophages around them. The sinus endothelium is separated from the liver cells proper by a narrow space, called the space of Disse. These endothelial lining cells have small open apertures that directly link the lumina of the venous sinuses with the space of Disse.

In addition to the endothelial lining cells, there are in the walls of the sinuses typically large macrophages, the cells of von Kupffer. Their cell bodies may protrude into the sinus lumina and carry pseudopodial extensions that may reach the opposite sinus wall. The endothelial cells and the von Kupffer cells are the essential components of the reticuloendothelial system of the liver. The endothelial cells phagocytose by means of small bristle-coated pits at the cell surfaces, while the von Kupffer cells can engulf large particles through their pseudopodial extensions.

(P.P.H. De B.)

As in other parts of the reticuloendothelial system, several kinds of pigment may be found in the von Kupffer cells, the most common of which is hemosiderin. Although only small amounts are found normally, increased blood destruction results in extensive hemosiderosis. Hemosiderin may be found in both liver and von Kupffer cells, but this apparently common site probably represents the passage of hemosiderin to the liver cells by von Kupffer cells in which it is formed.

The malarial parasite forms a pigment, hematin, from hemoglobin that resembles hemosiderin in appearance but does not give chemical reactions for iron as hemosiderin does. Hematin may be found in von Kupffer cells. Occasionally malarial parasites are also present, but only in small numbers. (See also LIVER, HUMAN.)

(M.N.R.)

FUNCTIONS OF THE SYSTEM

Metabolism of iron

The reticuloendothelial system in metabolism. The metabolism of iron starts with its absorption in the intestinal tract and continues through its utilization in the formation of hemoglobin and its final degradation. The part played by the reticuloendothelial system in this sequence has been mentioned briefly in connection with the presence of hemosiderin in phagocytes. This process is the visible and therefore the most obvious functional activity in this connection. The role of reticulum cells in hemoglobin synthesis is not readily observed. It has been shown that in the sites of early red-cell formation, nucleated precursors of red cells, or normoblasts, may be seen in small groups with a reticulum cell in the centre. It is here that reticulum cells function directly in hemoglobin synthesis by transferring iron-containing material in the form of ferritin, an amber-coloured protein, directly to the surrounding normoblasts.

When hemosiderin is formed after disintegration of red cells, the reticuloendothelial system has the additional function of storage. The hemosiderin contained in the reticulum cells is available for further use in the formation of more hemoglobin. There is, thus, an important difference between the loss of hemoglobin and its contained iron from the body by bleeding and the loss of hemoglobin, with retention of iron, in hemolytic disorders. The implications for therapy in the latter case are obvious, for storage makes the administration of additional iron unnecessary.

The reticuloendothelial system also plays a role in the formation of bile. It has been shown experimentally in animals that bile formation may occur in the absence of a liver. In human beings this function can be observed in the vicinity of hemorrhages. Non-iron-containing fragments of hemoglobin, as well as hemosiderin, may be found in local phagocytes. Hematoidin, which is one of these fragments, has been identified as bilirubin.

The metabolism of lipids (fats) is also an important function of the reticuloendothelial system as is evidenced by its prominent involvement in disturbances of lipid metabolism. The foam cells characteristic of Niemann-Pick disease, for example, are reticuloendothelial cells, as are the Gaucher's cells of Gaucher's disease.

The reticuloendothelial system in defense. The reticuloendothelial system plays an important part in the body's defense against disease in two ways: first, by the phagocytic activity of its cells; second, by the apparent necessity of the presence of macrophages for the effective interactions and activities of immunologically competent cells, such as stimulated lymphocytes.

Phagocytosis of microorganisms has already been described. The process may also be directed to insoluble foreign bodies; examples of the activity are the reactions to splinter fragments or to surgical sutures. If a fragment is small enough, it may be ingested by a single cell. If the object is too large, several macrophages collect around it, and their cell bodies coalesce to form a single giant cell. These large cells may contain from a few to two or three dozen nuclei, indicating the number of macrophages involved in their formation. The foreign substance may often be identified in the cytoplasm (the cell substance outside the nuclei).

In chronic reactions (*e.g.*, granulomas) and in the healing stages of acute inflammation, the cellular response is predominantly by mononuclear cells from the connective tissue mingled with cells from the blood. The result is the production of a tissue referred to as granulation tissue. The process is described in the article INFLAMMATION. Here only the part played by the reticuloendothelial system is considered.

(M.N.R./P.P.H. De B.)

As a response to a local noxious stimulus, such as the entry of bacteria into the connective tissue, white blood cells leave the circulation through capillaries and small veins and arrive at the site where the stimulus occurs. The granulocytes immediately phagocytose bacteria or other particulate material, and they can destroy digestible foreign material by enzymes present in their granules (lysosome~).Nongranulocytic cells (monocytes) also leave the circulation and, once arrived in the extravascular connective tissue, transform into macrophages. In addition to these macrophages derived from the peripheral blood (hematogenous macrophages), the histiocytes present in the connective tissue become actively phagocytic and thus form a pool of histogenous macrophages.

(P.P.H. De B.)

The second method by which the reticuloendothelial system participates in body defense is through immune reactions. (The complex mechanism of this participation is discussed in detail in the article IMMUNITY.) In brief, one class of lymphocytes (B-cells, believed to be derived from bone marrow) can synthesize and secrete antibody molecules (immunoglobulin) with the help of another class of lymphocytes (T-cells, derived from the thymus.) The T-cells are also capable of other immunological reactions not involving antibody production. In these immunological responses of lymphocytes, the presence of macrophages appears to be a required factor, although their precise function is not entirely clear.

METHODS FOR STUDY OF RETICULOENDOTHELIAL FUNCTION

Tests of phagocytosis

The classical method of identifying the cells of the reticuloendothelial system is the injection of colloidal dyes or carbon particles. Quantitative measurements of functional activity can be obtained. The general principle is to utilize the phagocytic capacity of the cells by determining the rate at which substances are removed from the circulation after injection into a vein.

Besides the dyes and carbon particles, the substances used in animal experiments have included materials such as radioactive chromium phosphate and thorium dioxide (Thorotrast) and other materials. Quantitative studies of reticuloendothelial function have been based on the theory that functioning of the cells can be blocked by saturating them with such harmless substances as lipids, albumin, and gelatin. The assumption is that cells already filled cannot ingest additional material. This "blockade" of the reticuloendothelial system is only partially successful. Cells filled with one substance may still take up another; the presence of ingested material may influence the rate of phagocytosis, phagocytized substances may be disposed of, or new phagocytes may be mobilized.

Nevertheless, a number of factors are known to influence the rate of phagocytosis, such as the chemical nature of the material, the size and number of the injected particles, and the state of the cells themselves. The species and condition of the animal used in the experiment affect the results because there are differences in the amount of reticuloendothelial tissue available, especially in the spleen and liver.

The metabolic activity of the cells is less readily measured. Inert substances remain within the tissues indefinitely, but others undergo metabolic changes within the cells. Serum albumin labelled with radioactive iodine is an example. The iodine portion is released by the reticuloendothelial cells and eliminated, so that reduction of radioactivity in an organ is an indication of the rate of metabolism.

NEOPLASMS

Tumours of the reticuloendothelial system are common. They may be relatively localized or widespread, and are nearly always malignant.

Reticulum-cell sarcoma

Reticulum-cell sarcoma is a malignant tumour composed of reticulum cells. Theoretically it may arise wherever reticulum cells occur, but the most common primary site is in lymph nodes. When tumours of lymph nodes are grouped together as lymphomas, reticulum-cell sarcoma is referred to as malignant lymphoma of the reticulum cell (or histiocytic) type. The tumour may metastasize (spread) either by blood or by lymph and thus appear in other locations in the body.

The different appearances of reticulum cells are reflected in differences among the tumours derived from them. The sarcoma cells may be round and little larger than lymphocytes, thus suggesting lymphosarcoma; the formation of tumour giant cells may give the appearance of Hodgkin's disease, the cells of which are also of reticuloendothelial origin.

Although the tumours metastasize readily, their cells are not ordinarily found on blood examination. Occasionally, however, cases of leukemia occur in which the leukemic cells in the blood have been identified as circulating reticulum cells. As noted before, there has been a controversy over the origin of the neoplastic cells in monocytic leukemia.

It is notable that the cells of reticulum-cell sarcoma seldom contain ingested material. There is, however, a condition that can readily be confused with reticulum-cell sarcoma or with Hodgkin's disease, and in which phagocytosis of red cells is a conspicuous feature. The condition, called histiocytic medullary reticulosis, is one of a group of diffuse proliferations of cells described under the general terms reticuloendotheliosis or reticulosis.

BIBLIOGRAPHY. For a historical review by the author of the concept of the reticuloendothelial system see L. ASCHOFF, *Reticuloendothelial System: Lectures in Pathology* (1924). Experimental studies are reported in the *Journal of the Reticuloendothelial Society* (1963–). Additional information on this subject is given in W. BLOOM and D.W. FAWCETT, *A Textbook of Histology,* 9th ed. (1968); R.H. JAFFE, "The Reticuloendothelial System," in H. DOWNEY (ed.), *Handbook of Hematology,* vol. 2 (1965); M.N. RICHTER, "Blood and Bone Marrow" and "Lymph Nodes, Spleen and Reticuloendothelial System," in W.A.D. ANDERSON (ed.), *Pathology,* 5th ed. (1966). A variety of aspects of the reticuloendothelial system are presented in R. VAN FURTH (ed.), *Mononuclear Phagocytes* (1970). and *Mononuclear Phagocytes in Immunity, Infection, and Pathology* (1975). The most comprehensive treatment of the reticuloendothelial system is found in a series of books by H. FRIEDMAN, M. ESCOBAR, and S.M. REICHARD (eds.), *The Reticuloendothelial System* (1980–).

(M.N.R./P.P.H. De B.)

Revelation

Revelation is a religious term that designates the disclosure of divine or sacred reality or purpose to human beings. In the religious view, such disclosure may come through mystical insights, historical events, or spiritual experiences that transform the lives of individuals and groups.

NATURE AND SIGNIFICANCE

Every great religion acknowledges revelation in the wide sense that its followers are dependent on the privileged insights of its founder or of the original group or individuals with which the faith began. These profound insights into the ultimate meaning of life and the universe, which have been handed down in religious traditions, are arrived at, it is believed, not so much through logical inference as through sudden, unexpected illuminations that invade and transform the human spirit. Those religions that look upon God as a free and personal spirit distinct from the world accept revelation in the more specific sense of a divine self-disclosure, which is commonly depicted on the model of human intersubjective relationships. In the "prophetic" religions (Judaism, Christianity, Islām, and Zoroastrianism), revelation is conceived as a message communicated by God to an accredited spokesman, who is charged to herald the content of that message to an entire people. Revelations received on behalf of the whole community of the faithful are often called "public" (as opposed to "private" revelations, which are given for the guidance or edification of the recipient himself).

The media of revelation

The ways by which revelation occurs are variously conceived. Most religions refer to signs, such as auditory phenomena, subjective visions, dreams, and ecstasies. In so-called primitive religions, revelation is often associated with magical techniques of divination. In the prophetic religions, revelation is primarily understood as the "Word of God," enabling the prophet to speak with certainty about God's actions and intentions. In mystical religions (*e.g.*, Islamic Ṣūfism, Tantric Buddhism), revelation is viewed as an ineffable experience of the transcendent or the divine.

TYPES AND VARIATIONS

Primitive religions. In so-called primitive cultures, revelation is frequently identified with the experience of supernatural power (mana) in connection with particular physical objects, such as stones, amulets, bones of the dead, and unusual animals. The sacred or holy is likewise believed to be present in sacred trees, groves, shrines, and the like and in elemental realities such as earth, water, sky, and the heavenly bodies. Once specified as holy, such objects take on symbolic value and become capable of mediating numinous (spiritual) experiences to the adherents of a cult. Certain charismatic individuals, such as shamans, who are believed to be in communion with the sacred or holy, perform functions akin to those of the prophet and the mystic in more developed religions.

Religions of the East. Eastern religions are concerned with man's struggle to understand and cope with the predicament of his existence in the world and to achieve emancipation, enlightenment, and unity with the Absolute. Western religions, on the other hand, lay more stress on man's obedient response to the sovereign Word of God. The notion of revelation in the specific sense of a divine self-communication is more apparent in Western than in Eastern religions.

Hinduism. In Hinduism, the dominant religion of India, revelation is generally viewed as a process whereby the religious seeker, actuating his deeper spiritual powers, escapes from the world of change and illusion and comes into contact with ultimate reality. The sacred books are held to embody revelation insofar as they reflect the eternal and necessary order of things.

A major form of Hindu thought, Vedānta, includes two main tendencies: the monistic (advaitn) and the theistic (*bhakti*). The leading sage of Advaita Vedanta, Śaṅkara (early 9th century), while acknowledging in principle the possibility of coming to a knowledge of the Supreme Reality (Brahman) through inner experience and contemplation of the grades of being, held that in practice a vivid apprehension of the divine arises from meditation on the sacred books, especially the *Upaniṣads*. In Bhakti, systematized by the philosopher Rāmānuja (*c.* 1050–1137), the Absolute is regarded as personal and compassionate. Revelation, consequently, is viewed as the gracious self-manifestation of the divine to those who open themselves in loving contemplation. The devotional theism of Bhakti, very influential in modern India, resembles the pietism and mysticism of the Western religions.

Buddhism. Buddhism, the other great religion originating on Indian soil, conceives of revelation not as a personal intervention of the Absolute into the worldly realm of relativities but as an enlightenment gained through discipline and meditation. Gautama the Buddha (6th to 5th century BC), after a striking experience of human transitoriness and a period of ascetical contemplation, received an illumination that enabled him to become the supreme teacher for all his followers. Although Buddhists do not speak of supernatural revelation, they regard the Buddha as a uniquely eminent discoverer of liberating truth. Some venerate him, some worship him, and all Buddhists seek to imitate him as the most perfect embodiment of ideal manhood — an ideal that he in some way "reveals."

Role of the Buddha

Chinese religions. Chinese wisdom, more world-affirming than the ascetical religions of India, accords little or no place to revelation as this term is understood in the Western religions, though Chinese traditions do speak of the necessity of following a natural harmony in the universe. Taoism, perhaps the most characteristic Chinese form of practical mysticism, finds revelation only in the transparency of the immanent divine principle or way (Tao). Confucianism, while not incompatible with Taoism, is oriented less toward natural mysticism and more toward social ethics and decorum, though it too is concerned with accommodating life to a balance in the natural flow of existence. Confucius (551–479 BC), who refined the best moral teachings that had come down in the tradition, was neither a prophet appealing to divine revelation nor a philosopher seeking to give reasons for his doctrine.

Religions of the West. In the three great religions of the West — Judaism, Christianity, and Islām—revelation is the basic category of religious knowledge. Man knows God and his will because God has freely revealed himself —his qualities, purpose, or instructions.

Judaism. The Israelite faith looked back to the Pentateuch (the first five books of the Old Testament) for its fundamental revelation of God. God was believed to have revealed himself to the patriarchs and prophets by various means not unlike those known to the primitive religions —theophanies (visible manifestations of the divine), dreams, visions, auditions, and ecstasies — and also, more significantly, by his mighty deeds, such as his bringing the Israelites out of Egypt and enabling them to conquer the Holy Land. Moses and the prophets were viewed as the chosen spokesmen who interpreted God's will and purposes to the nation. Their inspired words were to be accepted in loving obedience as the Word of God.

Revelation through events

Rabbinic Judaism, which probably originated during the Babylonian Exile and became organized after the destruction of the Temple by the Romans, concerned itself primarily with the solution of legal and ethical problems. It gradually developed an elaborate system of casuistry resting upon the Torah (the Law, or the Pentateuch) and its approved commentaries, especially the Talmud (commentaries on the Torah), which was regarded by many as equal to the Bible in authority. Orthodox Judaism still recognizes these authoritative sources and insists on the verbal inspiration of the Bible, or at least of the Pentateuch.

Christianity. The New Testament took its basic notions of revelation from the contemporary forms of Judaism (1st century BC and 1st century AD)—*i.e.*, from both normative rabbinic Judaism and the esoteric doctrines current in Jewish apocalyptic circles in the Hellenistic world. Accepting the Hebrew Scriptures as preparatory revelation, Christianity maintains that revelation is brought to its unsurpassable climax in the person of Jesus Christ, who is God's own Son (Heb. 1:1–2), his eternal Word (John 1:1), and the perfect image of the Father (Col. 1:15). The Christian revelation is viewed as occurring primarily in the life, teaching, death, and Resurrection of Jesus, all interpreted by the apostolic witnesses under the illumination of the Holy Spirit. Commissioned by Jesus and empowered by the divine spirit, the apostles, as the primary heralds, hold a position in Christianity analogous to that of the prophets in ancient Israel.

The Apostle Paul, though not personally a witness to the public life of Jesus, is ranked with the Apostles by reason of his special vision of the risen Christ and of his special call to carry the Gospel to the Gentiles. In his letters, Paul emphasized the indispensability of missionary preaching in order that God's revelation in Christ be communicated to all the nations of the world (Rom. 10:11–21).

Christianity has traditionally viewed God's revelation as being complete in Jesus Christ, or at least in the lifetime of the Apostles. Further development is understood to be a deeper penetration of what was already revealed, in some sense, in the 1st century. Periodically, in the course of Christian history, there have been sectarian movements that have attributed binding force to new revelations occurring in the community, such as the 2nd-century Montanists (a heretical group that believed they were of the Age of the Holy Spirit), the 13th century Joachimites (a mystical group that held a similar view), the 16th-century Anabaptists (radical Protestant sects), and the 17th-century Quakers. In the 19th century the Church of Jesus Christ of Latter–day Saints (popularly known as Mormons) recognized, alongside the Bible, additional canonical scriptures (notably, the Book of Mormon) containing revelations made to the founder, Joseph Smith.

Islām. Islām, the third great prophetic religion of the West, has its basis in revelations received by Muhammad (*c.* 7th century AD). These were collected shortly after his death into the Qur'ān (Koran), which is regarded by Muslims as the final, perfect revelation — a human copy of the eternal book, dictated to the Prophet. While Islām accords prophetic status to Moses and Jesus, it looks upon the Qur'ān as a correction and completion of all that went before. More than either Judaism or Christianity, Islrim is a religion of the Book. Revelation is understood to be a declaration of God's will rather than his personal self-disclosure. Insisting as it does on the absolute sovereignty of God, on man's passivity in relation to the divine, and on the infinite distance between creator and creature, Islām has sometimes been inhospitable to philosophical speculation and mystical experience. Yet in medieval Islām there was both a remarkable flowering of Arabic philosophy and the intense piety of the mystical Siifis. The rational-

Revelation through a sacred book

ism of some philosophers and the theosophical tendencies of some of the Ṣūfīs came into conflict with official orthodoxy.

Zoroastrianism. A fourth great prophetic religion, which should be mentioned for its historic importance, is Zoroastrianism, once the national faith of the Persian Empire. Zoroaster (Zarathushtra), a prophetic reformer of *c.* 7th century BC, apparently professed a monotheistic faith and a stern devotion to truth and righteousness. At the age of 30 he experienced a revelation from Ahura Mazdā (The Wise) and chose to follow him in the battle against the forces of evil. This revelation enabled Zoroaster and his followers to comprehend the difference between good (Truth) and evil (The Lie) and to know the one true God. Later forms of Zoroastrianism apparently had an impact on Judaism, from the time of the Babylonian Exile, and, through Judaism, on Christianity.

THEMES AND FUNCTIONS

Recurrent questions concerning revelation include the relationship between general and special revelation; the relationship between word and deed as media of special revelation; the authority of the sacred books; the revelatory value of tradition; the nonverbal component in revelation; the interpersonal dimension of revelation; and the relationship between faith and reason.

General revelation: the role of nature. The Eastern religions, on the whole, differ from Western religions in that they place less emphasis on a special or exclusive revelation received by a "chosen people" and rather speak of the manifestation of the Absolute through the general order of nature. There is, however, no irreconcilable opposition between general and special revelation. Vedanta Hinduism and Buddhism, even if they do not speak of special revelation, believe that their religious books and traditions have unique value for imparting a saving knowledge of the truth. The Bible and the Qurʾān, conversely, proclaim that although God has specially manifested himself to the biblical peoples, he also makes himself known through the order of nature. The failure of some nations to acknowledge the one true God is attributed not to God's failure to disclose himself but rather to the debilitating effects of sin on the perceptive powers of man.

Special revelation: the role of history. The Western religions differ somewhat among themselves in the ways in which they understand how special revelation occurs. Some focus simply on the direct inspiration of the divinely chosen prophets. The Judeo-Christian tradition, however, characteristically looks upon the prophets as witnesses and interpreters of what God is doing in history. Revelation through deeds is conceived to be more fundamental than revelation through words, though the words of the prophets are regarded as necessary to clarify the meaning of the events. Since the Old Testament term for "word" (davar) signifies also "deed" or "thing," there is no clear line of demarcation between word-revelation and deed-revelation in the Bible. The biblical authors look upon the national fortunes of Israel as revelations of God's merciful love, his fidelity to his promises, his unfailing power, his exacting justice, and his readiness to forgive the penitent sinner. The full disclosure of the meaning of history, for many of the biblical writers, will occur only at the end of time, when revelation will be given to all peoples in full clarity. The Judeo-Christian notion of history as progressive revelation has given rise to a variety of theological interpretations of world history, from St. Augustine (AD 354–430) to G.W.F. Hegel (1770–1831) and other modern thinkers.

The importance of canons of scriptures

Revelation and sacred scriptures. In those religions that look for guidance to the ancient past, great importance is attached to sacred books. Theravāda Buddhism, while it professes no doctrine of inspiration, has drawn up a strict canon (standard or authoritative scriptures)—the "Pāli canon'—in order to keep alive what is believed to be the most original and reliable traditions concerning the Buddha. Mahāyāna Buddhism, while it has no such strict canon, considers that all its adherents must accept the authority of the *sūtras* (basic teachings written in aphorisms). Zen Buddhism, in many ways the broadest development of Mahāyāna thought, sometimes goes to the point of rejecting any such written authority. Many religions view their holy books as inspired and inerrant. According to a very ancient Hindu tradition, the sages of old composed the Vedas by means of an impersonal type of inspiration through cosmic vibrations. Judaism, on the other hand, looks upon the Bible as divinely inspired. The idea of verbal dictation from God, which occurs here and there in the Bible, was applied by some rabbis to the Pentateuch, which was believed to have been written by Moses under verbal inspiration, and even to the whole Bible. Christianity, which generally accepts both the Old and New Testaments as in some sense inspired, has at times countenanced theories of verbal dictation. According to the Mormons, the Book of Mormon was composed in heaven and delivered on tablets of gold to Joseph Smith. Islām holds that the Qurʾān, an eternal heavenly book, was dictated verbatim to Muhammad. The Prophet's companions testified that he would often turn red or livid, sweat profusely, and fall into trances while receiving revelations.

Apostolic tradition

Revelation and tradition. The great religions frequently make a distinction between those scriptures that contain the initial revelation and others, at the outer fringe of the canon, that contain authoritative commentaries. In Hinduism, the four Vedas and three other ancient collections—the *Brāhmaṇas,* Aranyakas, and *Upaniṣads*—are *Śruti* ("that which has been heard"; *i.e.*, constitutive revelation); the other sacred writings (the *sūtras*, the lawbooks, *Purāṇas*, and the Bhagavadgītā and the *Rāmā*yana, the two great epics) are *Smṛti* ("that which has been remembered"; *i.e.,* tradition). Later Judaism, while recognizing the unique place of the Bible as the written source of revelation, accords equal authority to the Talmud as traditional commentary. Among Christians, Roman Catholics and the Eastern Orthodox believe that revelation is to be found not only in the Bible but also, by equal right, in the apostolic tradition. Protestants emphasize the objective sufficiency of Scripture as a source of revelation, but many Protestants today are careful to add that Scripture must always be read in the light of church tradition in order that its true message be rightly understood. Islām holds that the Qurʾān alone contains revelation in the strict sense (wahy), but it accepts tradition (Hadith) as a supplementary source of Islāmic law. Special significance is attached to the practice (sunnah) of the Prophet himself and to the traditions handed down by his immediate companions.

Revelation and experience. In most religions nonverbal communication plays an important part in the transmission of revelation. This can occur in art (notably in icons, statues, and idols), in sacred music, in the liturgy, and in popular dramas, such as the mystery plays common in medieval Europe or those still performed in Indian villages. For a deeper initiation into the revelation, it is believed necessary to live under the tutelage of a guru (teacher), monk, or holy man. To the extent that revelation is identified with a profound and transforming personal experience, the spiritual preparation of the subject by prayer and asceticism is stressed. Among the great living religions of the world, there is wide agreement that revelation cannot be fully communicated by books and sermons but only by an ineffable, suprarational experience. In Hinduism the *Upaniṣads* emphasize the hiddenness of God. Leaving behind all created analogies, the adept is led to the point where he comes to praise God in an adoring silence more exalted than speech. Buddhism of the Mahāyāna, especially its Zen varieties, likewise advocates ecstatic contemplation. The Eastern mystics are here in close agreement with the Jewish Hasidim (mystical pietists), with the Islāmic Ṣūfīs, and with the great Christian mystics, such as Pseudo-Dionysius, the Areopagite, Meister Eckehart, and St. John of the Cross. Many theologians within Judaism (*e.g.*, Maimonides) and Eastern Christianity (*e.g.*, St. John Chrysostom, St. John of Damascus) have contended that God is best known through a negative, or "apophatic," theology that makes no positive statements about God. This idea, never absent from the medieval scholastic (intellectualist) tradition, was newly emphasized by Martin Luther, who insisted that the revealed God ***(Deus revelatus)*** remains the hidden God ***(Deus absconditus)***, before whom man must stand in reverent awe. Contemporary Roman Catholic theologians, such as Karl Rahner, maintain that even in heaven God will not cease to be, for man's finite mind, an unfathomable mystery. Revelation makes man constantly more aware of the depths of the divine incomprehensibility.

The revealed God as the hidden God

Revelatory relationships. In certain forms of mysticism, particularly prevalent in the Eastern religions, the envisioned goal is an absorption into the divine, involving the loss of individual consciousness. In the Western religions and in Bhakti Hinduism the abiding distinctness of the individual personality is affirmed. Islāmic orthodoxy, looking upon revelation as a declaration of the divine will, stresses not so much the communion of man with God as rather man's obedient submission to the creator. Islāmic Ṣūfīsm, however, resembles Hasidic Judaism and Christianity in its aspiration for personal union with God. For many contemporary religious thinkers, such as the Jewish philosopher Martin Buber and the Roman Catholic philosopher Gabriel Marcel, revelation involves a mutual self-giving of the revealer and the believer in personal intercommunion. According to Karl Rahner, revelation consists primarily and essentially in God's gracious communication of his own divine life to man as a personal spirit. In his view, the articulation of revelation in the scriptures and creeds is a secondary stage, presupposing an experiential encounter with the divine. This secondary phase, however, is viewed as necessary in order that man may realize himself in his humanity as a believer and achieve solidarity with his fellow believers. In general, the Western religions tend to attach more importance to the idea of a community of faith than do the Eastern religions. Revelation in the biblical and Islāmic view is addressed not to individuals as such but to a whole people, which achieves its identity, in part, by articulating its faith in writings that are approved as authentic expressions of what God has revealed.

Revelation and reason. The problem of the relationship between revelation and reason arises, on the one hand, because revelation transcends the categories of ordinary rational thought and, on the other hand, because revelation is commonly transmitted by means of authoritative records, the contents of which cannot be verified by the believer. Buddhism, since it does not attribute inspiration or inerrancy to its canonical sources, allows some scope for individual reason to criticize the authoritative writings, but, like other religions, it has to face the charge that the illumination to which it aspires may be illusory. Orthodox Hindus, giving full authority to the Veda, hold that human reason errs whenever, on the grounds of perceptual experience, it takes issue with the sacred writings. Hinduism, however, allows for great freedom in the exegesis (interpretation) of its sacred books, some of which are more poetic than doctrinal.

Tension between faith and reason

The tension between faith and reason has been particularly acute in the Western religions, which find revelation not simply in holy books but in prophetic words that call for definite assent and frequently command a precise course of action. The ambiguities of scripture in these religions are frequently cleared up by creeds and dogmas of the community, calling for the assent of true believers. Judaism, Christianity, and Islām, moreover, came into close contact with Hellenistic culture, which held up the ideal of rationally certified knowledge as the basis for the good life. They, therefore, had to face the problem: could assent to an authoritative revelation be justified before the bar of reason? Some theologians took a "fideist" (faith-based) position, maintaining that reason must in all things submit to the demands of revelation. Others, such as the Arabic philosopher Averroës and his followers (both Muslim and Christian), accepted the primacy of reason. They reinterpreted the content of revelation so as to bring it into line with science and philosophy. A third school, in which the medieval Jewish philosopher Maimonides and the medieval Christian scholastic theologian Thomas Aquinas may be included, sought to maintain the primacy of faith without sacrificing the dignity of reason. According to the Thomist theory, human reason can discern the credibility of revelation because of the external signs by which God has authenticated it (especially prophecies and miracles). Reason, moreover, makes it possible for the believer to understand, in some measure, the revealed mysteries. This intellectualist position continues to appeal to many Christians; but some maintain that it overlooks the qualitative differences between faith—as a transrational assent to mystery—and scientific knowledge, which operates within the categories of objectivizing reason.

CONCLUSION

In some theological circles the concept of revelation is rejected on the ground that it is bound up with mythological and anthropomorphic conceptions and introduces an unassimilable element into the history of religions. It would seem, however, that the concept can be purified of these mythical elements and still be usefully employed. In the sphere of religion, wisdom is often best sought through privileged moments of ecstatic experience and through the testimony of those who have perceived the sacred or holy with unusual purity and power. The self-disclosure of the divine through extraordinary experiences and symbols is fittingly called revelation. Because of the pervasiveness of the idea of revelation in the world's religions and because the various religions have had to cope with similar theological problems concerning revealed knowledge, revelation has become a primary theme for dialogue among the great religions of mankind.

Perception of the sacred or holy

BIBLIOGRAPHY. R.C. ZAEHNER, ***At Sundry Times*** (1958), a sympathetic approach by an accomplished scholar who finds anticipations of Christian revelation not only in Judaism but also in Hinduism, Buddhism, and Zoroastrianism; N. SODERBLOM, ***The Living God: Basal Forms of Personal Religion*** (1933), a discussion of revelation as found prevalent in the "prophetic" faiths and most of all in the historical, incarnational faith of Christianity; J.H. WALGRAVE, ***Un salut aux dimensions du monde,*** trans. from the Dutch (1970), an apologetically oriented work that attempts to bring out the distinctive qualities of the Christian view of revelation in comparison with Buddhism, Hinduism, and Islām.

Primitive religion: MIRCEA ELIADE, ***Traite' d'histoire des religions*** (1948; Eng. trans., ***Patterns in Comparative Religion,*** (1958), a discussion of hierophanies, myths, and symbols as pertinent to the theme of revelation; G. VAN DER LEEUW, ***Phänomenologie der Religion*** (1933; Eng. trans., ***Religion in Essence and Manifestation,*** 2 vol., 1963), a phenomenological approach influenced by Rudolf Otto and others that deals with ***mana*** and related notions and holds that revelation eludes phenomenology and can be grasped only in faith; C.H. LONG, ***Alpha: The Myths of Creation*** (1963), a discussion of the value of myth as a form of expression of religious experience.

Christianity: ETIENNE GILSON, ***Reason and Revelation in the Middle Ages*** (1938), an historical survey by a contemporary Thomistic philosopher; A.R. DULLES, ***Revelation Theology: A History*** (1969), a brief historical survey of Catholic and Protestant views; J. BAILLIE, ***The Idea of Revelation in Recent Thought*** (1956), a sketch of trends in 20th-century Protestant theology; R. LATOURELLE, ***The'ologie de la révélation*** (1963; Eng. trans., 1966), a full historical and systematic study by a Roman Catholic; W. PANNENBERG (ed.), ***Oflenbarung als Geschichte*** (1965; Eng. trans., ***Revelation as History,*** 1968), a symposium by younger German Lutheran scholars.

Islām: A.J. ARBERRY, ***Revelation and Reason in Islam*** (1957), a concise and learned treatment of the medieval controversies; K. CRAGG, ***The Call of the Minaret,*** pt. 2, pp. 33–171 (1956), a very objective presentation of Muslim faith and piety, including some discussion of the doctrine of revelation.

Hinduism: K.S. MURTY, ***Revelation and Reason in Advaita*** *Vedānta* (1959), an exposition and evaluation of Śaṅkara's position in the light of modern Western philosophy; R.C. ZAEHNER, ***Hindu and Muslim Mysticism*** (1960), on the love relationship to God in Bhakti and Ṣūfīsm.

Buddhism: W.L. KING, ***Buddhism and Christianity: Some Bridges of Understanding*** (1962), an objective comparison between Christianity and Theravāda Buddhism; good discussion of the revelatory role of the Buddha.

Judaism: A.J. HESCHEL, ***God in Search of Man,*** pt. 2, pp. 167–278 (1956), presentation of modern Judaism by a prominent rabbinic scholar.

(A.Du.)

Revolution, Political

"Revolution," its connotations in astronomy apart, has in political and social thought come to mean a sudden, major, and hence typically violent alteration in government and in related associations and structures. It is used by analogy in such expressions as the industrial revolution, where it refers to a radical and profound change in economic relationships and technological conditions.

DEFINITIONS OF "REVOLUTION"

Though the idea was originally related to the notion of cyclical alterations in the forms of government, it now connotes a novel departure, as epitomized in Leon Trotsky's assertion that "revolutions are the mad inspiration of history." It is the idea of a "new order," which predominates since the American and French revolutions, rather than that of a return to an old one, as suggested by the cosmic revolution of heavenly bodies. An early use of the word around 1450, toward the end of the Middle Ages, has been noted; but it came into more general use after 1600, first meaning a return to the old order, but afterward in the modern sense. These dates show that revolution as a major political term parallels the emergence of the modern state around 1600. Hence in its core meaning a revolution constitutes a challenge to the established political order and the eventual establishment of a new order radically different from the preceding one. One recent writer defines it in its most common sense as "an attempt to make a radical change in the system of government." It might be added that a successful revolution is more than an attempt; it does radically change such a system. The great revolutions of European history, especially the Reformation and the English, French, and Russian revolutions changed not only the system of government but also the social system in all its ramifications, economic, social, and cultural.

THEORIES OF REVOLUTION

Classical theories of revolution

Ancient and medieval theories. In light of such a definition, it is hardly surprising that the history of political thought consists of a series of theories of revolution. In classical antiquity the preoccupation of Plato, Aristotle, and later writers was profoundly antirevolutionary, even though in point of fact Plato's (and Socrates') thought was revolutionary in its implications, or perhaps more correctly it was counter-revolutionary, in that it sought to recapture the old order and rejected the democratic order that prevailed in the Athens of their day. The Sophistic revolution, opposed by Plato, was seen to have destroyed the old *nomos,* or tradition, of right and law. It has been persuasively argued that Plato saw history as a regress and combatted the "liberal temper in Greek politics." Surely Socrates' sharp criticism of Pericles as the corrupter of the people (reported by Plato in the Gorgias) was counter-revolutionary in sentiment and intent. The implicit theory of revolution among the Greeks is that it results from the development of *anomia,* a disintegration of the traditional values and beliefs, founded in religious convictions, as to what is right and wrong. In a sense Plato's Republic seeks to show that only a fixed and unalterable system of values, based upon the rule of guardians who have learned philosophical truth through the cognition of eternal ideas, can hope to escape from the revolutionary turmoil that unstable values cause.

Aristotle, in his Constitution of Athens, displays a concrete insight into what appears as a sequence of revolutions and counter-revolutions in the political order of that great city. He generalized from that story, and presumably from similar (but now lost) studies of comparable stories of other Greek cities, and distilled what is commonly believed to be the first general theory of revolution. Empirical material of great interest is also contained in the History of the *Peloponnesian* War of Thucydides, who, although sophistical in his philosophical outlook, is also anti-revolutionary. The constant oscillation between oligarchic and democratic regimes in the Greek cities, bound up with the rivalry of Athens and Sparta, seemed to him an important cause of the lethal conflict of the Peloponnesian War, and he gives some highly critical descriptions of revolutionary upheavals—*e.g.,* "And so there fell upon the cities on account of revolutions many grievious calamities. . . ." Aristotle, writing after these calamities, formulated his theory of revolution in light of these experiences (Book V of Politics). Aristotle relates the causes of revolutions to his typology of forms of government and their cyclical sequence, which Plato had hinted at before him. He stresses a fourfold set of conditions, in keeping with his general theory of causation. The material conditions he recognizes are in the distribution of wealth and the related class structure, simplistically seen as the division between rich and poor. The ideological aspect (to use a modern term) he sees provided by the ideas of justice that mold party divisions. The efficient cause consists of the makers of revolutions, that is to say the leaders of the movements that are seeking power and, therefore, the transformation of the political order. This latter is itself the fourth "cause": the end or purpose of a revolution is a radically altered political order. Aristotle's theory can be stated more generally by saying that any radical alteration in basic value or beliefs provides the ground for a revolutionary upheaval.

From this it can be seen that Aristotle's was an elaboration of the Platonic view, and it is not surprising that Cicero too should have shared this outlook. In his fragmentary De *republica* he says that "a disagreement among the citizens is termed a breaking up because the citizens divide into several factions." He introduces the term sedition at this point, further emphasizing his distrust of revolution and the importance of avoiding it. It is the task of the statesman to be "always armed to meet emergencies which unsettle the constitution," he argues.

Revolution in medieval thought

This preoccupation with stability continued into the Middle Ages. Political thinkers, however, were concerned over the problem of resistance to revolution, and more particularly the right to resist, which was seen in the perspective of the division of medieval authority between the two swords of secular and priestly office and the related question of tyranny and monarchy. Accepting the latter as the natural and good form of government, in the image of God's rule over the universe, they had to face the issue of the abuse of power and the possibility of a usurper. These were the two kinds of tyranny medieval thinkers saw as perversions that the church had to prevent—if need be, by excommunicating the offending prince and calling for the resistance of his feudal vassals. The actual revolutionary in the medieval sense was, thus, the offending prince himself; and it was the task of his subjects to redress the balance with all available means, including tyrannicide. Even as conservative a thinker as St. Thomas Aquinas could, therefore, justify the killing of a tyrant when sanctioned by the ecclesiastical authorities. Part of his commentary on the fifth book of Aristotle's Politics is characteristically entitled "For what causes tyrannies and kingdoms are dissolved," and he asserts that injustice is the main cause. In keeping with Aristotle's theory of justice as expressed in his Ethics, Aquinas holds that seditions are caused by inequality. Revolution is seen as sedition and is understood to spring from the corruption that outrages the sentiment of justice and equality. Other medieval thinkers, such as Marsilius of Padua, leaned in the same direction. This theory of revolution was reinforced by the Augustinian indifference to the worldly city and the general belief that its iniquities must be endured by the good Christian. It is clear that the antirevolutionary tendency had become even more pronounced in the Middle Ages than it was in classical thought. Even the conciliar movement of the 15th century, though in many ways subversive of the medieval tradition in church and state, did not develop any theory of revolution.

The beginning of the modern view

Renaissance and early modern theories. Such a theory only became possible in the secular thought of humanism, with its Renaissance interest in classical antiquity. Niccolb Machiavelli especially, but others before and after him as well, were inclined to see revolution in a new light. In general, Machiavelli, like Plato, expounded revolutionary thoughts without theorizing about revolution. The term does not appear in his writings; he speaks of rebellion *(seditio)*, as the Middle Ages had done, and of changes of political order—Cicero's *mutatio rerum*—because like the ancients he was primarily interested in stability. He sought the means to found a lasting and stable political order, as Plato and Aristotle had, and as Polybius believed the Romans had succeeded in doing. But in Machiavelli's writings there is no evolutionary notion of a meaningful historical progress of which a revolution might be a crucial step. As Hannah Arendt has said, "the specific revolutionary pathos of the absolutely new . . . was entirely alien to him." Even so, he was, if not "the spiritual father of revolution," then at any rate the thinker who cleared the road for its appearance by the stress he laid upon the founding process, and the willingness to condone the violence associated with it.

The radical recognition of the creative potential of revolution first appears in the course of the English Revolution, and more particularly in the writings of John Milton, later popularized by John Locke. Milton, in his defense of what the Puritans had done, proclaimed the "right of revolution." This was no longer merely the right of resisting a tyrannous ruler, but the right to "choose" their own form of government. In ***The Tenure of Kings and Magistrates (1649)*** Milton speaks of the right of a free nation "as oft as they judge it for the best, either to choose or reject" a ruler, "though he be no tyrant," "merely by the liberty and right of free-born men to be governed as seems to them best." Milton's is the decisive breakthrough to the modern position, which asserts that revolution is vitally linked to the realization of freedom. Milton mocks those who think that a nation can be free which does not claim this right, those "whose government, though not illegal, or intolerable, hangs over them as a lordly scourge. . . ." Milton's contemporary, the political philosopher, James Harrington, shared this belief in the right of a free people to organize itself, and John Locke expresses the core idea of the right of revolution in what he calls an "appeal to heaven." But Locke did not maintain Milton's vigorously libertarian view. In his ***Second Treatise of Government,*** though he rejects the critics of the revolutionary doctrine of Milton and Harrington, he resumes the old argument about the "ill usage of arbitrary power" and suggests that "such ***revolutions happen"*** not "upon every little mismanagement in publick affairs." He argues about "illegal attempts made upon their [the people's] Liberties and Properties," and about "unlawful violence" and "tyranny."

Revolution as the agent of freedom was, of course, the concept behind both the American and French revolutions. Not the defense of old and established liberties but the conquest of new and fuller ones was seen as the essence of the revolutionary thrust. In order to argue thus, the doctrine of natural rights was developed out of the old natural-law notions and given fuller scope than it had hitherto possessed. Jean-Jacques Rousseau, though by no means an advocate of revolution, which, he tells us in his ***Confessions,*** he had come to detest, nonetheless became the revolutionary thinker par excellence. For what he sketched was a republican and democratic ideal, rooted in Swiss tradition, but with profoundly revolutionary implications for France and other European monarchies. Rousseau himself argued that only some people in some historical constellations were capable of genuine republican freedom, but if so, a revolution to achieve it was natural. The German philosopher Immanuel Kant, dubbed the "red Jacobin of Konigsberg" by his contemporaries, went a step further and argued that though a revolution may be accompanied by "miseries and atrocities" (as in the French Revolution) it was still a "sign" of the "moral propensity in man," and hence related to human progress. In spite of all its excesses, he saw the French (as well as the American) Revolution as an attempt "to secure freedom under law through the establishment of a republican constitution." For Kant, a revolution was a fact of nature, not morally or legally allowable as such but to be accepted as "natural" if directed toward a higher moral goal.

The 19th-century theories. The French and American revolutions, so vitally related to the progress of freedom, raised the question that has dominated political theory ever since, namely that of the ***historical meaning*** of a revolution. The broad cultural implications of these revolutions eventually led to a new definition of "revolution" as a spiritual upheaval through which a group seeks to establish a new foundation for its existence, no longer a merely political process but part of the unfolding of human potentiality. Though Kant and the Marquis de Condorcet had suggested this possibility, the decisive philosophical turn toward this view was G.W.F. Hegel's philosophy of history. The Hegelian metaphysics, which saw the progressive unfolding of human history as a manifestation of the world spirit and as such as the realization of freedom, though it did not explicitly employ the term revolution as a key concept, was in fact a great revolutionary interpretation of mankind's progress. The idea of world revolutionary individuals, who fulfill the destiny of nations by bringing about a new and broader freedom, is linked to this revolutionary interpretation. Hegel therefore interpreted the Reformation as a revolution using such language as this: "The ancient and ever-preserved inwardness of the German people had to effect this revolution out of the simple, modest heart." The political results are incidental to this spiritual revolution, and as such they are not of primary importance.

Revolution and historical progress

Passing over the English Revolution, as do so many continental thinkers, Hegel then concentrates on the French Revolution, which he saw as a manifestation of the French genius for abstraction. The revolution is "world-historical" in its significance, "the principle of the French Revolution permeated almost all modern states" and particularly all the Latin nations "fell under the dominion of liberalism." "Thus liberalism traversed the Latin world as an abstraction emanating from France." Regarding England, Hegel claims: "Abstract and general principles consequently have no attraction for Englishmen." A close student of Edmund Burke and his famous ***Reflections on the Revolution in France,*** he echoes here a notion that has persisted, namely that English people are resistant to revolutionary upheavals and prefer gradual and evolutionary development. At the end of his life Hegel sharply condemned the reform movement in England and expressed the view that it would destroy the English constitution's basic freedom. In the case of England he did not perceive the revolutionary thrust in positive terms.

The Hegelian theory of revolution as part of an evolutionary process of world-historical dimensions was made the basis of what is unquestionably the most dominant modern theory of revolution, the Marxian. By focussing attention upon the material factor of the control of the means of production and the class that exercises such control, Karl Marx believed he had given the Hegelian notions a scientific basis. He saw history as proceeding according to ascertainable and regular laws, and revolutions as serving a decisive function in this process: a new and ascendant class overthrows an obsolete one and a new social order replaces the outworn one. The productive forces of a certain stage clash with the prevailing "relations of production." That is, the existing social order hinders the further development of these relations, and, as tensions increase, a revolutionary situation develops. In Marx's view power and politics are crucial; the seizure of power by the revolutionaries is the first step that determines all the rest. The struggle of the classes culminates in civil war and revolution. But unlike Hegel's, Marx's vision is eschatological. The process is not conceived as continuing indefinitely; rather, the revolution of the class-conscious Communist elite of the proletariat is seen as achieving a final stage which puts an end to all class rule and exploitation. The class struggle has then come to an end, and with it the chance of further revolutionary upheavals.

Marxian theory

Marx thought that this final stage would be reached in the most advanced industrial nations at the end of their period of bourgeois class rule. In fact, the Communist revolution has succeeded in preindustrial and backward countries, and this success did not bring the classless society. It has been argued with persuasive insight that a new class has arisen: the ***apparat,*** that is to say the managerial bureaucracy, controlling the means of production in the name of the Communist Party. Nor did the industrial proletariat play a decisive role in these revolutions, which hardly represented a greater realization of freedom. In the former colonial regions, freedom from foreign dominion was achieved in this way, but usually at the price of a reduction in civil liberties.

THE CAUSES OF REVOLUTION

Unlimited revolutions

The processes of revolution. The problem of what causes a revolution, first explicitly stated by Aristotle, has had considerable vogue. The sharp contrast between strictly governmental revolutions, bordering on the coup d'etat, and the great revolutions of European history make it desirable to separate the two in a discussion of causation. For the great revolutions, in view of their culturally global character, it seems futile to search for definite causes, since the entire preceding state of a society is involved, and all one can hope to do is to describe the phases in the unfolding of the revolutionary process. A number of attempts have been made to sketch this pattern of revolution. One impressive attempt at overall interpretation, in the Hegelian tradition of seeing the great revolutions as part of the march of history, was made by Eugen Rosenstock-Huessy. He links the unlimited revolutions to the development of Western culture and the formation of nations. According to him, each of these revolutions was made by one of the major nations of Europe in the course of becoming a nation and discovering its own style of life, a particular version of Western culture.

This interpretation fits the English and French experience reasonably well, but becomes less tenable when applied to other revolutions, and it has been bitterly criticized for its historical inaccuracies. Yet Rosenstock-Huessy has contributed some important insights; he has shown how each of these total revolutions came to see itself as a "world revolution" whose makers expected that its values and beliefs would become universal throughout Europe. This bold theory of Europe's great revolutions treats the Russian Revolution as the last of its kind, after which the world is moving "out of revolution." The theory contains many fascinating details, such as the widening group of active revolutionaries, culminating in the one or two million proletarians of the Soviet Revolution. Rosenstock-Huessy sees each revolution as in a sense directed against the preceding one, echoing to some extent Marxian thought patterns in this respect. This striking synthesis of Hegelian and Marxian elements in terms of an autonomous process of nation forming is primarily a cultural theory of revolution, but the political element occupies the centre of the analysis, for in its violence and suddenness the political revolution epitomizes all the rest. Yet, it must be distinguished from the limited type of political revolution (such as the coup d'etat) that signifies little more than overthrow of the government.

Limited revolutions

Some contexts seem especially favourable to this kind of revolution. According to one count, there have been 68 revolutions during 65 years of Bolivian history. Actually many of these were mere coups d'etat. But even after discounting these, enough remain in Bolivia and other Latin American countries to conclude that Latin America, like ancient Greece, is favourable to revolutionary change. By contrast, Britain and the United States, Switzerland, and the Scandinavian countries have been rather free of it. These, in turn, form a striking contrast to France and a number of other European countries. The experience of France would suggest that if a regime represses much needed change and is overthrown, it may become very difficult to reestablish a political order that will last, unless provisions are made for its recurrent change. It may be necessary, as in Switzerland, to go as far as to allow a total revision of the constitution, but even more limited amending arrangements may accomplish the required flexibility.

Turning now to the process of limited political revolution, one modern historian (Crane Brinton) has constructed what he considers an "anatomy" of revolutions by abstracting the political from the broader cultural and social context. This political pattern, taken by itself, appears rather uniform. Brinton based his tentative generalizations upon some of the great, "unlimited" revolutions—the English, the French, and the Russian. We may summarize his findings as follows: The signs of an approaching revolution are not very distinctive. There are more and more stresses and strains, some of which are endemic in any society, but whose multiplications eventually leads to a breakdown of the political order. Both legitimacy and authority disintegrate; such power as is still effectively wielded depends more and more upon coercive means and is thereby limited. At the same time, hitherto unrecognized groups gain power and influence. Such groups eventually fan out into revolutionary movements with the openly declared goal of destroying the existing order and replacing it by another.

The phases of revolution

There is a definite sequence of stages or phases to the revolutionary process, but it appears doubtful that what holds of the unlimited revolutions necessarily applies to the limited ones. Brinton noted that in the first stage utopian expectations run high, and the revolutionaries engage in much perfectionist rhetoric. But this phase does not last very long. The practical tasks of governing have to be faced, and a split develops between moderates and radicals. It ends in the defeat of the moderates, the rise of extremists and the concentration of all power in their hands. This is the second stage. Such extreme concentration of power is followed by the "terror," and the ever more radical deployment of violence. The third stage is the desperate effort to realize the revolutionary goals at all costs. This overextension engenders a reaction, after the French case often called the Thermidorean reaction (from Thermidor, French name of the month in which the reaction took place), bringing a period of convalescence during which the revolutionary fervour subsides. This period of resignation and dissension then leads to the setting up of a dictator who is still animated by some of the revolutionary aspirations; his rule constitutes the fifth stage. Gradually, the practical problems of governing gain the upper hand and old habits of life reappear. The revolutionary symbols lose their hold and the dictatorship appears as naked power. At this point, the time is ripe for a restoration, an attempt to reestablish the old regime. But the restoration never succeeds in restoring what has been; new institutions are designed for the changed social structure the revolution has brought into being. Brinton concludes that the overall result of such a revolution is the achievement of greater governmental efficiency. He notes that, of the countries upon which his generalizations are based, all "emerged from their revolutions with more efficient and more centralized governments."

Only the last of these aspects of the revolutionary process, and possibly the first, are to be observed in the American Revolution. That this strengthening of government is a common result of revolutions is understandable enough in light of the fact that revolutions are made against weak governments. A good many other political revolutions do not run through the phases Brinton abstracted from the unlimited revolutions of European history. The explanation may lie in the absence of a sharp clash between moderates and radicals that in turn is related to the absence or weakness of the ideological factor. The Roman revolution by which the Republic was replaced by the empire illustrates the point. Like the Roman revolution, many political revolutions result from the breakdown of an established system.

If there is no ideological ferment, what are the causes of revolution? No particular useful analysis results from listing abstract general values such as freedom, security, equality, or justice as causes of revolution, on the ground that they provide a basis for revolutionary sentiment. Such general motivations will enter the thinking of revolutionaries as of other political leaders; they may be real motivations or mere rationalizations. It has already been noted that a multiplication of stresses and strains, resulting from a structure that is too rigid, is the most usual cause of revolutionary sentiment. George Pettee called these stresses and strains "cramps," implying a "maladjustment with accompanying strain." He then proceeded to analyze the prerevolutionary situation in terms of economic, ideological, social, and political cramps, resulting from institutional decay and the decadence of the elite. Pettee insists that only a cumulation of such stresses and strains will produce a revolutionary outbreak. A clash of loyalties develops which in due course produces a revolutionary crisis.

The collapse of the state

In the prerevolutionary situation, ever larger groups of people are alienated from the established political order. Existing law loses its legitimacy and seems arbitrary and its enforcement unjustified. The old established sense of right that the Greeks called ***nomos*** fades, and the resulting state of affairs is one without *nomos*; and the Greek word *anomia* has once again become fashionable: anomie, though often used rather vaguely, is meant to describe the present-day moral confusion and lack of a firm sense of values. In such a prerevolutionary period, usually efforts at reforming the political and social order are attempted, but they often fail and thus enhance the sense of frustration and alienation. Each such period has produced one or more striking figures who try hard: Strafford in England under Charles I, Turgot and Necker in France under Louis XVI, and Stolypin in Russia under Nicholas II were genuine reformers, but their very ability and dedication having produced no success served to prove that the system was beyond reform. These reformers highlight its rottenness and corruption. Not long after this sort of effort, the political order (the state) collapses. It is often some minor matter that serves to demonstrate that the system has ceased to function, but it may be a major policy that helps to make manifest the breakdown, such as France's Algeria policy before Charles de Gaulle.

If one considers these facts and compares a large number of revolutions, both unlimited and limited, it is striking how often they have been precipitated by an unsuccessful war. In fact, losing a war is dangerous for any political regime, but especially for one with prerevolutionary tensions. It would be superficial, however, to consider the war the cause of the revolution. The war is itself the result of the deeper strains that produced the prerevolutionary situation. It has often been remarked that autocratic regimes when confronted with internal conflict and restlessness have sought to assuage the aggressive elements by engaging in attacks upon other states. Mussolini's aggression against Ethiopia, Hitler's against Poland, and Stalin's against Finland are merely recent examples of a propensity that has prevailed through the ages. If the aggression is successful, it may serve as the safety valve for the regime, but if it is not, it may precipitate a revolution.

The medieval doctrine of tyrannicide, with its stress on the abuse of power, pointed to another "cause" of revolutionary overthrow. But as in the case of war, one must ask further what caused such abuse. The old answer that power tends to corrupt, while not without some value, is inadequate. The tyrant may be a self-seeking bully; more probably he is a man afraid who is trying to master an unmanageable situation resulting from failure of the regime to adapt to changing circumstances.

The place of ideology

The aims of revolutionaries. So far, the discussion of causation has been largely in terms of the "efficient cause^{w}—the so-called scientific and value-free analysis of antecedents. But there is also the teleological causation, which we reach by inquiring about the end or purpose of the revolutionaries. For it is not merely a matter of considering the antecedents but also of analyzing the revolution's thrust. A radical alteration of the political and social order may occur, and appear in retrospect to have brought about a revolution, but more typically, the revolutionaries have a global goal, such as a "free" society or a constitutional order or a socialist community. Such goal definitions usually contain a utopian element, and the more unlimited the revolution, the more utopian the goal. The utopianism, often eschatological, that is to say, conceived as a final state in mankind's evolution, usually occasions the excesses of violence, not only because superhuman efforts are required to achieve the unattainable and the ineffable, but also because such goals beyond rational communication become a fertile ground for factionalism among the revolutionaries themselves. It has been said that "the revolution devours its children," because the attempt to achieve the utopian goal brings on the fanaticism of violence. It is a collective frustration-aggression psychosis that causes these self-destructive propensities. Even so, the goal, end, or purpose of the revolution, unattainable as it may be, has been a powerful factor in the etiology of revolutions. The goal predominant in the minds of the leaders may be the *goût* de ***pouvoir,*** but such love of power is usually reinforced by ideological fixations. To the scientific observer, the English, American, and French revolutions, as well as many postcolonial revolutions, are unusual in that their goal of establishing a constitutional order called for a considerable degree of self-denial on the part of the leaders of the revolution. The drama and the tragedy of this situation is epitomized in Oliver Cromwell's bitter struggles with parliament to make and abide by a constitution that they had made at his dictatorial demand.

A particular kind of constitutional revolution has occurred in recent times, the so-called negative revolutions in France, Italy, and Germany following World War II. In these cases constitutional orders were fashioned by people who had no great enthusiasm for or belief in constitutionalism but preferred it to the dictatorial regimes that had been overthrown. The same might be said of the quasi-revolution that brought de Gaulle to power in 19.58. While approved in a referendum, the constitution of the Fifth Republic was the work of legal technicians and has evolved quite differently from what its engineers intended. It unquestionably revolutionized the French government, and to some extent French society; but it was a "revolution from above," imposed by the imperious will of a leader who was not the maker but the beneficiary of the revolutionary ferment.

Revolutions, whatever their immediate occasion, would seem to be the result of deep-rooted and slowly evolving political and social malformations rather than the sudden outbreak that they appear to be at the surface. But in their last, culminating phase they are sudden and violent. And after the seizure of power by the revolutionaries, revolutions tend to run through certain phases, especially the unlimited, global revolutions of modern Europe. The more limited political revolutions do not exhibit all or even part of the phases; some of them closely resemble a coup d'etat. Such a coup is a stroke at the particular persons wielding power, but it is made merely for the purpose of replacing them by others, without any intended or actual change in the system. It is the violent substitute for what an election normally accomplishes, and many an overturn (stasis) discussed by Aristotle in his writing on revolution is merely such a coup. Coups often occur in the succession of revolutionary leaders: the replacement of Soviet Premier Nikita Sergeyevich Krushchev, for example, was clearly a coup d'etat. Before a new political order becomes established, a coup is the only way of changing the leadership, and much revolutionary violence occurs in connection with such coups and their prevention.

THE STAGES OF REVOLUTIONARY CONFLICT

It has been noted that there is an interlude of good feeling immediately after the outbreak of a revolution. For a time enthusiasm and optimism reign supreme, but this is a short-lived period that cannot last. The practical task of realizing the revolutionary program and attending to the everyday business of government leads eventually to sharp controversy and in turn to power struggles.

In the next phase the ideological issues come to the fore. The programmatic certainty of the early days vanishes and is replaced by strife and disagreement. Many possible avenues of implementation are explored, advocated, and discarded. All the French revolutionaries may have been followers of Rousseau; still they guillotined each other. Intertwined with this ideological phase is a succession of coups d'etat, which eventually leads to an exhaustion, after having precipitated one or more civil wars. In some cases, such a civil war constitutes a separate phase, but in many others it remains latent or overlaps with the succession struggles. It is true that in the French Revolution the radicals won in these wars and succession crises, but this is by no means generally true. Neither in England (Crom-

The phase of exhaustion

well) nor in America nor in Russia did the radicals triumph; Trotsky had to yield to the more conservative Stalin. The exhaustion from internecine conflict also may be preceded by foreign intervention as a separate phase; but this again depends upon circumstances. The greater the threat the revolution poses for other states, the more likely is such intervention, provided that the neighbouring states are not preoccupied with other matters, as they were in the case of the English Revolution, or weakened from a long war, as in 1917.

It is arguable whether the phase of exhaustion, the Thermidorean reaction, is still part of the revolutionary cycle. It was out of this period in the French Revolution that the dictatorship of Napoleon emerged. The Soviet leadership, for example, having determined to avoid the Thermidorean reaction and consequent dictatorship and restoration, declared that their regime was a permanent revolution. Some critical observers of the Soviet scene were inclined to doubt their success and to interpret Stalin's regime as a typical Thermidorean phase. This interpretation seems in retrospect rather doubtful, since the terror culminated during his leadership. Although Stalin may have had the chance of becoming the Napoleon of the Russian Revolution, and though he did in fact build a new system of autocracy, he appears to have been the executioner of the revolution. Like Napoleon, he excused the violence of his rule by reference to external threats. But it remains to be seen how close the evolution of the Russian Revolution will come to the pattern of phases deduced from the French revolution.

Revolutions of the limited kind often have not had a period of exhaustion and hence no Thermidor. After achieving the new order, the revolutionaries were able to continue it without any restoration. Thus Hitler's revolution led to a new Fascist regime in Germany, as had Mussolini's before him in Italy, and then on into war and destruction of the new order. No period of this kind seems in sight in China, though the death of Mao Tse-tung may initiate one.

CONCLUSION

In conclusion, one might say that the phenomena of political revolution and resistance are endemic in any political order, and that they are closely related. To avoid them or reduce their menace to a minimum, effective change by means of gradual transformation has to be organized to make possible recurrent adaptations of the institutions and processes of a political order to evolving values, interests, and beliefs. Otherwise violence will spread, sporadically as resistance at first, globally and all-engulfing as revolution afterward. Political orders resemble forests and families. They contain the potentiality of self-renewal, but this potentiality does not exclude the chance of failure and ultimate death. Revolution, when successful, signalizes the passing of such a political order. It is not in itself a good, as contemporary political romantics are inclined to feel, but it is better than the death of the society that such an order is intended to serve. The old Roman adage still holds: *Videant consules ne respublica detrimentum capiat* ("Let the consuls see that the republic suffer no harm").

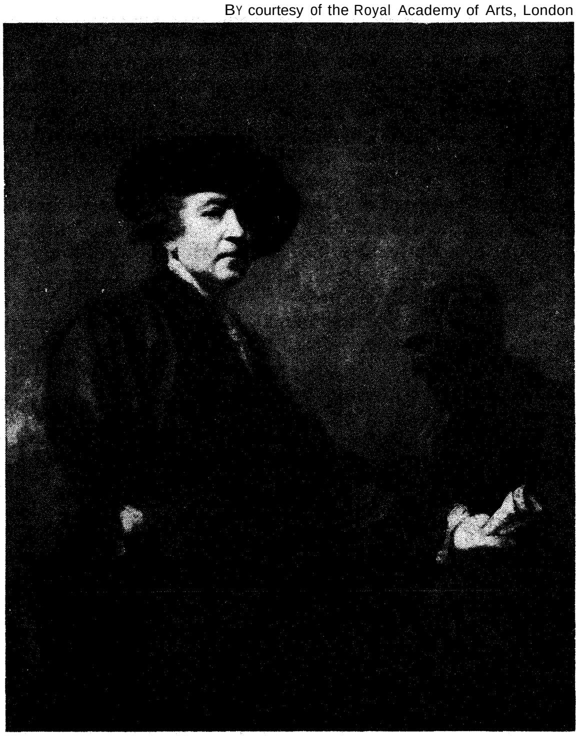
BIBLIOGRAPHY. Two classic statements on the subject of revolution are JOHN LOCKE, *Second Treatise on Government* (1689), on the right of revolution; and EDMUND BURKE, *Reflections on the Revolution in France* (1790), *a* famous criticism of revolutionary violence. Attempts to define the concept of revolution may be found in the essays of HANNAH ARENDT, *On Revolution* (1963); in a collection of articles edited by C.J. FRIEDRICH, *Revolution* (1966); and in C.A. JOHNSON, *Revolutionary Change* (1966). L.S. FEUER (ed.), *Mars and Engels: Basic Writings on Politics and Philosophy* (1959), contains excerpts from the writings of Marx and Engels that are essential to an understanding of the Marxist theory of revolution. V.I. LENIN, *The State and Revolution: The Marxist Theory of the State and the Tasks of the Proletariat in the Revolution* (1917), presents a significant extension of the thought of Marx and Engels. Comparative studies of revolutions include C.C. BRINTON, *The Anatomy of Revolution,* rev. ed. (1952), an analysis of the process of revolution as exemplified by the French, American, English, and Russian Revolutions; and R.B. MERRIMAN, *Six Contemporaneous Revolutions* (1938), an historian's comparative description and evaluation of a revolutionary age. C. MALAPARTE, *Coup d'Etat: The Technique of Revolution,* trans. by S. SAUNDERS (1932), is concerned primarily with the lessons of Fascism. An interesting study of the interaction between revolution and the relation between states is G.A. KELLY and L.B. MILLER, *Internal War and International Systems* (1969). Other contemporary theoretical discussions include: B. MOORE, *Social Origins of Dictatorships and Democracy* (1966); N.R.C. COHN, *The Pursuit of the Millennium: Revoiutionary Messianism in Medieval and Reformation Europe and its Bearing on Modern Totalitarian Movements,* 2nd ed. (1961; rev. paperback ed., 1970), a general study with stress on Utopian objectives; S.P. HUNTINGTON, *Political Order in Changing Societies* (1968), a study of contemporary political development; H.J. LASKI, *Reflections on the Revolution of Our Time* (1943); H.D. LASSWELL and D BLUMENSTOCK, *World Revolutionary Propaganda* (1939); and E. ROSENSTOCK-HUESSY, *Out of Revolution* (1940, pa. 1969).

(C.J.F.)

Reynolds, Sir Joshua

The British art world of the middle and late 18th century was dominated by Sir Joshua Reynolds, one of England's greatest portrait painters and artistic theorists. Through his art and teaching, he attempted to lead British painting away from the indigenous anecdotal pictures of the early 18th century toward the grandiloquent formal rhetoric of the continental tradition of the "Grand Manner." In the *Discourses,* among the most important art critical writings of the time, he outlined the essence of grandeur in art and suggested the means of achieving it through rigorous academic training and study of the old masters of art.

BY courtesy of the Royal Academy of Arts, London



Reynolds, self-portrait, oil painting. 1773. In the Royal Academy of Arts, London.

Joshua Reynolds was born on July 16, 1723, at Plympton in Devonshire. He was educated at the Plympton grammar school of which his father, a clergyman, was master. The young Reynolds became well read in the writing of classical antiquity and throughout his life was to be much interested in literature, counting many of the finest British authors of the 18th century among his closest friends. Influenced by the essays of the prominent English portrait painter Jonathan Richardson, Reynolds early aspired to become an artist, and in 1740 he was apprenticed for four years in London to Richardson's pupil and son-in-law, a conventional portraitist named Thomas Hudson. In 1743, he returned to Devon and began painting at Plymouth naval portraits that reveal his inexperience. Returning to London for two years in 1744, he began to acquire a knowledge of the old masters and an independent style characterized by bold brushwork and the use of impasto, a thick surface texture of paint, such as in his portrait of "Capt, the Hon. John Hamilton."

Italian influence

Back in Devon in 1746 he painted a large group portrait of the "Eliot Family" (c. 1746/47; Earl of St. Germans, Port Eliot, Cornwall), which clearly indicates that he had studied the large-scale portrait of the "Pembroke Family" (1634–35; Wilton House, Wiltshire) by the Flemish Baroque painter Sir Anthony Van Dyck (1599–1641), whose style of portrait painting influenced English portraiture throughout the 18th century. In 1749 he sailed with a friend to Minorca, one of the Balearic Islands off the Mediterranean coast of Spain. A fall from a horse detained him for five months and permanently scarred his lip—the scar being a prominent feature in his subsequent self-portraits. From Minorca he went to Rome, where he remained for two years, devoting himself to studying Italian art from ancient to his own times. Returning to England via Florence, Bologna, and Venice, he became absorbed by the compositions and colour of the great Renaissance Venetian painters of the 16th century: Titian, Jacopo Tintoretto, and Paolo Veronese. The Venetian tradition of atmospheric painting with its emphasis on colour and the effect of light and shading had a lasting influence on Reynolds, and, although all his life he preached the need for young artists to study the sculptural definition of form characteristic of Florentine and Roman painters, his own works are redolent of the Venetian style.

In 1753 Reynolds settled in London, where he was to live for the rest of his life. His success was assured, from the very first, and by 1755 he was employing studio assistants to help him execute the numerous portrait commissions he received. The early London portraits have a vigour and naturalness about them that is perhaps best exemplified in a likeness of "Hon. Augustus Keppel." The pose is not original, being a reversal of the "Apollo Belvedere," an ancient Roman copy of a mid-4th-century-BC Hellenistic statue he had seen in the Vatican. But the fact that the subject is shown striding along the seashore introduces a new kind of vigour into the tradition of English portraiture. In these first years in London, his knowledge of Venetian painting is very apparent in such works as the portraits of "Lord Cathcart" (1753/54; Trustees of Earl Cathcart, on loan to Manchester City Art Gallery, Lancashire) and "Lord Ludlow" (1755; Woburn Abbey, Bedfordshire). Of his domestic portraits those of "Nelly O'Brien" and of "Georgiana, Countess Spencer, and Her Daughter" (1761; Earl Spencer, Althorp, Northamptonshire) are especially notable for their tender charm and careful observation.

After 1760 Reynolds' style became increasingly classical and self-conscious. Falling under the influence of the classical Baroque painters of the Bolognese school of the 17th century and the archaeological interest in Greco-Roman antiquity that was sweeping Europe at the time, the pose and clothes of his sitters took on a more rigidly antique pattern, in consequence losing much of the sympathy and understanding of his earlier works.

Presidency of the Royal Academy

There were no public exhibitions of contemporary artists in London before 1760, when Reynolds helped found the Society of Artists and the first of many successful exhibitions were held. The patronage of King George III was sought, and in 1768 the Royal Academy was founded. Although Reynolds' painting had found no favour at court, he was the obvious candidate for the presidency, and the King confirmed his election and knighted him. Reynolds guided the policy of the academy with such skill that the pattern he set has been followed with little variation ever since. His yearly Discourses clearly mirrored many of his own thoughts and aspirations, as well as his own problems of line versus colour and public and private portraiture, and gave advice to those beginning their artistic careers.

From 1769 nearly all of Reynolds' most important works appeared in the academy. In certain exhibitions he included historical pieces, such as "Ugolino" (1773; Knole House, Kent), which were perhaps his least successful works. Many of his child studies are tender and even amusing, though now and again the sentiment tends to be excessive. Two of the most enchanting are "Master Crewe as Henry VIII" (1775–76; Lord O'Neil, London) and "Lady Caroline Scott as 'Winter' " (1778; Duke of Buccleuch, London). His most ambitious portrait commission was the "Family of the Duke of Marlborough (1777; Blenheim Palace, Oxfordshire).

In 1781 Reynolds visited Flanders and Holland, where he studied the work of the great Flemish painter of the Baroque style Peter Paul Rubens. This seems to have affected his own style, for in the manner of Rubens' later works the texture of his picture surface becomes far richer. This is particularly true of his portrait of the "Duchess of Devonshire and Her Daughter" (1786; Chatsworth House, Derbyshire). Reynolds was never a mere society painter or flatterer. It has been suggested that his deafness gave him a clearer insight into the character of his sitters, the lack of one faculty sharpening the use of his eyes. His vast learning allowed him continually to vary his poses and style. In 1782 Reynolds had a paralytic stroke and about the same time he was saddened by bickering within the Royal Academy. Seven years later his eyesight began to fail, and he delivered his last Discourse at the academy in 1790. On February 23, 1792, he died in London and was buried in St. Paul's Cathedral.

Social life

Reynolds preferred the company of men of letters to that of his fellow artists. Although his 14th Discourse (1788) is a tender and moving appreciation of his rival, the painter Thomas Gainsborough, who stood for so much that he himself disliked in painting, it was in the company of the lexicographer and author Dr. Samuel Johnson, the statesman and philosopher Edmund Burke, and the dramatist, novelist, and poet Oliver Goldsmith that Reynolds was happiest. When Goldsmith died, Reynolds could not bring himself to paint for a whole day, and a moving essay he wrote on his friend showed that he could write a portrait as well as he could paint one. Reynolds and his friends were members of The Club, which he established in 1764. He never married, and his house was kept for him by his sister Frances.

Reynolds' state portraits of the King and Queen were never considered a success, and he seldom painted for them; but the Prince of Wales patronized him extensively, and there were few distinguished families or individuals who did not sit for him. Nonetheless, some of his finest portraits are those of his intimate friends and of fashionable women of questionable reputation.

Unfortunately, Reynolds' technique was not always entirely sound, and many of his paintings have suffered as a result. After his visit to Italy, he tried to produce the effects of Tintoretto and Titian by using transparent glazes over a monochrome underpainting, but the pigment he used for his flesh tones was not permanent and even in his lifetime began to fade, causing the overpale faces of many surviving portraits. An example of this can be seen in the "Roffey Family" (1765; City of Birmingham Museum and Art Gallery). This paleness often has been increased by injudicious cleaning. In the 1760s Reynolds began to use more extensively bitumen or coal substances added to pigments. This practice proved to be detrimental to the paint surface. Though a keen collector of old master drawings, Reynolds was never a draftsman, and indeed few of his drawings have any merit.

MAJOR WORKS

"Captain the Honourable John Hamilton" (1746; Duke of Abercorn Collection, Ireland); "Lady Chambers" (1752; Iveagh Bequest, Kenwood House, London); "Honourable Augustus Keppel" (1753–54; National Maritime Museum, Greenwich, London); "Anne, Countess of Albemarle" (1757–59; National Gallery, London); "Nelly O'Brien" (1760–62; Wallace Collection, London); "Lord Ligonier" (1760; Tate Gallery, London); "Garrick Between Comedy and Tragedy" (1760–61; Lord Rothschild Collection, Rushbrooke, West Suffolk); "Mrs. Abingdon As 'The Comic Muse' " (1764–65; Waddesdon Manor, Buckinghamshire); "The Honourable Henry Fane with His Guardians, Inigo Jones and Charles Blair" (1766; Metropolitan Museum of Art, New York); "Mrs. Richard Hoare with Her Son" (1767–68; Wallace Collection, London); "Dr. Samuel Johnson" (1770–80; Tate Gallery, London); "William Robertson" (1772; Scottish National Portrait Gallery, Edinburgh); "Three Ladies Adorning a Term of Hymen" (1773–74; National Gallery, London); "Self-Portrait" (1773; Royal Academy, London); "St. John in the Wilderness" (1776; Wallace Collection, London); "The Infant Samuel" (1776; Tate Gallery, London); "Lady Bamfylde" (1776–77; Tate Gallery, London); "The Society of Dilettanti I & II" (1777–79; Society of Dilettanti, London): "Lady Elizabeth Delmé and Children" (1777; National Gallery of Art, Washington, D.C.): "John Musters" (1777–80; Museum of Fine Arts, Boston); "Colonel George Cousssmaker" (1782; Metropolitan Museum of Art, New York); "Sarah Siddons As the Tragic Muse" (1784; Huntington Art Gallery, San Marino, California); "Heads of Angels" (1787; Tate Gallery, London); "Selina, Lady Skipwith" (1787; Frick Collection, New York); "Lady Gertrude Fitzpatrick As 'Sylvia' " (1787; Museum of Fine Arts, Boston); "Lord Hcathfield, Governor of Gibraltar" (1788; National Gallery, London); "The Age of Innocence" (1788; Tate Gallery, London); "Mrs. Abbingdon As 'Roxalana' " (1788; Peter Vaughan Collection, London).

BIBLIOGRAPHY. E.K. WATERHOUSE, *Reynolds* (1941), an indispensable, scholarly work with 300 plates; SIR JOSHUA REYNOLDS, *Discourses Delivered at the Royal Academy* (1769–91; reprinted as *Discourses on Art*, 1966); J.S. NORTHCOTE, *Memoirs of Sir Joshua Reynolds* (1813) and *Supplement* (1815), an account by an artist who both knew him and worked in his studio; A. CRAVES and W.V. CRONIN, *A History of the Works of Sir Joshua Reynolds*, 4 vol. (1899–1901), a scarce but fundamental sourcebook; DEREK HUDSON, *Sir Joshua Reynolds* (1958), the best biography for both the scholar and general reader; JOHN O. WOODWARD, *A Picture History of British Painting* (1962), reproduces 12 of Reynolds' finest paintings together with one admirable colour plate; MALCOLM CORMACK, *The Ledgers of Sir Joshua Reynolds*, vol. 42 of *The Walpole Society* (1971); *Catalogue of Reynolds Exhibition, Birmingham* (1961), very informative and well documented—rich in paintings from the 1750s, the period in which it could be argued that the artist did his finest work.

(J.Wo.)

Reza Shah Pahlavi of Iran

An officer of the Iranian Army, who rose from the rank of private to that of general and finally became the sovereign of Iran, Reza Shah Pahlavi (name at birth Reza Khan) is regarded as having begun the regeneration of his country. He was officially granted the title of the "the Great."

Keystone

Reza Shah Pahlavi.

Reza Khan was born in 1878 in the village of Alasht in the Savadkuh district of Mazanderan province. He was of a family of chiefs of a clan named Pahlevan. After the death of his father, Col. Abbas Ali Khan, Reza's mother took him to Teheran, where he eventually enlisted as a private in an Iranian military unit under Russian instructors. Tall and powerfully built, the young soldier, from the beginning, showed an uncommonly strong will, remarkable intelligence, and a capacity for leadership. He was highly regarded by his seniors.

After centuries of misrule by its former rulers and the ravages of the war waged by foreign belligerents on its soil from 1914 to 1919, Iran in 1921 was prostrate, ruined, and on the verge of disintegration. The last of the *shāhs* of the Qājār dynasty, Ahmad Shah, was young and incompetent, and the Cabinet was weak and corrupt. Patriotic and nationalist elements had long been outraged at the domination of Iran by foreign powers, especially Great Britain and Russia, both of which had strong commercial and strategic interest in the country. This situation led Reza Khan to decide on an attempt at putting an end to the chaos by taking over power and forming a strong government, bolstered by an effective and disciplined military force. He contacted some young, progressive elements and on February 21, 1921, occupied Teheran at the head of 1,200 men. A young journalist, Sayyid Zia od-Din Tabatab'ai, became prime minister, while Reza Khan took command of all the military forces and was appointed minister of war a few weeks after.

Coup of February 1921

Reza Khan cherished the idea of regenerating the Iranian nation and leading it on the path of progress. Many had imagined that Reza Khan, whom they took to be an unsophisticated regimental officer, would be content with a high-sounding title and a sword of honour given by the Shah. But he was not about to step aside to allow a mixed group of inexperienced though sincere idealists and foreign-influenced opportunists to rule the country. His progress toward supreme power was extraordinarily rapid. Of a forbidding appearance, he talked very little and never revealed his intentions. Displaying great political talent against his opponents, he divided and weakened them. He also understood that to reach his ultimate objective he had to have complete control over a military force and that. that required money. Able to levy some taxes, he built up the army with the proceeds and then used the army to collect more taxes, until finally he had gained control over the entire country. As war minister, he was the real power behind several prime ministers in succession until 1923, when he became prime minister himself.

The sovereign, Ahmad Shah Qājār, was ill and undergoing a lengthy cure in Europe. In spite of the entreaties of Reza Khan and the speaker of the Majles (Iranian parliament), the Shah refused to return to Iran. Reza Khan then considered proclaiming a republic but was dissuaded by the strong opposition to the idea by the majority of the people. In 1925 the Majles deposed the absentee monarch, and a constituent assembly elected Reza Khan as *shdh*, vesting sovereignty *in* the new Pahlavi dynasty.

Policies as *shāh*

After his coronation in April 1926, Reza Shah continued the radical reforms he had embarked on while prime minister. He broke the power of the tribes, which had been a turbulent element in the nation, disarming and partly settling them. In 1928 he put an end to the one-sided agreements and treaties with foreign powers, abolishing all special privileges. He built the Trans-Iranian Railway and started branch lines toward the principal cities (1927–38). He emancipated women and required them to discard their veils (1935). He took control of the country's finances and communications, which up to then had been virtually in foreign hands. He built roads, schools, and hospitals and opened the first university (1934). His measures were directed at the same time toward the democratization of the country and its emancipation from foreign interference.

His foreign policy, which had consisted essentially of playing the Soviet Union off against Great Britain, failed when those two powers joined together in 1941 to fight the Germans. In order to be able to supply the Soviet forces with war material through Iran, the two allies jointly occupied the country in August 1941.

Reza Shah then decided to abdicate, to allow his son and heir, Mohammed Reza Shah, to adopt a policy appropriate to the new situation, and to preserve his dynasty. He wanted to go to Canada, but the British government sent him first to Mauritius and then to Johannesburg, South Africa, where he died in July 1944.

BIBLIOGRAPHY. H.I.M. MOHAMMED REZA SHAH PAHLAVI, *Mission for My Country* (1961), the reigning sovereign's interpretation of his father's role in the evolution of Iran; P. AVERY, *Modern Iran* (1965), the rise of Reza Shah and his work as seen by an English intellectual trying to be unbiassed; H. ARFA, *Under Five Shahs* (1964), contains little known facts about the Shah and his policy; R.W. COTTAM, *Nationalism in Iran* (1964), includes many points of interest; I.M. UPTON, *The History of Modern Iran: An Interpretation* (1960), an interesting synthesis; D.N. WILBER, *Contemporary Iran* (1963), facts as seen by a Westerner.

(H.Ar.)

Rhamnales

The Rhamnales, or buckthorn order, is an order of flowering plants containing about 1,550 species in three families. General interest in this group rests primarily on the cultivated grape and the fruits known as jujubes. Both products have been consumed since prehistoric times in the Mediterranean countries. Tradition has it that Jesus' crown of thorns was fashioned from jujube twigs.

Bitropical distribution of Rhamnales plants

Scientific interest in the group is especially piqued by the number of genera (but no species) native to both Old World and New World tropics or subtropics. This kind of distribution is by no means rare, but the Rhamnaceae (buckthorn) and Vitaceae (grape) families are among the dozen families of flowering plants with the highest proportion of genera to be found in both tropical regions. Distributions such as these have yet to be satisfactorily explained, but at least they seem to give some evidence of great antiquity, perhaps 150,000,000 years or more according to some writers. It has been conjectured that such families probably arose before Cretaceous times (*i.e.*, more than 136,000,000 years ago), but there is no fossil evidence either to support or to undermine this opinion. In contrast to the Rhamnaceae and Vitaceae families, which are widely distributed in temperate and tropical regions, the Leeaceae family, comprising a single genus of about 100 species, is restricted to tropical areas from southern Asia to Africa.

General features. *Size range and diversity of structure.* Trees, shrubs, and woody vines predominate in the buckthorn order, herbaceous plants being extremely rare. Climbing plants are common. Perhaps the most bizarre forms encountered are the nearly leafless shrubs of *Colletia paradoxa* (Rhamnaceae family) of South America, whose short, opposite green twigs are modified to thick, laterally flattened thorns; this species is widely planted as a curiosity. The one notable tendency in the Rhamnaceae family is that of xeromorphism, meaning adaptation to a dry climate, which encompasses a syndrome of adaptations including reduced size of leaves, crowding of leaves, shortening of some branch axes, thorniness or spininess; and a low, shrubby, intricately branched habit. In the Vitaceae family the most notable tendency is that toward trailing and twining, in general a vine-like habit. There are no true aquatics, epiphytes, or parasites in the order.

Economic uses. Two important members of the buckthorn order that furnish edible products are the grape and the jujube. The grape (genus *Vitis,* family Vitaceae) has been a source of fresh and dried fruit and has been used in wine making in Mediterranean lands since prehistoric times. The jujubes (genus *Ziziphus,* family Rhamnaceae), spiny shrubs that abound in arid areas from the Mediterranean region eastward, also furnish fresh and dried fruit and their culture probably dates as far back as that of grape culture. Other species furnishing edible products include *Hovenia dulcis* (Rhamnaceae), widely grown in China and neighbouring countries; the stalks of the flower clusters are succulent and raisin-like. The fruits of species of *Sageretia* (Rhamnaceae) are eaten fresh in parts of India and China. The leaves of *Sageretia thea* furnish a substitute for tea in southern China.

Green and yellow dyes are obtained from the fruits of various Old World species of *Rhamnus* (Rhamnaceae). Two species of *Rhamnus* furnish the famous "Chinese green" or lokao dye, which is the only natural plant dye giving a green colour in cotton.

Laxatives are obtained from the berries and bark of species of *Rhamnus,* including the famous cascara sagrada from the North American *Rhamnus purshiana.*

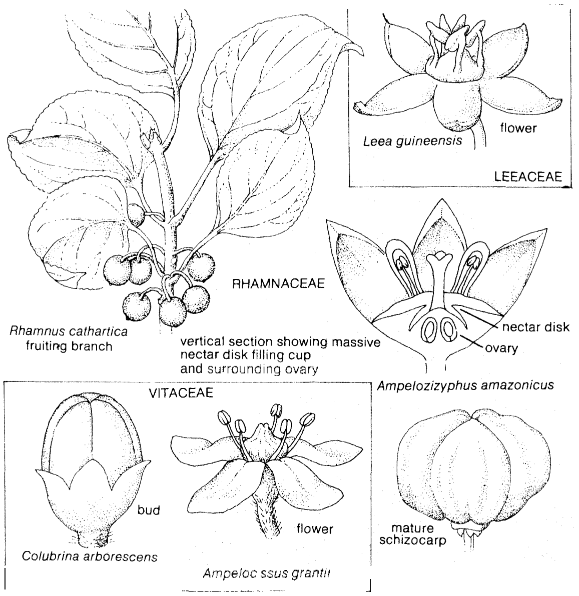
The fruits of *Colubrina asiatica* (Rhamnaceae) of tropical shores and of the Mexican *Ziziphus amole* are used as soap.

Ornamental plants, other than shade trees, are rare in this order, owing to the usually small size and scattering of the flowers. But a number of species of the North and Central American genus *Ceanothus* (Rharnnaceae) are cultivated for their showy flower clusters and are sometimes called "wild lilac" or "French lilac." The Virginia creeper *(Parthenocissus quinquefolia,* family Vitaceae) is widely planted as a wall and ground cover.

Wood products

A number of kinds of plants from the Rhamnaceae family furnish useful timber. Perhaps the most important economically is the tropical African *Maesopsis eminii,* which is now widely planted in Indonesia and Malaya as well as Africa; the lumber is used in building houses and boats. The wood of the genus *Ziziphus* is often hard and dense. Among the Old World species the wood of *Ziziphus mauritiana* is used for houses and for furniture; that of *Z. spina-christi* and of *Z. jujuba* for cabinetmaking; of *Z. xylopyra* and *Z. mucronata* for making wagons, agricultural implements, etc. Two Jamaican species, *Z. sarcomphalus* and *Z. chloroxylon* yield good wood, the former for construction and the latter a hard, heavy, elastic wood excellent for all purposes. The wood of *Hovenia dulcis* is valuable for furniture and musical instruments. *Pomaderris apetala* of southern Australia is especially used in cooperage. Some West Indian species have such dense, strong wood that it is often called "ironwood" in the trade—for example, *Reynosia latifolia* and *Krugiodendron ferreum;* the latter, air-dried, has a specific gravity of about 1.42 and is, thus, one of the densest woods known. The wood of *Colubrina arborescens* is used in the West Indies; that of the related *Colubrina oppositifolia* is so dense and durable that it was used as money by the ancient Hawaiians. *Rhamnus cathartica* and *R. frangula* of Europe furnish limited quantities of high-quality wood for various small objects, lathework, and furniture, as do *Berchemia zeyheri* of Africa and the North American *Condalia hookeri.*

Drawing by M Pahl


Leea guineensis
flower
LEEACEAE
RHAMNACEAE
Rhamnus cathartica fruiting branch
vertical section showing massive nectar disk filling cup and surrounding ovary
nectar disk
ovary
Ampelozizyphus amazonicus
VITACEAE
bud
flower
mature schizocarp
Colubrina arborescens
Ampeloc ssus grantii

Representative plant structures of the order Rhamnales.

Natural history. Little is known of the natural history of the Rhamnales order. Of the few species that have been intensively studied in culture, it can only be said that their vital success seems to be related, at least in part, to a lack of specialization. Seed germination is rarely checked, but in those species widely cultivated it seems to be prompt and easily achieved. Seed dispersal is usually by means of animals attracted to the fruits, although some genera have winged fruits adapted to wind dispersal. Reproduction is primarily by sexual means, although root sprouting is not uncommon. Pollination, in spite of the drabness and small size of the flowers, is accomplished by insects, which apparently are attracted to the nectar of the flowers. Extraordinarily little is known of the identity of the pollinators. Various small wasps, flies, bugs, bees, beetles, butterflies, and moths have been observed to visit the flowers of various species of the Rhamnaceae and Vitaceae families, but the extent to which any one species accomplishes pollination is unknown. In contrast to the situation in some groups, birds, ants, bats, and wind and water currents apparently are never involved in the pollination process in the buckthorn order. Special soil characteristics or other environmental requirements, such as a dependency upon fires, have not been found in the Rhamnales order. In fact, the members of the order are apparently so diverse as to their habitat requirements or at least their preferences that no cogent generalization is possible on this subject, though it can be said that, on the average, tropical forested areas are favoured. Many members of the order occur in forests and many in low forests or scrubby vegetation or desertic scrub. Few occur in grasslands. Certain spiny jnjubes abound in some desertic and semidesertic scrub formations in the Old and New Worlds, but apparently nowhere does any member of the buckthorn order so dominate the vegetation that species of other groups are essentially excluded. While each species of the order seems to have a narrower range of habitat tolerances and preferences, the order, as a whole, ranges from exceedingly arid areas to rain forests and from sea level to elevations and latitudes near the perpetual snow line. Perhaps the most specialized habitat requirements are exhibited by *Colubrina* asiatica, which is restricted to tropical beaches. The seeds of this species are known to float unharmed in seawater for long intervals, and this apparently accounts for its dispersal to numerous islands scattered in a vast expanse of the Pacific Ocean.

Wide range of habitats

Form and function. Members of the buckthorn order share the distinction that the stamens, which are the male flower parts that produce the pollen, are exactly as numerous as the flower petals and, even more important, are attached opposite (just toward the floral axis from) the petal attachments. The number and placement of stamens as found in the buckthorn order, known as obstemony, is a relatively rare condition among flowering plants. This is another reason for the considerable botanical interest in these plants. In most groups in which the stamens are just as numerous as the petals, they are attached alternate to the petals; *i.e.*, nearly between them. Other characteristics that these plants share include a disk (a ring-shaped structure of several distinct masses of tissue) that is located toward the central axis from the stamen attachments (intrastaminal) and usually produces nectar. The ovules (immature seeds) are anatropous (*i.e.*, the micropyle or depression through which the pollen tube bearing the male sexual cells enters for fertilization is at the same end of the ovule as the stalk by which the ovule is attached), bitegmic (the micropyle is defined by a double flange of seed-wall tissue), and erect (they are attached at the bottom of the seed cavities of the ovary). Each of the cavities of the ovary has only one or two ovules.

Evolution. Fossil record. Numerous fossil leaf impressions and a lesser number of fossil woods, fruits, and seeds have been referred to this order. The oldest such remains are from Cretaceous Period beds (about 100,000,000 years old), but the most numerous are from younger beds, especially those of the Miocene and Pliocene Epochs (about 10,000,000 to 20,000,000 years old). Unfortunately, there is no single character or set of characters of the foliage that, beyond a shadow of a doubt, distinguishes leaves of this order from those of some other orders. None of the leaves has been found in organic connection with flowers or fruits. Thus, the assignments of fossils to the Rhamnaceae or Vitaceae families range in probability of correctness from high to infinitesimal. In general, the oldest fossils are of the most dubious identity, but even some of the most recent fossils are not certainly referred to the Rhamnales order. Numerous Cretaceous and early Tertiary leaf impressions (about 30,000,000 to 100,000,000 years old) previously referred to the genus *Ziziphus* have more recently been referred to *Cercidiphyllum*, which is not a close relative of the Rhamnales order. This illustrates the pitfalls of hasty assignment of fossil leaf impressions to modern kinds of plants. Fossils that are assignable to this order with a higher degree of probability have been found in Oligocene or younger beds (about 30,000,000 years old or less). By Miocene times (about 20,000,000 years ago), apparently both the Rhamnaceae and the Vitaceae families were large and diverse. This seems to indicate a much greater antiquity for both families, but it is impossible to trace them further back through the fossil record with any certainty. Evaluation and correlation of the fossil wood, fossil pollen, and fossil leaf evidence is not yet in a sufficiently advanced state to permit the presentation of a coherent evolutionary story.

Phylogeny. The idea that the buckthorns and the grapes are closely related dates at least as far back as 1671 and continues to be accepted by present-day botanists. The gathering consensus that the buckthorn and grape families have an especially close affinity by now amounts to a botanical act of faith. There have been important dissenters but relatively few. The placing together of the buchthorn and grape families (the latter including the genus Leea as a subfamily) as the narrowly construed order Rhamnales dates from 1892. Previously, these families had been associated at the ordinal level with families now placed variously in the Celastrales, Sapindales, Rutales, and Cornales orders. The assumption implicit in the close modern taxonomic placement of the families Rhamnaceae, Vitaceae, and Leeaceae is that they represent the products of a single ancestral population that was not the same as those from which plants of neighbouring taxonomic groups arose, though it may in turn, further back in time, have sprung from a population from which plants of neighbouring taxonomic groups also arose. This is thought by many to be a reasonable assumption, and, if it is accepted, then the relationships of such presumably closely related groups as the Sapindales and Celastrales orders can only be described as collateral to the Rhamnales order.

Relation of the buckthorn and grape families

Classification. *Distinguishing taxonomic* features. The following characters are especially useful to delineate groups within the order: the degree of lobing of the leaves; the presence or absence of a floral cup; the degree of bud closure by the calyx (sepals) or corolla (petals); the number of ovules per ovary cavity and the dorsal, lateral, or ventral placement of the raphe (a ridge along the ovule); the formation of the micropyle either by the inner or outer integument; the nature of the fruit, and, especially, whether the endocarp is soft or indurated; and the relative size of the embryo to the seed.

The classification below is generally acceptable as a reflection of current taxonomic thought (see, however, the remarks at the end). None of the families is of doubtful affinity except insofar as the question of the mutual affinity of the two suborders is concerned.

Annotated classification.

ORDER RHAMNALES

Giant trees, large woody vines or lianas to subshrubs or weak vines, rarely herbs (annual herbs in only 1 species, the Guatemalan *Crumenaria steyermarkii* of the Rhamnaceae family); tendrils often present. Leaves simple or palmately compound or rarely twice or thrice pinnately compound; stipules usually present. Inflorescence (the arrangement of the flowers) basically cymose, but the cymes commonly aggregated into 2nd-order inflorescences of various shapes, and these usually leafless or nearly so, occasionally the cymes so reduced that the flowers are merely fasciculate (clustered) or even solitary. Flowers small, rarely more than 5 mm in diameter, rarely brightly coloured (exceptions in *Ceanothus* and Leea), radially symmetrical, 4- or 5-merous (parted, lobed), hypogynous, perigynous, or epigynous (*i.e.*, the floral parts arise at the base of, from a cuplike organ around, or from the top of the ovary, respectively), perfect, polygamous, or dioecious (*i.e.*, flowers have both sexes in 1 flow-er, or on 1 plant there are separate flowers of either sex, or there are separate plants containing flowers all of 1 sex or the other). Sepals free or less commonly joined in the lower part and never overlapping even in bud. Petals rarely any other colour than white, pale yellow, or cream colour, as numerous as sepals and alternate with them, free or uncommonly joined to each other along the margins in the distal moiety (in which case the corolla falls as a unit just before anthesis) or (in *Leea*) forming a short tubular corolla, or petals rarely absent. Stamens as numerous as sepals and alternate with them; anthers minute, usually opening toward the central axis of the flower (introrse), filaments thin, usually inbowed in bud and rarely exceeding the petals in length; pollen production not copious, the pollen binucleate (rarely trinucleate) when shed, usually tricolporate (having 3 equidistant germinal grooves) and angulapertuate, the outer wall being usually, perhaps always, reticulate. Intrastaminal disk present, usually producing nectar, either lining or filling the cup in perigynous flowers or usually lobed and closely associated with the ovary or along with the filaments forming a tube. Gynoecium (female flower parts) of a single pistil (ovary, style, and stigma, the female structure) of 2 or 3, less commonly 4, rarely up to 8 carpels; locules (chambers) as numerous as carpels or fewer by abortion; each locule with 1 or 2 basal, erect, anatropous, bitegmic ovules; the dicotyledonous embryo straight, flat or nearly so, embedded in abundant, firm to corneous endosperm; fruit various, never very large, averaging from a few millimetres to a few centimetres in maximum dimensions. Two suborders, 3 families, about 59 genera and 1,550 species in tropical and temperate regions.

Suborder Rhamnineae

Leaves simple and unlobed; flowers with a floral cup at the rim of which are borne the sepals, petals, and stamens, this floral cup either free (and the flower then perigynous) or its lower part joined to the gynoecium and the flower then epigynous; calyx (sepals) valvate and enclosing the corolla (petals) in bud; petals never exposed in bud, never valvate, smaller than the sepals, occasionally absent; disc never deeply lobed and always intimately associated with or lining the floral cup; anthers usually introrse, rarely extrorse (opening outwards, away from floral axis); ovule usually 1 per locule with a dorsal or lateral raphe; micropyle usually formed by the outer integument; fruit a schizocarpous (splitting apart into 1-seeded units) capsule or dry or fleshy drupe (a "stone-seeded" fruit), with the endocarp always indurated or at least cartilaginous, never a berry; embryo of seeds usually large, the endosperm very rarely lobed or ruminate.

Family Rhamnaceae

Trees, shrubs, seldom vines, rarely herbs, with monopodial (having a main axis) growth and with the subordinal characters given above; distributed in all temperate and tropical regions. About 45 genera and 850 species.

Suborder Vitineae

Leaves simple and deeply lobed or commonly palmately or pinnately compound; flowers hypogynous (floral cup absent); calyx reduced, never valvate and never enclosing the bud; petals never absent, much larger than the calyx, valvate, and closed over the stamens, disc, and pistil in bud; disc usually deeply lobed (the lobes alternate with the petals) and intimately associated with the basal part of the ovary; anthers introrse or extrorse; ovules 1 or 2 in each locule, with a ventral raphe; micropyle formed by the inner integument; fruit always a berry, usually soft, seldom firm, rarely hard, and with the endocarp of the same texture as the rest; embryo of seeds usually comparatively small and embedded in the endosperm near the rnicropyle; endosperm lobed or in a few members even ruminate.

Family Vitaceae

Woody vines or lianas, seldom shrubby or treelike; tendrils usually present; growth sympodial (with an apparent main axis not produced from a terminal bud but from one fork of a dichotomy, the other being suppressed); stipules present; petioles never winged in the lower part; leaves simple and deeply lobed or commonly palmately compound, rarely pinnate; inflorescences often "lateral" (*i.e.*, terminal on the short axes that appear lateral because of the sympodial habit); petals either free and persistent or else joined marginally in their upper moieties in which case the corolla falls as a unit just prior to anthesis; filaments free; lobes of disc never equalling filaments in length; anthers introrse; locules of ovary 2 (rarely 3), each with 2 ovules; endosperm lobed, rarely ruminate. About 12 genera and 600 species, widely distributed in temperate and tropical regions.

Family Leeaceae

Shrubs or subshrubs, rarely small trees, never lianas; tendrils absent; growth monopodial; stipules absent; petioles winged in their lower part; leaves simple or twice or thrice pinnately compound; inflorescences always terminal; petals joined marginally in their lower moieties, the corolla persistent at least until after pollination; a tube or corona formed partially by the lower parts of the filaments and apparently partially by disc-lobe tissue is united to the corolla tube at base, becoming free and prolonged above, approximately equalling the anthers in elevation, and adaxially bearing an annular flange; upper parts of filaments free; anthers at first appearing introrse because of inbowing of filaments over the corona, but later erect and definitely extrorse; locules of ovary 3 to 8, rarely 2, each with 1 ovule; endosperm ruminate. One genus (Leea) and about 100 species, in tropical areas from Africa to southern Asia.

Critical appraisal. The internal coherence and assumed origin from a single ancestral population of each of the three families circumscribed above is unquestioned. Also, most authorities consider the Vitaceae and Leeaceae families to be naturally and closely associated, although it has been recently suggested that the family Leeaceae has as much in common with the Rhamnaceae family as it has with the Vitaceae family. The Leeaceae family, however, is probably a specialized derivative of the Vitaceae.

The internal coherence of the Rhamnales order as a whole, however, is not such as to convince all taxonomists that their close association at the ordinal level, and their segregation from other orders such as Sapindales, are required. Nor will the careful phylogenist find it necessary to postulate a common origin short of that which is common to various other orders such as Sapindales, Rutales, etc. Erect ovules and tricolporate pollen (pollen with three equidistant germinal grooves) are scarcely rare in such other orders, and obstemony could easily have arisen in two or more diplostemonous ancestors (*i.e.*, ancestors with two series of stamens, one series opposite, and one series alternate with, the petals).

The true measure of the interrelatedness of Rhamnaceae, Vitaceae, and Leeaceae families and the various groups with which they have been associated in the last 200 years is perhaps indicated by the gathering consensus that all these plants were derived ultimately from plants similar to some of the more primitive members of the Rosidae subclass. The character combinations by which families of the loosely coherent Celastrales, Sapindales, and Rutales orders are distinguished from one another are often of about the same magnitude and diversity as those which distinguish the Rhamnaceae family from the Vitaceae and from the Leeaceae families.

In fact, the suggestion is not at all far fetched that the ancestors of the Rhamnaceae may be found in the vicinity of the Rosales order, whereas those of the suborder Vitineae (including Vitaceae and Leeaceae families) may be found, as suggested by some authorities, in the neighbourhood of the ancestors of the Sapindales and Celastrales orders, that is to say, near Saxifragales order.

If this suggestion is accepted as having greater probability of correctness than the widespread assumption of a rather approximate common ancestry of the families here included in the Rhamnales order, then the taxonomic action indicated is the dismemberment of the traditional Rhamnales group. The Rhamnaceae family would then be referred to the Rosales order, and the suborder Vitineae either to the Saxifragales or to the Sapindales or, more likely, placed near these two orders as a distinct order, Vitales.

BIBLIOGRAPHY. HERMANN HARMS, "Reihe Rhamnales: Geschichtliche Entwicklung der Ansichten über die Umgrenzung der Reihe," in A. ENGLER and K. PRANTL (eds.), Die *Natürlichen Pflanzenfamilien*, 2nd ed., 20d: 1–6 (1953), the classical summary of the historical underpinnings of the modern taxonomic circumscription of the order; KARL SUESSENGUTH, "Rhamnaceae," "Vitaceae," and "Leeaceae," ibid., 2nd ed. 20d:7–390 (1953), the basic modern reference to these families with an extensive bibliography indispensible for study of these groups; GEORGE K. BRIZICKY, "The Genera of Rhamnaceae in the Southeastern United States," *J.* Arnold Arbor., 45:439–463 (1964), and "The Genera of Vitaceae in the Southeastern United States," J. Arnold Arbor., 46:48–67 (1965), valuable sources for the entire order even though they give details only for the genera occurring in southeastern United States.

(M.C.J.)

Rheinland-Pfalz

Rheinland-Pfalz, in English Rhineland-Palatinate, the sixth largest *Land*, or state, of the Federal Republic of Germany, has had a long history of division and possession by foreign powers. The modern state was created only after World War II and is therefore still reworking its economy. industry has overtaken agriculture as the primary wage earner, and trade links with the rest of West Germany have been forged and strengthened.

The state is bordered by the states of Nordrhein-Westfalen to the north, Hessen and Baden-Württemberg to the east, and Saarland to the southwest, and by the countries of France, Luxembourg, and Belgium to the south and west. It covers 7,659 square miles (19,838 square kilometres), an area about the size of Israel. The population of 3,645,000 represents 6 percent of the inhabitants of West Germany. its capital is Mainz, one of the oldest German cities.

The border of Rome

History. The oldest archaeological remains of the region are tools from the Stone Age that are at leas? 100,000, and may be as much as 300,000, years old. During the Neolithic Period between 3000 and 1800 BC, large areas along the Rhine were settled by Celts and Germanic peoples. Incorporated into the Roman Empire in the 1st century BC, the Rhineland formed the northeastern border region of the Gallic Provinces for about 500 years. The cities of Moguntiacum (Mainz), Augusta Treverorum (Trier), and Colonia Agrippinensis (Cologne) were founded during the Roman period. From the 5th to the 9th centuries AD, the Rhineland belonged to the frankish kingdom of the Merovingians and later the Carolingians. In 843 the kingdom was divided in half, and the Rhineland became the western border region of the East Frankish or German kingdom. During this period, the region was fractured into a large number of small independent states with temporal and religious governments. The most powerful of these states were the archdioceses of Trier, Cologne, and Mainz and the Rhenish Palatinate, which was ruled from the 13th century by the Bavarian Wittelsbach dynasty.

Incorporation into the French Rhine

The Reformation of the 16th and 17th centuries saw further territorial division that originated in the religious conflicts of Protestantism, Catholicism, and Calvinism and led to the Thirty Years' War (1618–48). Foreign nations—particularly Bavaria, Spain, Austria, Sweden, and France—determined the political development of the Rhineland. In the 17th and 18th centuries the Palatinate had close political and cultural ties with France. According to the truce of Campo Formio on October 17, 1797, the lands to the west of the Rhine were incorporated into the French territories, the individual states were dissolved, and religious possessions were secularized. In 1815 the Rhineland became a part of the newly founded German Confederation, and the region was divided by the Congress of Vienna of 1815 among Prussia, Bavaria, Hesse-Darmstadt, and Hesse-Nassau. After World War II, the region was again divided by order of the American, French, and English military governments, and the Rhineland-Pfalz was created.

The landscape. ***Relief.*** The northern portion of the state consists of woodlands and cultivated fields that are crossed by deeply eroded river valleys. Part of the Rhine Massif, its average elevation lies between 1,300 and 1,900 feet. The Rhine River crosses the region from the southwest to the northeast and receives the Moselle from the southwest and the Lahn from the northeast. The rivers effectively create four distinct regions—the Hunsriick, Eifel, Westerwald, and Taunus.

The southwestern region, bordered on the north by the Nahe River, is broken by the Saar-Nahe Mountains and the escarpments of the Pfalzer Wald, a national forest. Open cultivated areas of heavy clay and marl (loose clay containing calcium carbonate) soils alternate with large wooded areas of porphyry (dark-red, mineral-bearing rocks) and sandstone. In contrast, the southeast contains the treeless Rhein-Hesse Plateau and the Rhine Rivet Valley. The plateau is composed of limestone that is covered by loess (a brownish mixture of clay, silt, and sand deposited by the wind), while the valley contains fertile alluvial soils.

Climate. At Mainz the average annual temperature is 50° F (10" C) and the average yearly rainfall is 20 inches. Regional variations occur, however: the northwestern part of the Eifel has an average annual temperature of 42" F (5.6" C) and an average yearly rainfall of 31 inches.

Settlement. The settlement pattern is characterized by very old, irregularly structured villages with mostly consolidated fields in the regions of favourable soil and climate, and by villages and hamlets in the mountains. In the mid-20th century, the function of the rural settlements changed from that of purely agrarian communities to dormitory towns containing a large number of commuters. The most important cities are Mainz, Ludwigshafen, Koblenz, Trier, Kaiserslautern, Worms, Pirmasens, and Neustadt an der Weinstrasse.

A changing population

The people. Although the majority of the population is of Frankish descent, the turbulent history of the Rhineland has produced a complex social structure that has also been influenced by Romanic (*i.e.*, French and Italian) origins. The latest influence has been that of the 403,000 refugees who have migrated from middle and east Germany since 1945 and comprise about 12 percent of the total population of 3,645,000. Over 56 percent of the people are Roman Catholic and about 42 percent are Protestant. The population has an annual rate of increase of 0.7 percent that is due to both natural increase and immigration. The immigrants are primarily foreign workers from Italy, France, Turkey, Spain, Yugoslavia, and Greece who enter Rheinland-Pfalz through the surrounding German states. The average population density of 480 persons per square mile is considerably lower than that for the entire nation; the greatest—over 800 persons per square mile—is in the large river valleys where the main cities are located, while the sparsely populated highlands support under 150 persons per square mile. There is a constant emigration from the rural areas to the urban regions, but this rural exodus is on a much smaller scale than in the rest of western Germany. In the early 1970s half the people lived in communities of under 5,000, and only 18 percent lived in cities of over 100,000.

Economy. Rheinland-Pfalz is one of the poorest of West German states. In the late 1960s, however, the rate of increase of the economy was higher than that of the rest of the country. The total gross national product of 30,000,000,000 Deutsche Marks (DM. 3.22 = $1 U.S.; DM 8.37 = £1 sterling, on December 31, 1971) in 1969 derived from several sources—50 percent came from secondary industry, 28 percent from services, 16 percent from commerce and transportation, and 6 percent from agriculture.

Agriculture. In the early 1970s about 17 percent of the employed population was engaged in agriculture. The number of farm workers and of small farms was constantly decreasing, while the number of large farms was growing steadily. In the most fertile agricultural regions of the Neuwied Basin, Bitburg, Rhinehessen, and the eastern Palatinate, potatoes, cereals, and sugar beets are the primary crops. In the less fertile highlands, however, stock farming is more important. The state is known for the many specialized crops of the river valleys. Besides the growing of fruit and tobacco, viticulture occupies a predominate place in the agriculture of the state, and the famous vineyards along the Rhine, Moselle, and Nahe rivers produce over 75 percent of West Germany's wines.

New industries

Industry. The pumice quarries of the Neuwied Basin represent the state's only important natural resource. The largest activities are the chemical industry in Ludwigshafen, Ingelheim, and Mainz and the engineering industry in Bad Kreuznach, Frankenthal, and Kaiserslautern. Over one-third of the nation's shoes are produced in Pirmasens, and 88 percent of West Germany's trade in precious stones is carried on in the town of Idar-Oberstein. An important tourist industry has developed in the middle Rhiie area.

In 1970 over 40 percent of the gross national product, or GNP, derived from commerce and services, and about 50 percent of the employed population was thus engaged. The centres of industry and administration are also the most important places for such activities.

Transportation. The most important transportation routes are the navigable waterways of the Rhine and the Moselle. In order to meet increasing demands, the Rhine is to be deepened in the 1970s. All of the main railways follow the Rhine Valley, and most of the less important junctions are to be discontinued. The primary highways also run from north to south. Roads, existing or planned, from Ludwigshafen to Saarbrucken, from Koblenz to Trier, and the Hunsriick Highway all join the primary routes.

Administration and social conditions. The state is divided into the three *Regierungsbezirke* (administrative districts) of Koblenz, Trier, and Rheinhessen-Pfalz. The legislative body, the Landtag, is composed of 100 deputies elected by the people every four years. The Landtag appoints a prime minister, who in turn appoints seven ministers. Under the state's judicial system, civil and criminal cases are tried by the provincial court of appeal and the county courts. There are also special courts for social, financial, and administrative problems, while the constitutional court deals with the legality of laws and administrative measures. Education is also subject to the sovereignty of the state. All children from six years of age are obliged to attend grammar school for four years. The student may then choose to attend five more years of *Hauptschule*, or common school; six years of *Realschule*, which qualifies him to enter technical schools; or nine years of *Gymnasium*, where a classical and scientific education are offered. Only from this school is entrance to the University possible. There are universities at Mainz (Johannes Gutenberg-Universitat) and Trier and Kaiserslautern (Universitat Trier-Kaiserslautern). Teachers for the primary schools are educated in the departments of special education universities. There are several technical schools, a university for the science of administration at Speyer (Hochschule fur Verwaltungswissenschaften), and academies for arts and sports that are a part of the university in Mainz. There are about 240 public and private hospitals in the state.

Cathedrals and festivals

Culture. The state's rich cultural tradition is reflected in the cathedrals at Speyer, Worms, and Mainz. The many museums include the Historische Museum der Pfalz ("Historical Museum of the Palatinate") in Speyer, the Rheinisches Landesmuseum ("State Museum of the Rhine") in Trier, and the Romisch-Germanisches Zentralmuseum ("Roman-Germanic Central Museum") and the World Museum of the Art of Printing (Gutenberg-Museum) in Mainz. National traditional festivals take place annually, highlighted by the pre-Lenten carnival in Mainz. There are theatres in Mainz, Trier, Koblenz, and Kaiserslautern. The academy of sciences and the Rheinisch-Naturforschende Gesellschaft ("Natural Research Society of the Rhine") promote scientific research. The large cities contain printing offices for 26 daily papers, with a circulation of 630,000. Mainz is the seat of a studio of the Southwestern German Broadcasting and the Second German Television networks.

Prospects for the future. The future prosperity of the Rheinland-Pfalz depends on its growing strength as an industrial state. New businesses and manufacturers are needed to absorb the influx of foreign immigrants to the cities, while the growth of commercial farming prompts internal movements to the urban centres. Improved communications and transportation routes with the rest of the country will aid in the future success of the economy. Although federal plans exist for another redivision of the state, it is apparent that the region is becoming a viable political and economic unit.

BIBLIOGRAPHY. W. GOTZ (ed.), *Rheinland Pfalz: Ursprung, Gestalt und Werden eines Landes* (1967), is a collection of several articles, dealing with the history, geography, politics, economics, and culture of the region; *Deutsche Planungsatlas*, vol. 8, *Rheinland-Pfalz* (1965), deals with the landscape, population, rural and industrial economy, transportation, administration, and culture; *Raumordnungsbericht* 1969 *der Landesregierung Rheinland-Pfalz* (1969), is a report of the structural development of the state between 1967 and 1969; while OBERSTE LANDESPLANUNGSBEHORDE (ed.), Landesentwicklungsprogramm *Rheinland-Pfalz* (1968), is a projection of future developments in the 1970s; L. PETRY (ed.), *Handbuch der historischen Stätten Deutschlands*, vol. 5, *Rheinland-Pfalz und Saarland* (1959), describes important historical sites; and *Luftbildatlas Rheinland-Pfalz* (1970) is *a* photographic collection of the state's varied landscape.

(O.Ka.)

Rhetoric

The term rhetoric has traditionally applied to the principles of training communicators—those seeking to persuade or inform others; in the 20th century it has undergone a shift of emphasis from the speaker or writer to the auditor or reader. This article will deal with the subject of rhetoric in both its traditional and its modern forms. For related information on the techniques of language in poetical contexts, see the article PROSODY. Rhetorical techniques are also discussed in the articles ARTS, CRITICISM OF THE; LITERARY CRITICISM; and LITERATURE, NONFICTIONAL PROSE.

Rhetoric in literature

THE NATURE AND SCOPE OF RHETORIC

Function of rhetoric

Traditional and modern rhetoric. From the traditional point of view, rhetoric might be limited to the insights and terms developed by rhetors, or rhetoricians, in the Classical period of ancient Greece, about the 5th century nc, to teach the art of public speaking to their fellow citizens in the Greek republics and, later, to the children of the wealthy under the Roman Empire. Because some sort of public performance was regarded as the highest reach of education proper, rhetoric as a discipline or as a principle of pedagogy and learning was at the centre of the educational process in western Europe for some 2,000 years—until well into the 19th century, as a matter of fact. *Institutio oratoria* (before AD 96; "The Training of an Orator"), by the ancient Roman rhetorician Quintilian, perhaps the most influential textbook on education ever written, was in fact a book about rhetoric. Inevitably, there were minor shifts of emphasis in so long a tradition, and for a long time even letter writing fell within the purview of rhetoric; but it has consistently maintained its emphasis upon creation, upon instructing those wishing to initiate communication with other men.

The mark of modern rhetoric, on the other hand, is its shift of focus to the auditor or reader. From the beginning, of course, literary criticism borrowed from rhetoric—stylistic terms such as antithesis and metaphor were invented by Classical rhetoricians. When language became a subject of sustained scholarly concern, as it has in the 20th century, it was inevitable that scholars would turn back to Classical theories of rhetoric for help. But modern rhetoric is far more than a collection of terms borrowed from Classical rhetoricians. The perspective from which it views a text is different from that of other disciplines, History, philosophy, literary criticism, and the social sciences are very apt to view a text as though it were a kind of map of the author's mind on a particular subject. The rhetorician, accustomed by his traditional discipline to look at communication from the communicator's point of view, regards the text as the embodiment of an intention, a design—not as a map. He knows that that intention in its formulation is profoundly affected by its audience. He knows, also, that the structure of a piece of discourse, the way its major parts fit together, is profoundly a result of its intention. A concern for audience, for intention, and for structure is, then, the mark of modern rhetoric. As a study or discipline, educational principle, or body of theory, modern rhetoric is as involved with the process of interpretation, or analysis, as it is with the process of creation, or genesis.

The situations of a message

In these two processes, the perspectives are similar. Rhetorical analysis is actually an analogue of traditional rhetorical genesis: both view a message through situations—the situation of the auditor or reader as well as the situation of the speaker or writer. Both view the message itself as compounded of elements of time and place, motivation and response. **A** text must reveal its context—that insistence automatically makes a rhetorician of the literary critic or interpreter and distinguishes his approach from the other kinds of verbal analysis. A group of literary critics that came into prominence around World War 11—the "New Critics"—insisted upon isolating, or abstracting, the literary text from the mind of its creator and from the milieu of its creation, but they found themselves unable to abstract it from the situation of its reader. Certain modern critics have joined with rhetoricians in denouncing the folly of all such attempts at abstraction. When modern man interprets any text—say a speech by Elizabeth I of England at Tilbury, Essex, or a play by the great Hindu poet of the 5th century, Kālidāsa—he must imaginatively re-create the original situation of that text as well as endeavour to understand those factors that condition his present understanding.

Moreover, all discourse now falls within the rhetorician's purview. The traditional rhetorician, even when his efforts are directed toward interpretation, is primarily concerned with oratory or with those forms of nonfictional prose that might have a didactic or persuasive intent. But the modern rhetorician identifies rhetoric more with critical perspective than with artistic product. In part, he justifies expanding his concerns into other literary provinces on the basis of a striking change that has occurred in thinking about the very nature of man's reason itself. Modern philosophers of the Existentialist and Phenomenologist schools, while seeking a way out of the oftentimes false dilemmas posed by traditional philosophical approaches, have strongly challenged the assumptions whereby such dualities as knowledge and opinion, persuasion and conviction, reason and emotion, rhetoric and poetry, and even rhetoric and philosophy have in the past been distinguished. The old line between the demonstrable and the probable has become blurred. According to these modern philosophers, man's basic method of judgment is argumentation, whether in dialogue with others or with a text, and the results are necessarily relative and temporal. Such modern philosophers again and again use legal battles in a courtroom as basic models of the process every man goes through, or should go through, in acquiring knowledge or opinion of any phenomenon. For some, philosophy and rhetoric have become conflated, with rhetoric itself being a further conflation of the subject matter Aristotle discusses not only in his *Rhetoric* but also in his *Topics,* which he had designed for dialectics, for disputation among experts. According to this view, a philosopher, like any lawyer or giver of knowledge, is engaged in a rhetorical transaction seeking to persuade through a dialogic process first himself and then, by means of his utterance, others. It is in something like this "argumentative" light that a rhetorically trained reader or auditor interprets all texts—philosophical, oratorical, poetical—and justifies their inclusion within the province of rhetoric.

Thus, having shifted its focus from the communicator to the communicant and having been profoundly influenced by the temper of modern intellectual movements, rhetoric has come to be understood less as a body of theory or as certain types of artificial techniques and more as an integral component of all human discourse. As a body of discursive theory, rhetoric has traditionally offered rules that are merely articulations of contemporary attitudes toward certain kinds of prose—perhaps that is the most charitable interpretation of the 17th-century English satirist Samuel Butler's charge that "all a rhetorician's rules/ But teach him how to name his tools." In this view rhetoric has tended to be identified with orations in which the specific intent to persuade is most obvious. But modem rhetoric is limited neither to the offering of rules nor to studying topical and transient products of controversy. Rather, having linked its traditional focus upon creation with a focus upon interpretation, modern rhetoric offers a perspective for discovering the suffusion of text and content inhering within any discourse. And for its twin tasks, analysis and genesis, it offers a methodology as well: the uncovering of those strategies whereby the interest, values, or emotions of an audience are engaged by any speaker or writer through his discourse. The perspective has been denoted with the term situation; the methodology, after the manner of certain modern philosophers, may be denoted by the term argumentation. It should be noted at the outset that one may study not only the intent, audience, and structure of a discursive act but also the shaping effects of the medium itself on both the communicator and the communicant. Those rhetorical instruments that potentially work upon an audience in a certain way, it must be assumed, produce somewhat analogous effects within the writer or speaker as well, directing and shaping his discourse.

Figures of speech

Elements of rhetoric. For the tasks imposed by the rhetorical approach some of the most important tools inherited from antiquity are the figures of speech: for example, the metaphor, or comparison between two ostensibly dissimilar phenomena, as in the famous comparison by the 17th-century English poet John Donne of his soul and his mistress's to the legs on a geometer's compass in his "A Valediction: Forbidding Mourning"; another is the allegory, the extended metaphor, as in John Bunyan's classic of English prose *Pilgrim's Progress* (1678, 1684), wherein man's method of earning Christian salvation is compared to a road on which he journeys, and the comparison is maintained to such an extent that it becomes the central structural principle of the entire work. Such figures may be said to pertain either to the texture of the discourse, the local colour or details, or to the structure, the shape of the total argument. Ancient rhetoricians made a functional distinction between trope (like metaphor, a textural effect) and scheme (like allegory, a structural principle). To the former category belong such figures as metaphor, simile (a comparison announced by "like" or "as"), personification (attributing human qualities to a nonhuman being or object), irony (a discrepancy between a speaker's literal statement and his attitude or intent), hyperbole (overstatement or exaggeration) or understatement, and metonymy (substituting one word for another which it suggests or to which it is in some way related—as part to whole, sometimes known as synecdoche). To the latter category belonged such figures as allegory, parallelism (constructing sentences or phrases that resemble one another syntactically), antithesis (combining opposites into one statement—"To be or not to be, that is the question"), congeries (an accumulation of statements or phrases that say essentially the same thing), apostrophe (a turning from one's immediate audience to address another, who may be present only in the imagination), enthymeme (a loosely syllogistic form of reasoning in which the speaker assumes that any missing premises will be supplied by the audience), *interrogatio* (the "rhetorical" question, which is posed for argumentative effect and requires no answer), and *gradatio* (a progressive advance from one statement to another until a climax is achieved). However, a certain slippage in the categories trope and scheme became inevitable, not simply because rhetoricians were inconsistent in their use of terms but because well-constructed discourse reflects a fusion of structure and texture. One is virtually indistinguishable from the other. Donne's compass comparison, for example, creates a texture that is not isolable from other effects in the poem; rather, it is consonant with a structural principle that makes the comparison both appropriate and coherent. Above all, a modern rhetorician would insist that the figures, like all elements of rhetoric, reflect and determine not only the conceptualizing processes of the speaker's mind but also an audience's potential response. For all these reasons figures of speech are crucial means of examining the transactional nature of discourse.

Rhetoric of or in a discourse. In making a rhetorical approach to various discursive acts, one may speak of the rhetoric *of* a discourse—say, Robert Browning's poem "My Last Duchess" (1842)—and mean by that the strategies whereby the poet communicated with his contemporaries, in this case the Victorians, or with modern man, his present readers; or one may speak of the rhetoric *in* a

The three voices of discourse

discourse and mean by that the strategies whereby the persona, the Duke of Ferrara who speak's Browning's poem in dramatic-monologue fashion, communicates with his audience in the poem, in this case an emissary from the father of Ferrara's next duchess. The two kinds of rhetoric are not necessarily discrete: in oratory or in lyric poetry, for example, the creator and his persona are assumed to be identical. To a degree Aristotle's distinction between the three voices of discourse still holds. A poet, according to Aristotle, speaks in his own voice in lyric poetry, in his own voice and through the voices of his characters in epic (or narrative), and only through the voices of his characters in drama. Thus, the speaker of oratory or of most nonfictional prose is similar to the lyric speaker, with less freedom than the latter either to universalize or to create imaginatively his own audience.

RHETORICAL TRADITIONS

Although knowledge of rhetorical traditions is essential to the modern student's work, it must be borne in mind that he is nonetheless divorced from those traditions in two important ways. First, there is an almost exclusive emphasis upon the speaker or writer in traditional rhetoric; and, second, there is an implicit belief that the truth can be detached from the forms of discourse and can be divided into the demonstrable and the probable. In both of these respects, modern rhetorical practice differs.

Ancient Greece and **Rome.** Since the time of Plato it has been conventional to posit a correlative if not causal relationship between rhetoric and democracy. Plato located the wellsprings of rhetoric in the founding of democracy at Syracuse in the 5th century BC. Exiles returning to Syracuse entered into litigation for the return of their lands from which they had been dispossessed by the overthrown despotic government. In the absence of written records, claims were settled in a newly founded democratic legal system. To help litigants improve their persuasiveness, certain teachers began to offer something like systematic instruction in rhetoric.

In this experience at Syracuse, certain identifiable characteristics become prototypal: the rhetor, or speaker, is a pleader; his discourse is argumentative; and his audience are participants in and judges of a controversy. Later, in Athens, these characteristics began to aggregate to themselves some serious intellectual issues.

Role of the Sophists in Athens

In Athens early teachers of rhetoric were known as Sophists. These men did not simply teach methods of argumentation; rather, they offered rhetoric as a central educational discipline and, like modern rhetoricians, insisted upon its usefulness in both analysis and genesis. With the growth of Athenian democracy and higher systematized education, the Sophists became very powerful and influential. Today the word sophistic refers to a shabby display of learning or to specious reasoning; it refers, consequently, to an image of the Sophists that resulted from the attacks upon them led by such reformers as Plato. The ideal rhetoric proposed by Socrates in Plato's dialogue the *Phaedrus,* however, is itself not unlike the ideal sought by the Sophists in general, Isocrates in particular. Though the Platonic-Socratic ideal is more specialized in its focus on creating discourse, nonetheless, like the Sophistic ideal, it sought a union of verbal skills with learning and wisdom. Specifically, Platonic-Socratic rhetoric became a means of putting into practice the wisdom one acquires in philosophy. In this way Plato and Socrates resolved one of the most serious intellectual issues surrounding the subject: the relationship between truth and rhetorical effectiveness. The resolution, of course, presupposes and maintains a bifurcation between the two.

Aristotle's contributions to rhetoric

Aristotle, too, presupposed and maintained the same division between truth, which was knowable to varying degrees of certainty, and verbal skills, which for Aristotle were primarily useful in assisting truth to prevail in a controversy. But Aristotle lived in a world different from Plato's, one that was closer to the present in the premium it placed upon literacy and upon those patterns of thought that literacy encourages. The literate function of Aristotle's brilliance at recording and categorizing is well captured in Donne's phrase, "Nature's Secretary." Aristotle's *Rhetoric* both recorded contemporary practice and sought its reform through fitting it into its proper category among the arts. One of the masterstrokes of Aristotle's thought on the subject is his teaching that rhetoric itself is not a productive art of making but is an art of doing, embodying a power which is employed in certain kinds of speaking. Further evidence of his brilliance on the subject is his division of speaking into the forensic, the deliberative, and the epideictic and of persuasive appeals into the ethical, the emotional, and the logical. His division of speaking into three kinds reflects his efforts to distinguish rhetoric and its counterpart, dialectics, from philosophy and science. Rhetoric and dialectics, he felt, are concerned with probable matters, in which there are several roads to truth; philosophy and science, on the other hand, are concerned with demonstrable matters, in which the roads are fewer but the truth more certain. In dividing persuasive appeals into three kinds, Aristotle indicated an unmistakable preference for the logical. This preference has been interpreted variously as a result of Aristotle's naive assumption about the rationality of most audiences and as 'an attempt to reform the emotionally charged rhetoric of his contemporaries. In discussing elements of style, Aristotle treated metaphor, perhaps the major figure of speech, in a way that was to plague rhetoricians and poets for centuries. He describes it not as an instrument of thought but as an ornamentation, an adornment that at best serves the functions of clarity and vividness. The effect is further reflection of the principle noted earlier: for Aristotle the truth with which rhetoric is concerned is not demonstrable. It is, moreover, detachable from the forms of argument, and it can be tested by such analytical means as dialectics, which is the counterpart of rhetoric but which does not have what Aristotle viewed as rhetoric's cloying concerns with that beast of many heads, the heterogeneous audience composed of experts and laymen alike.

Rhetorical issues in the age of Quintilian

The Sophistic doctrine that rhetoric should be the central discipline in the educational scheme had a long history, rising to its fullest statement in the writings of Quintilian in Rome of the 1st century AD. By the age of Quintilian three intellectual issues had become firmly fixed within the orbit of rhetoric. Two of these were consciously faced: (1) the relationship between truth and verbal expression and (2) the difficulties of achieving intellectual or artistic integrity while communicating with a heterogeneous audience. In a sense, both of these issues were not faced at all but dodged, as they had been in the past, with the implicit assumption that wisdom and eloquence were not necessarily synonymous and that truth and integrity were ultimately dependent upon the character of the speaker. The orator, according to Cato the Elder, must be a good man skilled in speaking. Through the writings of Cicero, the ancient Roman orator of the 1st century BC whom later ages were to adulate both for his statesmanship and for his prose style, Cato's doctrine was spread in the Western world for centuries. Quintilian's tediously prescriptive *Institutes* is built on Cato's thesis: it offers an educational program for producing generations of Ciceronian statesmen. But for all its importance and influence, the work never found its time so far as being used as a text for political leaders to follow. Quintilian's program was impossible to achieve in the age of tyranny in which he lived, and it was impracticable in the Renaissance. Nevertheless, it was in the Renaissance that the *Institutes* began to be revered as the greatest educational treatise ever written.

A third issue arose in part as a consequence of literacy and in part as a consequence of social change: rhetoric became a productive art, but one whose role and status were unclear. The audience was no longer quite the full partner in the creative event that it had been in older days of freer public discussion; subsequently, from the classical period through the Middle Ages rhetoricians began to conceive of their art as a kind of methodical, solitary progress toward literary creation. Rhetoric was thought of less in terms of a power and more in terms of certain products of that power—orations; elaborate rules were given for distinguishing the kinds of orations and'for arranging the material in them. Accompanying this shift, the entire creative process taught by the rhetoricians became linear and sequential in concept, with some activities located at further and further removes from the serious operations of the mind. A certain linearity, or step-by-step procedure, is evident in Aristotle's *Rhetoric*, but the attendant dangers of compartmentalization and fragmentation into increasingly trivial matters did not make themselves felt for centuries. By the time of Cicero, rhetoric was considered to be a discipline that encompassed five "offices": invention, analyzing the speech topic and collecting the materials for it; disposition, arranging the material into an oration; elocution, fitting words to the topic, the speaker, the audience, and the occasion; pronunciation or action, delivering the speech orally; and memory, lodging ideas within the mind's storehouse. Not only orations but also poems, plays, and almost every kind of linguistic product except those belonging peculiarly to logic (or dialectics) fell within the rhetoricians' creative art. Thus, the function of rhetoric appeared to be the systematic production of certain kinds of discourse, but the significance of this now clearly productive art became increasingly dubious in ages when governments did not allow public deliberation on social or political issues or when the most significant speaking was done by church authorities whose training had been capped by logic and theology.

The five "offices" of rhetoric

The Middle Ages. The early Church Father St. Augustine made one of the earliest efforts to write a rhetoric for the Christian orator. Book IV of *On Christian Doctrine* is usually considered the first rhetorical theory specifically designed for the minister. Of course, the kind of truth to which Augustine sought to give verbal effectiveness was the "revealed" truth as contained in the Scriptures. The first three books of On *Christian Doctrine*, which describe procedures for a proper interpretation of the Bible, actually set forth the invention part of Augustine's rhetoric. There is no basis here for replacing either logic or theology with rhetoric as the capstone of professional training. The work does represent, however, one of the first theoretical efforts to bring together interpretation—that is, interpreting a text, as opposed to interpreting the facts of a case—and rhetoric.

Late in the 13th century, two students of the German philosopher Albertus Magnus produced a great impact upon the thought—particularly the educational thought—of succeeding generations. Thomas Aquinas, who became in effect the preceptor of the theological curriculum, and Peter of Spain, the preceptor of the general or "arts" curriculum, gave articulate force to the current educational practice of making logic the specialty toward which the professional student advanced beyond rhetoric. Thomas wrote on the logic of abstract, symbolic thought, and Peter wrote on the logic of dialectics, disputation among experts.

The Renaissance and after. In the 16th century, at a time marked by a tremendous growth of interest in creating vernacular rhetorics to satisfy a new self-consciousness in the use of native tongues, the French philosopher Petrus Ramus and his followers merely completed the incipient fragmentation of rhetorical theory by affirming the offices as discrete specialties. invention and disposition were assigned to dialectics, by now largely a silent art of disputation which in the Ramist system placed a premium upon self-evident, axiomatic statements. Memory was considered not a matter of creating sound effects to enhance the memorization of the orator's ideas but a matter of effective disposition, so that separate attention to memory disappeared. Elocution and pronunciation were considered the only two offices proper to rhetoric, and these fell under peculiar opprobrium.

The Ramist system

Elocution, or style, became the centre of rhetorical theory, and in Ramist hands it was almost solely concerned with figures of speech. Actually, a strong emphasis upon the figures of speech had been evolving since the late Middle Ages. When responsibly taught, as linguistic postures, stances, gestures of the mind in confrontation with external reality, the figures served a useful purpose; and in Renaissance education they were widely employed, as in the modern manner, in the interpretation or analysis of discourse. Less responsibly taught, the figures became merely an ornamentation, like the metaphor in Aristotle. In the Ramistic system, the figures ranged between serving as arguments and serving as extrinsic decorations. The figures of speech fell into greater disrepute in the new culture of the Renaissance, which was marked not only by an enthusiasm for printed vernacular discourse in a "plain" style but also by an increasing perplexity over doctrines of the passions. For centuries rhetoricians had taught figures of speech as means of "amplifying" ideas so that they would appeal to the passions in an audience. With Ramus, rhetoric discarded its principles of amplification, leaving the passions to be discussed primarily by "moral philosophers," who battled heatedly over which were ordinate and which were inordinate passions. Ultimately, the passions themselves became subjects, or objects, of the new scientists, who divorced them from moral or religious dogma. It was the end of the 18th century before doctrines of the passions fell once more within the rhetorician's purview; however, at that time the figures were regarded less as appeais to an audience's passions and more as manifestations of the author's or speaker's psychology—or, to use the metaphor employed earlier, as places on the map of his mind.

The other part of the fragmented Ramist rhetoric, pronunciation or action, was rarely mentioned in the Renaissance; it hath not yet been perfected, was the excuse the Ramists gave. The first real impetus for a scientizing of English oral delivery came at the beginning of the 17th century from Francis Bacon, who, in touching on rhetoric in his writings, called for a scientific approach to the study of gesture. The Ramists had created a context within which Bacon's call would have peculiar force and meaning. John Bulwer's *Chirologia* (1644) was the first work to respond, and in its wake came a host of studies of the physical, nonverbal expression of ideas and passions, including works by Charles Darwin and Alexander Melville Bell in the 19th century and modern writings on "silent language" by the American linguist Edward T. Hall.

The elocutionary movement

But, so far as rhetorical theory is concerned, even more significant attempts to specialize in the study of pronunciation or action came in the elocutionary movement of the 18th century, which was the first large-scale, systematic effort to teach reading aloud (oral interpretation). The elocutionists named their study for the third office of rhetoric partly because "pronunciation" was coming to refer solely to correct English phonation and partly because "elocution" had traditionally referred to the decorous expression of previously composed material. The most important elocutionists were actors or lexicographers, such as Thomas Sheridan and John Walker, both of whom acted in London and went on to write dictionaries in the late 18th century. At first glance, their efforts to describe or prescribe the oral delivery of written or printed discourse (poems, plays, as well as speeches) appear to operate on extremely inadequate theory: exactly how one discovered the meaning on the page seems mysterious, almost divinatory. Some of their efforts produced such absurdities as statue-like posing or a contempt for the verbal later associated in America with the 19th-century French teacher of dramatic and musical expression François Delsarte. Yet, their efforts may also be seen as attempts to restore the voice to that entire language process which the page abstracted—as attempts to bridge the gap left in concepts of "natural" meaning by the decay of the oral traditions. Moreover, it is most significant that of all theorists within the history of rhetoric, the elocutionists were the first to place an exclusive concern upon interpreting discourse. Indeed, it was through the elocutionary emphasis upon interpretation that something like a meaningful restoration of pronunciation occurred within the rhetorical tradition.

Sheridan had found within the teachings of the 17th-century English philosopher John Locke a foundation on which the study of elocution could be built: words are the signs of ideas, tones the signs of passions. A new, virtually irrevocable split had apparently occurred between spoken language and printed or written discourse. But the split did not produce in other rhetoricians quite the anxiety it produced in the elocutionists. Other rhetoricians began to discover faculty psychology (*i.e.*, the obsolete notion that supposed faculties of the mind such as will and reason account for all human behaviour) and associationism (*i.e.*, the philosophy expostulated by the 18th-century Scot David Hume and others that most mental activity is based on the association of ideas). In these concepts they found a fragmented, compartmentalized means whereby a fragmented, compartmentalized rhetorical theory could recover part of its earlier vast province, as, for example, doctrines of the passions. Pathetic appeals could simply become, as in Hugh Blair's Lectures on Rhetoric *and Belles Lettres* (1783), something like the sixth office of rhetoric. Besides Blair's, the most important rhetorical treatises of the period were George Campbell's *Philosophy* of *Rhetoric* (1776) and Richard Whately's Elements of *Rhetoric* (1828). All three books were written by Protestant clerics, and all reveal the pervasive assumptions of the Age of Reason. Though rhetoric may involve the whole man—indeed, that is the very reason Campbell believed rhetoric properly seen is naturally allied with a science of the mind—nonetheless, man was viewed as an animal with higher and lower faculties, whose intellect was susceptible to being disordered by his passions and whose noble achievement was the creation of rational, preferably written, discourse.

Theories of rhetorical invention of the 18th and 19th centuries seldom treated the exigencies of oral composition before live audiences or even involved an imaginative projection of oneself into a public situation. Rather, they posited an inventive process that was silent, solitary, meditative—a process of conducting solitary, or inward, dialogues. Imagination, that faculty by which man may potentially synthesize what faculty psychology termed his rational and sensory experiences, was not vindicated philosophically until the Romantic movement of the 19th century (and perhaps never effectively). By that time, rhetoric had fallen into discredit. Printed matter had proliferated to such an extent that traditional principles of invention had become antiquated. Eventually all traditional techniques of style and all organized rhetorical study were devalued by interest in experiments; in Switzerland, cultural historian Jacob Burckhardt described antiquity's interest in rhetoric as a "monstrous aberration." In America, the Delsarteans, who stressed gesture rather than words, spread an antirhetorical approach to imagination, the passions, sensory experience, and delivery. Thus, well into the 20th century, "elocution" in popular speech meant florid delivery and "rhetoric," because of its principal concern with oratory, meant purple prose. In academic circles, "rhetoric" referred largely to principles of "belles lettres" until "belletristic" became a pejorative; then "rhetoric" in a host of college "composition" courses referred to less philosophically troublesome principles of paragraph development and thematic arrangement. More than the medieval logicians, more than Ramus, more than all Rationalist philosophers, and more than even the new philosophies of science, it was probably the very momentum of the revolution begun by Johannes Gutenberg's invention of the printing press that caused traditional rhetoric, both as an educational principle and as a theory, to go under.

TOWARD A NEW RHETORIC

These extremely negative views toward rhetoric prevailed until the 1930s, when attention to the importance of studying how language is used was stimulated by Logical Positivism, the philosophical movement that insists that all statements be verifiable by observation or experiment, and that movement had ironically been stimulated in turn by the very scientism that had earlier disparaged rhetoric. Substantial attempts were made, particularly in the United States, to develop an art of discourse suitable for teaching in schools and universities.

In the opening decades of the 20th century, an attempt was made in American universities to restore rhetoric to the serious study of communication (that is, of creating discourse). Teachers of public speaking were the first to turn to rhetorical traditions for help, followed by teachers of writing. (The teaching of speaking had been divorced from the teaching of writing in America since the third quarter of the 19th century—a divorce that has been recognized by modern universities but challenged by the temper of modern life.) Appropriately, considering the impetus of Logical Positivism, the restored rhetoric was largely Aristotelian, an Aristotelianism that was filtered through centuries of faculty psychology, that was becoming part of a doctrinaire stance against the Romantics and the elocutionists, and that was interpreted in terms of lingering presuppositions of a typographical age. Nonetheless, the rhetoric offered through the tenets of a restored Aristotelianism was potentially more comprehensive—more inclusive of all the offices of rhetoric—than any in Western education since the Renaissance. The political facts of modern life, however, made the Rationalist proclivity of this rhetoric appear naïve. The new media—films, radio, and television—and the new orality of modern life was felt by those interested in rhetoric as a challenge to older linguistic notions, not simply those of the print-oriented teachers of written or spoken composition but those of the Aristotelian Positivists as well.

Restored Aristotelianism

Moreover, the restoration of traditional rhetoric was at first—within speech departments and then later within English departments—an attempt to serve as an emphasis upon training students in how to communicate. When modern rhetoricians shifted their emphasis to interpretation and shifted their concerns from the speaker or writer to the auditor or reader, traditional rhetoric was seen in a new perspective and the subject itself was given its strongest modern impetus and relevance. As noted earlier, the latter effect was the combined result of the work of modern philosophers and literary critics as well as educators.

The 20th century witnessed the publication of some highly provocative works on rhetoric, which potentially carry the subject beyond its Aristotelian confines and give it new relevance to an age dissatisfied with older epistemologies (or theories of knowledge) and their curious, divisive assumptions about truth and verbal expression or about oral and written discourse. Particular attention must be called to the work of the American critic and philosopher Kenneth Burke. A controversial writer, partly because of his extension of rhetoric into the study of nonverbal transactions and sensations, he has perhaps done more than anyone else to create a theoretical basis for the use of rhetoric in interpretation.

As noted at the opening of this essay, modern literary critics have helped to free rhetoric from its traditional emphasis by proving its instrumentality in literary analysis—"practical criticism," as the English critic I.A. Richards called his 1929 book on the subject. But in turn the practical critic has helped preserve traditional rhetoric for the analyses of traditional literature, and through his work on modern literature, he has stimulated the demand for a new rhetoric.

THE RHETORIC OF NON-WESTERN CULTURES

Freed, too, of the parochialism engendered by its Western traditions, rhetoric could undertake a variety of analytical endeavours, even "cross-cultural" studies—for example, the mingling of Malaysian and Western cultures in the political oratory of the Philippines, structure and intention in the oral literatures of Africa, or the communicative strategy of the Japanese verse form haiku.

The search for the rhetoric of non-Western cultures

Indeed, the search for the rhetoric of non-Western cultures has become a crucial scholarly and political endeavour, as men seek bases for understanding the politics as well as the poetry of other lands—and, hopefully, bases for dialogue across tribal and national boundaries. The avenues this search has taken thus far reveal a significant fact both about rhetoric and about the nature of its Western tradition: the true rhetoric of any age and of any people is to be found deep within what might be called attitudinizing conventions, precepts that condition one's stance toward experience, knowledge, tradition, language, and other people. Searching for those precepts, the scholar realizes the extent to which Western culture has become secularized and compartmentalized. In Western culture one may seek out a body of writing under such special rubrics as "rhetoric," "religion," "ethics." But in some Oriental or Middle Eastern cultures, the search may begin and end with religious thought and practices. The Talmudic rabbis, with their disputatious hermeneutics and their attitudes toward Oral Law, gave centuries of Jews a pattern of reasoning and communication. No less so did the *Tao-te-Ching*—the basic text of the Chinese religious system of Taoism—shape a mentality that is as inherent in certain Chinese poetry as in the oratory, dance, painting, architecture, and government of that ancient culture. And for all the Western studies one might encourage into the haiku, surely only one thoroughly grounded in the mysterious doctrines of Zen Buddhism can fully understand how that imagistic poetry itself "works." Moreover, as rhetorical doctrine, the form and function of the "sayings" of a modern, secular Oriental revolutionary may not be so far distant from the form and function of the ancient analects of the sage Confucius. Though rhetoric is to be found in every use of language, only Western man has attempted to divide its precepts discretely from the great body of ethical, moral, or religious precepts that condition the very nature of his culture.

In sum, the basic rhetorical perspective is simply this: all utterance, except perhaps the mathematical formula, is aimed at influencing a particular audience at a particular time and place, even if the only audience is the speaker or writer himself; any utterance may be interpreted rhetorically by being studied in terms of its situation—within its original milieu or even within its relationship to any reader or hearer—as if it were an argument. (T.O.S.)

Rhetoric in philosophy: the new rhetoric

There is nothing of philosophical interest in a rhetoric that is understood as an art of expression, whether literary or verbal. Rhetoric, for the proponents of the new rhetoric, is a practical discipline that aims not at producing a work of art but at exerting through speech a persuasive action on an audience.

NATURE OF THE NEW RHETORIC

Theory of argumentation

The new rhetoric is defined as a theory of argumentation that has as its object the study of discursive techniques and that aims to provoke or to increase the adherence of men's minds to the theses that are presented for their assent. It also examines the conditions that allow argumentation to begin and to be developed, as well as the effects produced by this development.

This definition indicates in what way the new rhetoric continues classical rhetoric and in what way it differs from it. The new rhetoric continues the rhetoric of Aristotle insofar as it is aimed at all types of hearers. It embraces what the ancients termed dialectics (the technique of discussion and debate by means of questions and answers, dealing especially with matters of opinion), which Aristotle analyzed in his Topics; it includes the reasoning that Aristotle qualified as dialectical, which he distinguished from the analytical reasoning of formal logic. This theory of argumentation is termed new rhetoric because Aristotle, although he recognized the relationship between rhetoric and dialectic, developed only the former in terms of the hearers.

It should be noted, moreover, that the new rhetoric is opposed to the tradition of modern, purely literary rhetoric, better called stylistic, which reduces rhetoric to a study of figures of style, because it is not concerned with the forms of discourse for their ornamental or aesthetic value but solely insofar as they are means of persuasion and, more especially, means of creating "presence" (*i.e.*, bringing to the mind of the hearer things that are not immediately present) through the techniques of presentation.

The elaboration of a rhetoric thus conceived has an undeniable philosophical interest because it constitutes a response to the challenge of Logical Empiricism. The Logical Empiricists proclaim the irrationality of all judgments of value—*i.e.*, those judgments that relate to the ends of men's actions—because such judgments can be grounded neither in experience nor in calculation, neither in deduction nor in induction. But it is not clearly necessary, after discarding the recourse to intuition as an insufficient basis for a judgment of value, to declare all such judgments equally arbitrary. This amounts to considering as futile the hopes of philosophers to elaborate a wisdom that would guide men in their public as well as their private lives. The alternative offered by the new rhetoric would furnish a complementary tool to traditional logic, which is limited to the technique of demonstration, or necessary proof according to the rules of deduction and induction; it would add the technique of argumentation. This would allow men not only to verify and to prove their beliefs but also to justify their decisions and their choices. Thus, the new rhetoric, elaborating a logic for judgments of value, is indispensable for the analysis of practical reasoning.

SYSTEMATIC PRESENTATION OF THE NEW RHETORIC

Adaptation to the audience

Personal relations with the audience. Argumentation, whether it be called rhetorical or dialectical, always aims at persuading or convincing the audience to whom it is addressed of the value of the theses for which it seeks assent. Because the purpose of all argumentation is to gain or reinforce the adherence of an audience, it must be prepared with this audience in mind. Unlike demonstration, it cannot be conceived in an impersonal manner. On the contrary, it is essential that it be adapted to the audience if it is to have any effectiveness. Consequently, the orator—the person who presents an argument either by speech or in writing to an audience of listeners or readers—must seek to build his argumentative discourse on theses already accepted by his audience. The principal fallacy in argumentation is the petitio principii ("begging of the question"), in which the speaker presupposes that the audience accepts a thesis that actually is contested by them, even implicitly.

Taken in a broad sense, the new rhetoric can treat the most varied questions and be addressed to the most diverse audiences. The audience may involve only the individual deliberating within himself or it may involve another person in a dialogue. The discourse may be addressed to various particular audiences or to the whole of mankind—to what may be called the universal audience —in which case the orator appeals directly to reason.

Classical rhetoric was traditionally addressed to an audience made up of a crowd of generally incompetent hearers gathered in a public place; argumentation, however, can be addressed to highly qualified audiences, such as the members of an academy or some learned society. As a result, effectiveness is not the only means of testing the value of an argument, for this value also depends on the quality and competence of the minds whose adherence is sought. An argument may persuade an audience of less informed persons and remain without effect on a more critical audience. For Plato, the argumentation worthy of a philosopher should convince the gods themselves.

Basis of agreement and types of argumentation. The orator, in order to succeed in his undertaking, must start from theses accepted by his audience and eventually reinforce this adherence by techniques of presentation that render the facts and values on which his argument rests present to the listener. Thus, the orator can have recourse to literary devices, using figures of rhetoric and other techniques of style and composition that are well-known to writers.

If the discourse is addressed to a nonspecialized audience, its appeal will be to common sense and common principles, common values, and common loci, or "places." Agreement about common values is general, but their object is vague and ill-defined. Thus, the appeal to universal values, such as the good and the beautiful, truth and justice, reason and experience, liberty and humanity, will leave no one indifferent, but the consequences to be drawn from these notions will vary with the meaning attached to them by the different individuals. Therefore, an agreement about common values must be accompanied by an attempt to interpret and define them, so that the orator can direct the agreement to make it tally with his purposes. If the discourse is addressed to a specialized group—such as a group of philosophers or jurists or theologians—the basis of agreement will be more specific.

Types of arguments

To pass from the premises accepted by the audience to the conclusions he wishes to establish, the orator can use arguments of various types of association and dissociation. A detailed analysis of such arguments would require a whole treatise; the best known, however, are arguments by example, by analogy, by the consequences, a pari (arguing from similar propositions), ***a fortiori*** (arguing from an accepted conclusion to an even more evident one), ***a*** *contrario* (arguing from an accepted conclusion to the rejection of its contrary), and the argument of authority. The traditional figures of rhetoric are usually only abridged arguments, as, for instance, a metaphor is an abbreviated analogy.

Associative arguments transfer the adherence from the premises to the conclusion; for example, the act-person association enables one to pass from the fact that an act is courageous to the consequence that the agent is a courageous person. Argumentation leads to the dissociation of concepts if appearance is opposed to reality. Normally, reality is perceived through appearances that are taken as signs referring to it. When, however, appearances are incompatible—an oar in water looks broken but feels straight to the touch—it must be admitted, if one is to have a coherent picture of reality, that some appearances are illusory and may lead to error regarding the real. Because the status of appearance is equivocal, one is forced to distinguish between those appearances that correspond with reality and those that are only illusory. The distinction will depend on a conception of reality that can serve as a criterion for judging appearances. Whatever is conformable to this conception of the real will be given value; whatever is opposed to it will be denied value.

Every concept can be subjected to a similar dissociation of appearance and reality. Real justice, democracy, and happiness can be opposed to apparent justice, democracy, and happiness. The former, being in conformity with the criteria of what justice, democracy, and happiness really are, will keep the value normally attached to these notions. The apparent—what is taken for real by common sense or unenlightened opinion—will be depreciated because it does not correspond to what actually deserves the name of justice, democracy, or happiness. By means of this technique of dissociating concepts, philosophers can direct men's action toward what they hold to be true values and can reject those values that are only apparent. Every ontology, or theory about the nature of being, makes use of this philosophical process that gives value to certain aspects of reality and denies it to others according to dissociations that it justifies by developing a particular conception of reality.

Scope and organization of argumentation. A discourse that seeks to persuade or convince is not made up of an accumulation of disorderly arguments, indefinite in number; on the contrary, it requires an organization of selected arguments presented in the order that will give them the greatest force. After its analysis of the various types of arguments, the new rhetoric naturally deals with the study of the problems raised by the scope of the argumentation, the choice of the arguments, and their order in the discourse.

Although formal demonstrative proof is most admired when it is simple and brief, it would seem theoretically that there would be no limit to the number of arguments that could be usefully accumulated; in fact, because argumentation is concerned not with the transfer from the truth of premises to a conclusion but with the reinforcement of the adherence to a thesis, it would appear to be effective to add more and more arguments and to enlarge the audience. Because the argumentation that has persuaded some may fail to have any effect on others, it would appear to be necessary to continue the search for arguments better adapted to the enlarged audience or to the fraction of the audience that has been hitherto ignored.

Limits to argumentation

In practice, however, three different reasons point to the need to set bounds to the scope of an argumentation: First, there are limits to the capacity and the will of an audience to pay attention. It is not enough for an orator to speak or write; he must be listened to or read. Few people are prepared to listen to a 10-hour speech or read a book of 1,000 pages. Either the subject must be worth the trouble or the hearer must feel some obligation to the subject or orator. Normally, when a custom or an obligation exists, it binds not the hearer but the orator, setting limits to the space or time allotted to the presentation of a thesis. Second, it is considered impolite for an orator to draw out a speech beyond the normally allotted time. Third, by the mere fact that he occupies the platform, an orator prevents other people from expressing their point of view. Consequently, in almost all circumstances in which argumentation can be developed, there are limits that are not to be overstepped.

It thus becomes necessary to make a choice between the available arguments, taking into account the following considerations: first, arguments do not have equal strength nor do they act in the same manner on an audience. They must be considered relevant for the thesis the speaker upholds and must provide valuable support for it. It is essential that they do not—instead of reinforcing adhesion--call the thesis into question again by raising doubts that would not have occurred to the audience had they not been mentioned. Thus, proofs of the existence of God have shaken believers who would never have thought of questioning their faith had such proofs not been submitted to them. Second, there is constant interaction between the orator and his discourse; thus, the speaker's prestige intensifies the effect of his discourse, but, inversely, if his arguments are weak, the audience's opinion of his intelligence, competence, or sincerity is influenced. Therefore, it is best to avoid using weak arguments; they may induce the belief that the speaker has no better arguments to support his thesis. Third, certain arguments, especially in the case of a mixed audience whose beliefs and aspirations are greatly varied, may be persuasive for only one part of an audience. Therefore, arguments should be chosen that will not be opposed to the beliefs and aspirations of some part of the audience. Thus, by stressing the revolutionary effect of a particular measure, for example, one stiffens the opposition to that measure on the part of those who wish to prevent the revolution, but one draws to the measure the favour of those who wait for the revolution to break out. For this reason arguments that have value for all men are superior to those that have more limited appeal; they are capable of convincing all the members of what could be called the universal audience, which is composed of all normally reasonable and competent men. An argumentation that aims at convincing a universal audience is considered philosophically superior to one that aims only at persuading a particular audience without bothering about the effect it might have on another audience in some other context or circumstances.

The need for order in argumentation

Further, for a discourse to be persuasive, the arguments presented must be organized in a particular order. If they are not, they lose their effectiveness, because an argument is neither strong nor weak in an absolute sense and for every audience but only in relation to a particular audience that is prepared to accept it or not. In the first place, the orator must have a certain amount of prestige, and the problem in question must raise some interest. Should the orator be a small child, a man of ill-repute, or one supposed to be hostile to the audience or should the question be devoid of interest for the audience, there is little chance that the orator will be allowed to speak or that he will be listened to. Thus, an orator is normally introduced by someone who has the public ear, and the orator then uses the exordium, or beginning portion of his discourse, not to speak about his subject but to gain the audience's sympathy.

Effective arguments can modify the opinions or the dispositions of an audience. An argument that is weak because it is ill-adapted to the audience can become strong and effective when the audience has been modified by a previous argument. Similarly, an argument that is ineffective because it is not understood can become relevant once the audience is better informed. Research into the effectiveness of discourse can determine the order in which arguments should be presented. The best order: however, will often be whatever is expected, whether it be a chronological order, a conventional order, or the order followed by an opponent whose argumentation has to be refuted point by point.

In all these considerations—concerning the techniques of presentation and argumentation and the arrangement of a discourse—form is subordinated to content, to the action on the mind, to the effort to persuade and to convince. Consequently, the new rhetoric is not part of literature; it is concerned with the effective use of informal reasoning in all fields.

It has been seen that common principles and notions and common loci play a part in all nonspecialized discourses. When the niatter that is debated belongs to a specialized field, the discussion will normally be limited to the initiated—*i.e.*, those who, because of their more or less extensive training, have become familiar with the theses and methods that are currently accepted and regarded as valid in the field in question. In such instances, the basis of the argumentation will not be limited to common loci but to specific loci. The introduction in some field of a new thesis or new methods is always accompanied by criticism of the theses or methods that are being replaced; thus, criticism must be convincing to the specialists if the new thesis or method is to be accepted. Similarly, the rejection of a precedent in law has to be justified by argumentation giving sufficient reasons for not applying the precedent to the case in question.

SIGNIFICANCE OF THE NEW RHETORIC

The new rhetoric introduces a fundamental change in the philosophical outlook. Insofar as it aims at directing and guiding human action in all of the fields in which value judgments occur, philosophy is no longer conceived as the search for self-evident, necessary, universally and eternally valid principles but, rather, as the structuring of common principles, values, and loci, accepted by what the philosopher sees as the universal audience. The way the philosopher sees this universal audience, which is the incarnation of his idea of reason, depends on his situation in his cultural environment. The facts a philosopher recognizes, the values he accepts, and the problems he attends to are not self-evident; they cannot be determined a priori. The dialectical interaction between an orator and his audience is imposed also on the philosopher who wishes to influence his audience. Therefore, each philosophy reflects its own time and the social and cultural conditions in which it is developed. This is the fundamental truth in the thought of G.W.F. Hegel, a German Idealist: the history of philosophy is not regarded as an abstract and timeless dialectic that proceeds in a predetermined direction but as an argumentation that aims at universality at a concrete moment in history.

To the extent that the new rhetoric views all informal discourse and all philosophical discourse from the viewpoint of its action on the minds of the hearers, it integrates into the analysis of thought valuable elements from both Pragmatism and Existentialism. In stressing the effectsof discourse it allows Analytical philosophy to be given the dynamic dimension that some scholars believe that it has heretofore lacked. The new rhetoric can thus contribute to the development of a theory of knowledge and to a better understanding of the history of philosophy. (C.Pe.)

BIBLIOGRAPHY. The following works may be regarded as fundamental to the points made in the preceding article: EDWIN BLACK, *Rhetorical Criticism* (1965); WAYNE C. BOOTH, *The Rhetoric of Fiction* (1960); WILLIAM J. BRANDT, *The Rhetoric of Argumentation* (1970); KENNETH BURKE, *The Philosophy of Literary Form* (1941), *A Grammar of Motives* (1945), and *A Rhetoric of Motives* (1950); CHAIM PERELMAN and LUCIE OLBRECHTS-TYTECA, *La Nouvelle Rhetorique: traité de l'argumentation*, 2 vol. (1958; Eng. trans., *The New Rhetoric: A Treatise on Argumentation,* 1969); JOHN CROWE RANSOM, *The New Criticism* (1941); and STEPHEN E. TOULMIN, *The Uses of Argument* (1958). See also CHAIM PERELMAN, "The New Rhetoric: A Theory of Practical Reasoning," in *The Great Ideas Today* (1970).

In addition, the following is helpful in understanding the modern critique of rhetorical traditions: LLOYD F. BITZER and EDWIN BLACK (eds.), *The Prospect of Rhetoric: Report of the National Developmental Project* (1971). RAYMOND F. HOWES (ed.), *Historical Stzidies of Rhetoric artd Rhetoricians* (1961); and R.S. CRANE (ed.), *Critics and Criticism, Ancient and Modern* (1952), are particularly useful in understanding respectively the critics and rhetoricians of Cornell and Chicago, the universities at which modern rhetoric received especially strong impetus. Other works useful in a study of the history of rhetoric include WILBUR SAMUEL HOWELL, *Logic and Rhetoric in England, 1500–1700* (1956); GEORGE KENNEDY, *The Art of Persuasion in Greece* (1963); and WALTER J. ONG, *Ramus: Method, and the Decay of Dialogue* (1958). In addition to Ransom's book, I.A. RICHARDS, *The Philosophy of Rhetoric* (1936), helped illuminate the early stages of the modern relationship between rhetoric and literary criticism. A book-length treatment of non-Western rhetoric is ROBERT T. OLIVER, *Communication and Culture in Ancient India and China* (1971).

(T.O.S./C.Pe.)

Rhine River

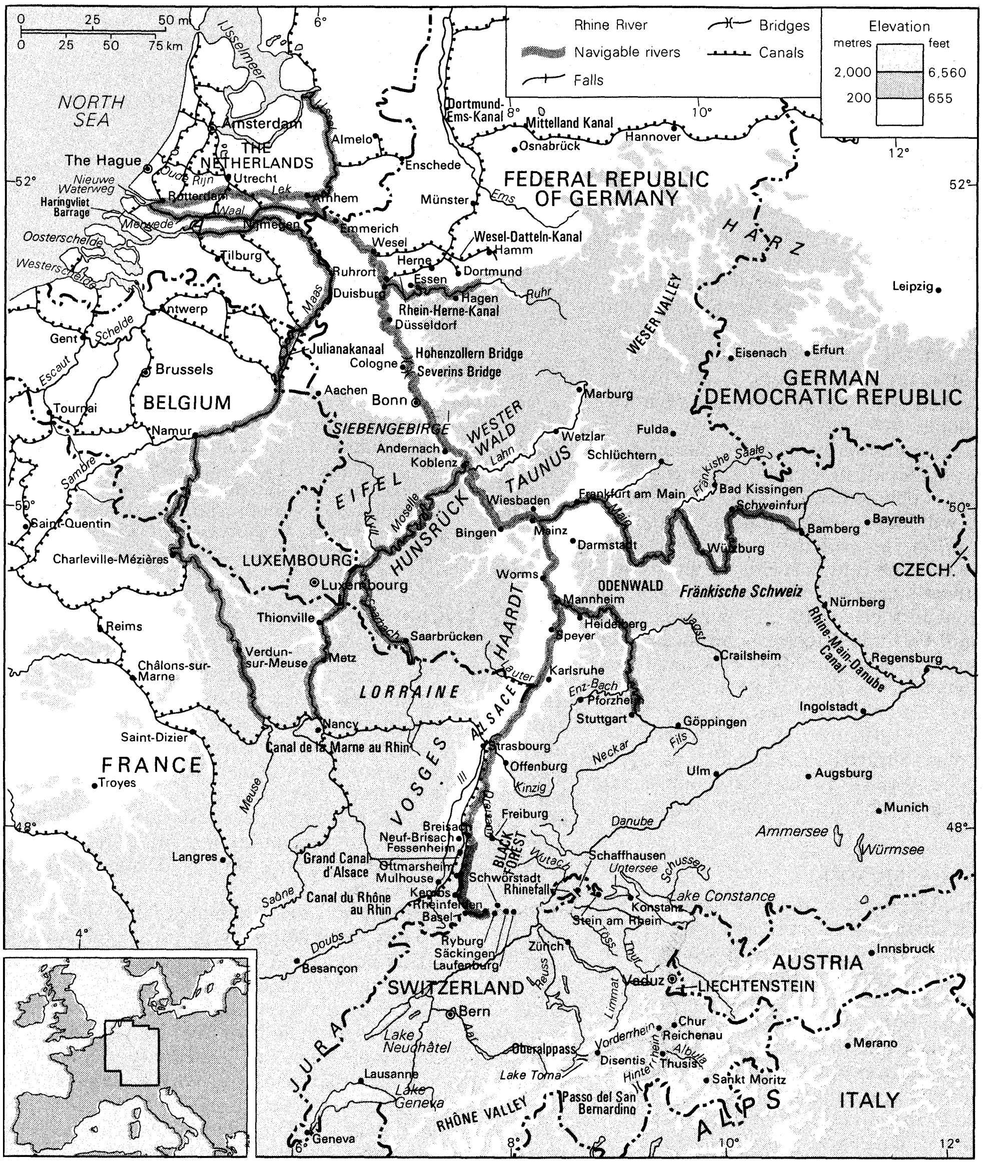
Culturally and historically one of the great rivers of Europe and the greatest European artery of waterborne traffic, the Rhine River flows 820 miles (1,320 kilometres) from east central Switzerland north and west to the North Sea, into which it drains through The Netherlands. An international waterway since the Treaty of Vienna in 1815, it is navigable overall for some 500 miles, as far as Lake Constance (Bodensee) in Switzerland. Its catchment area including the delta area exceeds 85,000 square miles (220,000 square kilometres), about twice the size of Liberia.

The character of the river

The Rhine has been a classic example of the alternating roles of great rivers as arteries of political and cultural unification and as political and cultural boundary lines. The river has also been enshrined in the literature of its lands, especially of Germany, as in the famous epic *Nibelungenlied.* In the second half of the 20th century, its importance as a trade route has increased, and political dissension about its role has given way to concern for ecological safeguards in the face of rising pollution levels.

The Alpine section of the Rhine lies in Switzerland, and, below Basel, the river forms the boundary between the German Federal Republic and France, as far downstream as the Lauter River. It then flows through German territory as far as Emmerich, below which its many-branched delta section epitomizes the landscapes characteristic of The Netherlands. The Alpine Rhine reaches its maximum flow in the spring and early summer, when its volume is swollen dramatically by snowmelt among the great peaks of the Alps. In this section, the beautiful Lake Constance acts as a filter, and the river emerges as a clear, translucent stream on its far side; the lake also helps to regulate river flow. The hydrological regime (highwater and lowwater) of the navigable Rhine is favoured by the well-distributed seasonal precipitation, with a winter maximum in the lower reaches balancing a summer maximum in the Alps. Winters are generally mild, and ice impedes navigation only in abnormally cold years. The scenic attractions of the German Rhine are marred by occasional industrial zones, with associated problems of pollution from industrial waste, but stretches of the river still present breathtaking vistas and attract tourists from near and far. For related information, see EUROPE, and also the articles on the states bordering the Rhine. See also BONN; COLOGNE.

The Alpine section of the Rhine. The Rhine rises in two headstreams high in the Swiss Alps. The Vorderrhein emerges from Lake Toma at 7,690 feet (2,344 metres), near the Oberalppass in the Central Alps, and then flows eastward past Disentis to be joined by the Hinterrhein from the south at Reichenau above Chur. (The Hinterrhein rises near the Passo del San Bernardino and



The Rhine River Basin and its drainage network.

is joined by the Albula River below Thusis.) Below Chur, the Rhine leaves the Alps to form the boundary first between Switzerland and the tiny principality of Liechtenstein and then between Switzerland and Austria, before forming a delta as the current slackens at the entrance to Lake Constance. In this flat-floored section, the Rhine has been straightened and the banks reinforced to prevent flooding. The Rhine leaves the Untersee arm of Lake Constance, and, below Stein am Rhein, the Swiss-German frontier deviates so that the Rhinefalls at Schaffhausen are entirely Swiss. Downstream, the Rhine flows swiftly between the Alpine foreland and the Black Forest region, its course interrupted by rapids, where—as at Sackingen, Laufenburg, and at Ryburg and Schworstadt—barrages (dams) have been built. In this stretch, the Rhine is joined by its Alpine tributaries, the Thur, Toss, Limmat, Reuss, and Aar (French Aare), and by the Wutach from the north. The Rhine has been made navigable between Basel and Rheinfelden since 1934 and will eventually provide a through waterway as far as Lake Constance.

The Lake Constance sector

The central course. Below Basel, the Rhine turns northward to flow across a broad, flat-floored valley, some 20 miles wide, held between, respectively, the ancient massifs of the Vosges–Black Forest and the Hardt–Odenwald. The main Alsatian tributary is the Ill, which joins the Rhine at Strasbourg, and various shorter rivers, such as the Dreisam and the Kinzig, drain from the Black Forest. Downstream, the regulated Neckar, after crossing the Odenwald in a spectacular gorge as far as Heidelberg, enters the Rhine at Mannheim; and the Main leaves the plain of lower Frankische Schweiz (Franconian Switzerland) for the Rhine opposite Mainz. Until the straightening of the Rhine in the early 19th century, the river described a series of great loops, or meanders, over its floodplain, and today their remnants, the old backwaters and cutoffs near Breisach and Karlsruhe, serve to mark the former course of the river. The important river docks were opened at Mannheim in 1840, and three years later the first steam tug, the "Matthias Stinnes," drew coal-laden barges through them. On the French side of the Rhine, a scheme (agreed upon in 1929) was evolved for the construction of the Grand Canal d'Alsace, designed to increase waterborne traffic above Strasbourg and also to generate electricity at a series of power stations located adjacent to the canal locks. The first barrage was built at Kembs, below Basel, in 1932, and, following World War 11—in which the Rhine played a major role in military strategy—installations were completed at Ottmarsheim, Fessenheim, and Neuf-Brisach. After 1956, attention was turned to the regulation of the Moselle, through a project involving agreement between France, the German Federal Republic, and Luxembourg. As a result, this formerly winding and swift-flowing river has been made navigable for vessels of 1,350 tons below Thionville in Lorraine and therefore serves both the French iron and steel region and the Ruhr–Rhine industrial area.

The great navigational improvements after World War II

The most spectacular and romantic reach of the Rhine extends between Bingen and Bonn, capital of the Federal Republic of Germany. In this 90-mile stretch the Rhine has cut a deep and winding gorge between the steep, slate-covered slopes of the Hunsruck to the west and the Taunus to the east. This is the Rhine of legend and myth, where the medieval Mouse Tower lies at water level near Bingen, and the castle of Kaub stands on an island in the river. Vineyards mantle the slopes as far as Koblenz, where the Moselle joins the Rhine at the site known to Romans as Confluentes. On the right bank the fortress of Ehrenbreitstein dominates the Rhine where the Lahn tributary enters. Downstream the hills recede, the foothills of the volcanic Eifel lying to the west and those of the Westerwald to the east. At Andernach, where the ancient Roman frontier left the Rhine, the basalt hills of the Siebengebirge rise steeply to the east of the river, where, as the English poet Lord Byron put it, "the castle crag of Drachenfels frowns o'er the wide and winding Rhine." At Bonn the modern structures of the Bundeshaus (meeting place of the federai assembly) have an impressive frontage flanking the Rhine. Below Bonn the valley opens out into a broad plain in which lignite (brown coal) is excavated for electricity generation.

The city of Cologne lies on the left bank of the Rhine. There the 1,500-foot-wide river is spanned by the modern Severin Bridge and by the rebuilt Hohenzollern railway bridge, which carries the line from Aachen to Dusseldorf and the Ruhr industrial region. Dusseldorf, on the right bank of the Rhine, is the dominant business centre of the Nordrhein-Westfalen coalfield, The twin towns of Duisburg and Ruhrort, which lie at the mouth of the Ruhr River, handle the bulk of the waterborne coal and coke from the coalfield. Following the pattern set by the construction of the Dortmund-Ems-Kana1 in 1899, the Rhine-Herne-Kanal was built in 1914, and the Wesel-Datteln-Kanal was built in 1930. Post-World War II developments have included the introduction of the sturdy vessel known as the push-pull tug, in preference to the diesel-powered barge. The heaviest traffic on the entire Rhine is, in fact, between the industrial section of the Ruhr and The Netherlands' system of inland waterways. More than half the river fleet is Dutch; a third is West German; and the rest is French, Swiss, and Belgian. Apart from the regular passenger-steamer service between Cologne and Mainz, a summer passenger service also operates between Rotterdam and Basel.

The delta region. The last section of the Rhine lies below the frontier town of Emmerich in the delta region of The Netherlands. There the Rhine breaks up into a number of wide branches, such as the Rijn, Lek, and Waal, known downstream as the Merwede. The Meuse (Maas) flows parallel to the Waal and finally forms the New Waterway on the north bank, where the port of Rotterdam lies. To the south lies the Haringvliet Barrage, part of the vast Delta Plan to reclaim and conserve the region's land.

BIBLIOGRAPHY. H.J. MACKINDER, ***The Rhine*** (1908), is a description of the Rhine Basin at the beginning of the 20th century. ALBERT DEMANGEON and LUCIEN FEBVRE, ***Le Rhin*** (1935), is a standard work of reference, especially on the history of the waterway and its regime. R.E. DICKINSON, ***Germany: A General and Regional Geography*** (1953), is a compilation based on a wealth of West German material. See also the chapter by CHRISTOPHER SYKES in ***Great Rivers of Europe*** (1966). A.A. MICHEL, "The Canalization of the Moselle and West European Integration," ***Geogrl. Rev.***, 52:475–491 (1962), deals with the regulation of the Moselle. ALICE F.A. MUTTON, ***Western Europe*** (1971), includes references to the Rhine waterway and the Ruhr coalfield. OSKAR BAR. ***Geographie der Schweiz*** (1971), is a modern textbook on Switzerland, with many illustrations. Atlases that include the Rhine are: ***Atlas der Schweiz*** (1965–78); ***Heimarailas der Südwestmark Baden*** (1934); ***Die Bundesrepublik in Karren*** (1953–72); and the ***Atlas van Nederland*** (1963–77).

(A.F.A.M.)

Rhode Island

Perhaps the most remarkable feature of Rhode Island, one of the six New England states in the northeastern corner of the United States, is its size. About 48 miles (75 kilometres) long and 37 miles wide at the maximum, it is the smallest and one of the most densely populated states in the nation. The nearly 950,000 Rhode Islanders, as reported in the 1980 census, live within an area of 1,214 square miles (3,144 square kilometres), of which about 14 percent is inland water. Many of the state's ample woodlands were once farms, that were abandoned during the 19th century; during the 1800s migrations to the cities resulted in an urban population that by 1900 was 90 percent. Rhode Island has since become one of the most heavily industrialized areas in the world.

This extreme compactness of area and the proliferation of people and activity have tied Rhode Island closely to its neighbours, Connecticut on the west and Massachusetts on the north and east. The Atlantic Ocean, which lies to the south, cuts deep into the state as Narragansett Bay.

The major islands within the state are Block Island and Rhode Island. On the latter is the famous yachting and music centre, Newport. A popular resort since West Indian and Southern planters discovered the island as a summer home in the mid-18th century, Newpor! became an international symbol of wealth and elegance in the 19th century as millionaires from across the country built the many mansions that stand as memorials to a gilded past.

Historical and contemporary character

Small though it is, Rhode Island has been uncompromisingly independent throughout its history. From its founding, in 1636, it was a haven for dissenters from the religious orthodoxies imposed by most of its fellow colonies. As a result, it became something of an outcast among the 13 colonies, but it was the first to declare its independence from Great Britain. After the American Revolution, however, it was the last to join the Union, doing so only under duress, and until 1842 it continued to be governed under terms of the royal charter of 1663. As the state became one of the nation's great textile and manufacturing centres, its cities and mill villages attracted an influx of immigrants whose numbers overwhelmed the original Yankee population. Major internal issues in the 20th century included the longtime withholding of the franchise from both propertyless and foreign-born citizens and the domination of the legislature by rural and small-city interests. (For information on related topics, see the articles UNITED STATES; and UNITED STATES, HISTORY OF THE.)

THE HISTORY OF RHODE ISLAND

Colonial period. In the state's official name—The State of Rhode Island and Providence Plantations—lies a clue to its founding. The first settlement was made by the minister Roger Williams and a few followers at Providence, near the head of Narragansett Bay, in 1636. They were either under edict of banishment from Massachusetts Bay Colony—Williams for advocating freedom of conscience in religion—or were in trouble with the authorities there. In 1638 a group of prominent Bostonians, in similar difficulties, purchased the island of Aquidneck, now Rhode Island, from Indians and settled Portsmouth. Factional strife split this settlement, and William Coddington and his adherents moved to the southern end of the island, where they founded Newport, leaving Anne Hutchinson and her followers in Portsmouth. In 1643 Samuel Gorton took a dissident group south of the boundaries of Providence Plantations and settled Warwick.

Williams went to England in 1643 and returned the following year with a royal patent for the colony, but the four towns could not agree on a form of government until 1647, when a loose confederacy was established. The colony was never accepted into membership in the United Colonies of New England—comprising Plymouth, Massachusetts, Connecticut, and New Haven—and it was constantly threatened with a takeover by one or another of these governments. Coddington, by going to England and having himself made ruler for life of the island towns, split the colony between the mainland and the island towns, Williams and John Clarke, the latter representing island elements unhappy about Coddington's commission, sailed for England in 1651, succeeded in getting the com mission rescinded, and in 1654 set up a reunited government. Clarke remained in England and, in 1663, won a royal charter that was to be the basis of colonial and state government for 180 years.

Although the colony never officially joined the other New England colonies in King Philip's War (1675-76), it suffered greatly. All mainland settlements were burned, including, in the spring of 1676, many houses in Providence. Most of the mainland settlers took refuge on Rhode Island, which was not attacked. The Great Swamp Fight, the battle in which the major portion of the power of the Narragansett Indians was broken, took place in December 1675, a few miles west of the present village of Kingston.

Early commerce with the West Indies

Rhode Island almost from the beginning had commerce with the West Indies, selling horses, barrel staves, and salt fish. Eventually, some of its merchants tried the triangular trade: taking rum to the African coast, where it was traded for slaves, carrying the slaves to the West Indies, where they were traded for molasses, and carrying the molasses to Rhode Island, where it was distilled into rum. The passage of the Sugar Act by Parliament in 1764 seriously affected this trade, and the colony, never overly anxious to obey unpopular laws, began to indulge in considerable smuggling of sugar and molasses. In 1772 the British customs vessel "Gaspee," patrolling Narragansett Bay, ran aground off Namquit (now Gaspee) Point while pursuing a suspected smuggler; that night it was burned by a group of townsmen from Providence. This has been widely regarded as the first act of outright violence against the British crown in the period leading up to the American Revolution.

Traumas of revolution and independence. During the war Newport was occupied by the British. In 1778 a land force under Gen. John Sullivan and the French fleet commanded by the Comte d'Estaing cooperated in an operation designed to dislodge them. Before the French troops could be landed, however, a British fleet appeared in the bay; d'Estaing halted the landing and set out in pursuit. Two days later, before the ships had actually engaged, they were dispersed by a storm. The American ground forces, lacking French assistance, were forced to retreat from the island. At Butts Hill they fought a strong rearguard action that became known as the Battle of Rhode Island and in which a battalion of freed slaves distinguished itself. A Rhode Islander, Gen. Nathanael Greene, distinguished himself as Washington's second in command and as commander of a brilliant campaign in the South.

After the war Rhode Island was reluctant to ratify the Constitution until the Bill of Rights was proposed in the form of 10 amendments. The state's largely agricultural population was opposed to joining the Union, while the merchants of Providence and Newport worked hard for ratification. When threats of commercial isolation from the other states were raised, Rhode Island accepted the document in May 1790, but it did so by a margin of only two votes.

Newport, preeminent before the war, lost much of its economic power during the British occupation, and Providence, led by such merchants as the four Brown brothers, John, Joseph, Nicholas, and Moses, assumed the leadership.

Conflicts over suffrage and equal representation

In 1842 a movement for widening the franchise, which had been limited under the 1663 charter to freeholders and their eldest sons, resulted in a civic conflict known as Dorr's Rebellion. Led by Thomas Wilson Dorr, the son of an aristocratic family, the faction favouring universal suffrage held a convention in 1841 and adopted a constitution embodying this principle. At an election held under this constitution, Dorr was elected governor in 1842, but the election was not accepted as legal by the legislature or the state Supreme Court. When his forces were repulsed in an attempt to seize the arsenal in Providence, Dorr fled the state. Upon his return, he was tried on a charge of high treason, convicted, and sentenced to life imprisonment; he served only one year, however, and was released in 1845. By that time the state had adopted a revised constitution considerably broadening the basis of the franchise, but it was not until the mid-20th century that full rights to vote in all elections were extended to all citizens at the age of 21 (later 18).

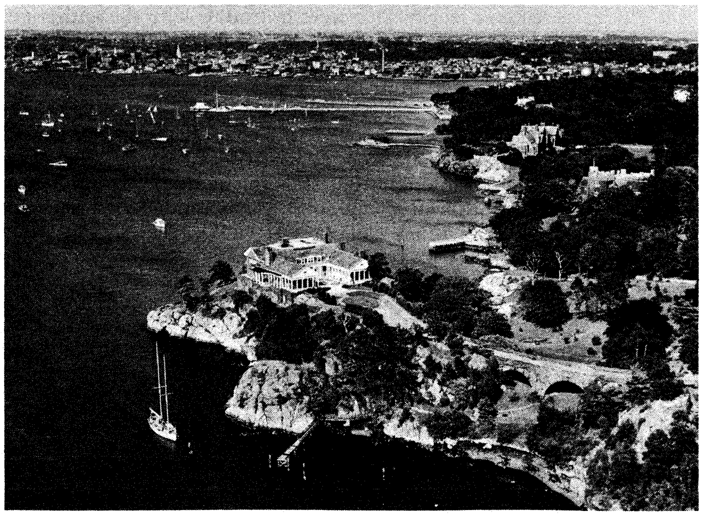
In the years after the Civil War the Republican Party, led by such political bosses as Gen. Charles R. Brayton, ruled the state completely, mainly because the cities, which were Democratic, were not represented in either chamber of the General Assembly in proportion to their population. Providence, by far the largest city, had one senator, as did a small town with only several hundred people.

Providence's representation has grown to comprise 11 senators in the state government, while often several small towns are grouped together with a single senator to represent them. Since the 1930s the Democratic Party has controlled the legislature, although Republicans have been elected governor on several occasions, and Dwight D. Eisenhower and Richard M. Nixon, among the Republican presidential candidates, were able to carry Rhode Island.

THE NATURAL AND HUMAN LANDSCAPE

Topography. Rhode Island's rocky soil is mostly glacial till deposited by the great ice sheets that covered the northern United States thousands of years ago. Glaciation also provided the material that the ocean later formed into the barrier beaches along the state's southern coast, where tidal ponds, open to the ocean through narrow

John T Hopf



Harbour and resort area of Newport, with a 19th-century summer mansion in the foreground.

breachways, stretch from the mouth of the Pawcatuck River at Westerly and Watch Hill to Point Judith. The best farmland is in Portsmouth, Middletown, and areas around the southern coast. There, potatoes and corn (maize) are the principal crops. Elsewhere the land is good only for pasture and dairy or poultry farming. Fields that have been cleared for farming are bounded by omnipresent stone walls. These old stone walls can still be found deep in the woods, evidence that the land once was under cultivation.

Although the western part of the state has hills rising as high as 800 feet (240 metres) above sea level, most of it is quite flat, and the average altitude is about 200 feet. The highest points are Jerimoth Hill, 812 feet, in the town of Foster and less than a mile from the Connecticut border; and Durfee Hill, 804 feet, in Glocester, a few miles farther north. The lowest part of the state, other than the ocean beaches, is in the vicinity of the Great Swamp in South Kingstown.

Climate. With a reasonably salubrious climate, the state is not normally subjected to great extremes of either heat or cold. The average mean temperature at Providence is 50° F (10° C), while the mean for summer is 70" F (21° C) and for winter 30° F (−1° C). Mean annual precipitation throughout the state is about 37 inches (940 millimetres).

Water-power and the mill villages

Waters of the state. Two large river systems drain Rhode Island. Most important is the Blackstone River and its tributaries. Rising near Worcester, Massachusetts, it cuts through the northeastern part of the state and was the source of waterpower for numerous textile mills that were built in the Blackstone Valley, at Woonsocket, Pawtucket, and a dozen mill villages between the two cities. The Pawcatuck River and its main branch, the Wood, provided power for numerous other small mills. Although it drains a smaller area, the Pawtuxet River was once of comparable economic importance, and its valley is crowded with mill villages.

These mill villages, which thrived in the 19th century, give rural Rhode Island its characteristic look. Small one- and two-family houses, set out in orderly rows along a single street, are not far from the mill where the residents worked. Formerly these houses were built and owned by the mill owners, but, with the demise of many of the mills or their conversion to kinds of manufacturing other than textiles, the mill village houses were sold off to the persons who lived in them.

Role of Narragansett Bay

Before the mills Narragansett Bay was the state's great asset, providing a convenient waterway running two-thirds of the length of the state and navigable as far up the Blackstone Valley as Pawtucket Falls. The small commercial trade on which the wealth of Newport, Bristol, and Providence was founded has given way to a larger ocean-borne commerce. Storage tanks for oil and gasoline dot both shores of the upper bay, and Providence has become one of the principal oil-distributing centres of the Northeast.

Narragansett Bay also has been useful in bringing U.S. Navy establishments to Rhode Island. The Naval War College has operated at Newport since the middle of the 19th century. Although before World War II the navy occupied about 500 acres (200 hectares) in Newport County, a Naval Operating Base was established and enlarged to more than 2,100 acres in 1941. The navy's presence continued to be strong through the late 1960s, when about 42,000 people were employed at Newport and other naval bases. By 1980, however, the number had been reduced to 9,500 people.

Other major waterways of the state are Mount Hope Bay, an arm of Narragansett Bay, providing navigable water to Somerset and Fall River, Massachusetts, and the Sakonnet River, a saltwater strait separating the island of Rhode Island from Little Compton and Tiverton on the east. Block Island Sound, lying between Block Island and the mainland, is highly regarded by sport fishermen seeking marlin and tuna and also yields commercial quantities of swordfish.

Patterns of settlement. Rhode Island comprises five counties, eight cities, and 31 towns. It sometimes is described as a city-state, and a large proportion of its population resides in Providence or in contiguous communities, such as Pawtucket on the north, North Providence and Johnston on the west, Cranston and Wanvick on the south, and East Providence across the river to the east. Except for Warwick, all of these communities are in Providence County, which in 1980 had a population density of 2,496 persons per square mile. Warwick and Cranston, the state's second and third largest cities, respectively, had populations in 1980 about one-half that of Providence.

THE PEOPLE OF RHODE ISLAND

Ethnic composition. Although the population of the state has undergone considerable change since World War II, with the exodus from the congested parts of such cities as Providence to the "bedroom communities" of Cranston and Warwick, much remains the same. Woonsocket still has a large population with French-Canadian ancestry whose first language is French. Pawtucket retains a large group of English whose ancestors came to work in the textile mills, and Providence has a strong Italian-American community as weil as a iarge number of Irish-Americans. In the small towns of the western and southern parts of the state, however, the typical Yankee farmers predominate, though mixed with an increasing number of people of Portuguese, Finnish, and Polish extraction and others who have fled the cities. By the 1980 census the state was 94.7 percent white. Providence had the greatest number of blacks and other nonwhites.

Although ethnic origins are a strong factor in political actions in Rhode Island, the state has returned to the principles of toleration on which the colony was founded. Despite much unwillingness to assimilate immigrant groups in the 19th century, the state has always clung to its ideals of religious freedom. In the 17th and early 18th centuries, Rhode Island was a refuge especially for Quakers and Jews, both of whom contributed substantially to the wealth and economic power of Newport. Touro Synagogue in Newport, a colonial building that has been carefully restored in decor and furnishings, is the oldest synagogue in the United States; it has been declared a national landmark.

Demography. The movement out of the urban centres caused population losses in six Rhode Island cities, according to the 1980 census. Newport lost more than one-sixth of its population and Providence nearly one-seventh, while Central Falls, Granston, Pawtucket, and Woonsocket lost smaller percentages. On the other hand, towns and cities near these centres or accessible by highways grew enormously, primarily East Greenwich, Narragansett, Portsmouth, and West Wanvick. During the 1970s Cranston replaced Pawtucket as the state's third largest city. While the state's population grew steadily until the 1970s, that of Providence reached its peak in 1925 and has decreased ever since. Like other predominantly urban states, Rhode Island has begun to feel the effect of the flight from the cities, which leaves behind more and more of the unemployed, the economically depressed, the unskilled, and those on the public welfare rolls. At the same time the cities continue to lose much of their economic and tax base.

THE STATE'S ECONOMY

Components of the economy. Rhode Island depends increasingly on brainpower and skilled labour to keep its economy afloat. Relying in its early years almost entirely on subsistence agriculture and seaborne trade, it became after the Revolution a pioneer manufacturing state, principally in the textile field. In the 20th century it has developed more diversified small industries, especially in the jewelry and allied industries, electronics, and such service fields as education and insurance.

Manufacturing employment and value

The largest group of workers is engaged in the jewelry and allied trades, whereas the once dominant textile industry has fallen to second place. The manufacture of machine tools and precision-measuring instruments is an important element in the economy; coupled with the fabrication of metal products and electrical equipment, it often employs large numbers of people, although any national retrenchment in heavy industry, such as automobile manufacturing, is quickly reflected in Rhode Island's economy. Since the jewelry industry is also remarkably sensitive to fashion trends and available money supplies, employment in this trade is frequently seasonal or otherwise erratic. One of the stablest elements of the economy is the state's educational institutions.

Rhode Island is not a rich state, and its per capita income places it in the middle ranks among the states. Jewelry and silverware manufacturing is Rhode Island's leading industry, with textiles second and metal-working machinery third.

Two Providence-based banks, the Industrial National and the Rhode Island Hospital Trust, through their branches throughout the state, are the dominant financial institutions, but there are several important banks of secondary rank.

Rhode Island has a sales tax, which exempts food and prescription drugs. In 1971, after many defeats in the General Assembly, the state adopted a personal income tax. Lotteries were widely used in colonial times and after the Revolution to raise capital for civic improvements; they later were prohibited but were reestablished in 1974. A large share of the state's income comes from its tax on pari-mutuel betting at two horse-racing tracks, Narragansett and Lincoln Downs.

Transportation. In addition to its bridges and ferries, Rhode Island has many other connections with states of the Eastern Seaboard. Providence is the centre of limited-access expressways to Boston, less than one hour away, to Cape Cod, and westward and southwestward into Connecticut. A main railroad provides the state with freight and passenger service, while interstate and intrastate bus lines run on its highways. There are several airports in Rhode Island, with most of the passenger service handled at the Theodore Francis Green State Airport, located at Hillsgrove.

ADMINISTRATION AND SOCIAL CONDITIONS

Atop the state capitol in Providence is the statue of "The Independent Man," a symbol of so much that is characteristic of Rhode Island. Roger Williams wrote that he had founded Providence as a place of refuge for "those distressed for cause of conscience," and the principle of absolute religious freedom has been an abiding article of Rhode Island's political philosophy. Because of the nature of its first settlements, which were not bound to any one religious faith, most Rhode Island towns do not have a central common, as do many New England towns, where the houses are usually clustered around a central group of churches or a single church.

Governmental traditions. Ever since its founding, the state has shown a reluctance to permit elected officials to exercise extensive powers. In the early years legislatures and elected officials served for only six months. Prior to 1854 the state had five capitals, Providence, Newport, East Greenwich, Bristol, and South Kingstown, and the General Assembly travelled from one to another. In 1900 Providence was chosen as the sole capital.

Executive, legislature, and judiciary

The governor has the power of the veto and the power to name certain department heads. Judges are elected by both houses of the General Assembly sitting as a committee of the whole, and since the 1930s they have had tenure.

Members of the General Assembly are paid by the day for 60 legislative days. A number of attempts to amend the state constitution to increase legislators' pay have been soundly beaten by the voters, who in 1968 defeated by a four-to-one margin a new constitution to replace the 126-year-old document.

In addition to the usual officers and the legislators, the state has a district court system, superior courts, a Supreme Court, which also can give advisory opinions if requested, and municipal and probate courts, the latter frequently identical with the town council in the smaller communities. There is also a Family Court, which handles both juvenile and domestic cases. In addition, a U.S. District Court sits in Providence.

Most of the cities in Rhode Island operate with a mayor and city council form of government, but some have a city manager, and the mayor is chosen from among councilmen to act as the ceremonial head of the local government. Likewise, most of the towns are governed by a town council, but in some cases operations are conducted by a town manager.

The social milieu. Early in its history, Rhode Island acted to stamp out slavery by limiting to a maximum of 10 years the period in which anyone could be held a slave. During the Revolution the General Assembly acted to raise a black regiment by decreeing that any slave who enlisted would be granted his freedom. Even in the 18th century there was a strong emancipation movement in the state, drawing much of its support from such men as Moses Brown and other prominent Quakers. In 1866 the General Assembly passed a law prohibiting the exclusion of any person from the public schools because of race or colour, thus closing separate schools maintained in Bristol, Newport, and Providence and admitting blacks to Providence High School.

Early traditions of toleration

The state maintains elaborate facilities to care for the sick and indigent, from children's centres to institutions for the elderly.

There are jails in cities and towns throughout the state, but most persons awaiting trial or sentenced to prison terms are sent to the Adult Correctional Institutions (ACI) at Howard. The ACI in recent years has adopted a work-release program to rehabilitate prisoners. Rhode Island has a force of more than 300 officers and policemen who operate out of seven barracks located throughout the state.

Education. In 1970 a Board of Regents was created and charged with responsibility over all public education, from elementary schools through the state-operated colleges and the university. A number of private preparatory schools, both sectarian and nonsectarian, send graduates to many of the major colleges and universities, especially in the East. Rhode Island is strong in its institutions of higher education. Brown University in Providence, founded in 1764 as Rhode Island College, is one of the major Eastern universities that make up the so-called Ivy League. It is noted for its library facilities, especially the John Carter Brown Library of early Americana. The Rhode Island School of Design (founded 1877) is widely known, primarily for its training in the visual and graphic arts. The University of Rhode Island, in Kingston, is a land-grant institution dating from 1892. Catholic colleges include Providence College (1917) and Salve Regina College (1947) in Newport. Rhode Island College, mainly for teacher training, dates from 1854.

Higher education

CULTURAL LIFE AND INSTITUTIONS

Cultural and historical preservation. Library facilities are plentiful throughout the state. The Redwood Library, in Newport, and the Providence Athenaeum, bbth proprietary institutions housed in architecturally important buildings, have roots going back to the mid-18th century. The public libraries of Providence and Westerly have important holdings, the former having special collections on whaling, printing, slavery, and Irish literature. The Rhode Island Historical Society, in Providence, has more than 1,000,000 manuscripts and is especially strong in its holdings of the state's newspapers. The society also operates John Brown House, a magnificent merchant's mansion in Providence; built in 1786, the house is furnished with masterpieces of the Newport school of cabinetmakers and with other 18th-century antiques.

The arts. The Museum of Art of the Rhode Island School of Design is especially strong in Greek and Roman sculpture and antiquities, Postimpressionist French paintings, American painting, and British watercolours. The School of Design also has a fine art library.

In both Providence and Newport, preservation societies have been active in restoring the surviving colonial homes. The Preservation Society of Newport County operates as museums several mansions that were formerly the summer homes of millionaires. The Newport Historical Society museum, with its fine collections; Touro Synagogue, a magnificent example of colonial architecture; Old Colony House; Redwood Library; Hunter House; the restored colonial homes of the Point section; and the National Lawn Tennis Hall of Fame and Tennis Museum in the Newport Casino building combine to give Newport an extraordinary and varied cultural heritage.

Music festivals

The Rhode Island Philharmonic Orchestra, the Rhode Island Civic Chorale and Orchestra, the Westerly Community Chorus, the Barrington Boys Choir, and numerous smaller groups and organizations are among the state's musical resources. It is visited regularly by the Boston Symphony Orchestra, and it has an annual chamber music series, as well as several series featuring renowned soloists. The State Ballet of Rhode Island performs regularly throughout the state. Many of the restored houses in Newport were the setting for the Newport Jazz Festival, a world-famous event begun with annual summertime jazz concerts in 1954 and joined in the 1960s by folk-music presentations. The festival was moved to New York City in 1971.

The Trinity Square Repertory Company, with its own home in Providence, is renowned for producing works by new piaywrights, as well as for staging novel productions of classic works. In 1968 it became the first regional theatre in the United States to be invited to perform at the Edinburgh International Festival of Music and Drama in Scotland.

Sports. Many of the recreational activities of the people of Rhode Island are centred around the state's water. For many years the waters off Newport have been the site of the yacht races for the America's Cup. An annual tuna tournament is held in Rhode Island and Block Island sounds. Newport Casino, one of the early centres of tournament tennis, has an annual grass-court tournament of national importance.

Communications. The state has several daily papers, the largest of which are *The Providence Journal* and *The Evening Bulletin.*

The state is served by three commercial television stations and by an educational station operated by the state. Boston television stations are also received in most of Rhode Island, and there are numerous radio stations.

Prospects. Rhode Island has had to live by its wits. Lacking any important natural resources except Narragansett Bay and the waterpower that could be generated on its river systems, the state prospered first by seaborne commerce and subsequently by manufacturing, especially textiles, machinery, and jewelry. The attractiveness of its colonial heritage in architecture, its assets as a summer resort, its suitability as a base in the Northeast for the navy, and its emphasis upon independence of thought and freedom of religion have given it a character unlike any other New England state.

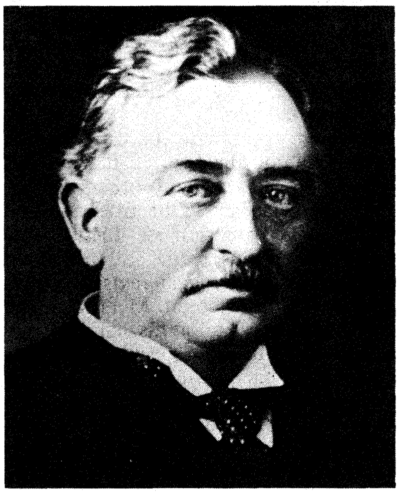
BIBLIOGRAPHY. The biennial *Rhode Island Manual* contains extensive factual data on governmental officials, committees, etc.; while the annual *Journal-Bulletin Rhode-Island Almanac* contains recent election results, economic data, and much miscellaneous information not easily obtainable elsewhere. The FEDERAL WRITERS' PROJECT, *Rhode Island* (1937, reissued 1973), is excellent for its factual material on the state's various communities. EDWARD FIELD (ed.), *State of Rhode Island and Providence Plantations at the End of the Century: A History,* 3 vol. (1902), though old, remains the most complete history for the period covered; HOWARD M. CHAPIN, *Documentary History of Rhode Island,* 2 vol. (1916), is indispensable for the earliest years of settlement; WILLIAM G. MCLOUGHLIN, *Rhode Island* (1978), is a shorter, interpretive history of the state. SAMUEL H. BROCKUNIER, *The Irrepressible Democrat, Roger Williams* (1940), is the best biography of Williams; while PERRY MILLER, *Roger Williams: His Contribution to the American Tradition* (1953), is the most important reassessment of Williams' position as a liberal thinker. JAMES B. HEDGES, *The Browns of Providence Plantations,* 2 vol. (1952–68), is a classic study of Rhode Island's most influential family. ANTOINETTE F. DOWNING and VINCENT J. SCULLY, JR., *The Architectural Heritage of Newport, Rhode Island, 1640–1915,* 2nd ed. rev. (1967), contains detailed studies of the architectural developments from colonial days to the "summer palaces" of the millionaires. Publications of the Rhode Island Historical Society, especially the quarterly *Rhode Island History.* are valuable for recent developments.

(B.F.S.)

Rhodes, Cecil

The life of Cecil John Rhodes, South African financier and statesman and one of the great empire builders, spanned the heyday of British imperialism. When he was born, central Africa was virgin territory, and it was still possible to niake a fortune and acquire power by what Rhodes called "philanrhropy plus five percent"; or to dream of "painting the map red" (to plant the British flag) from the Cape of Good Hope to Cairo. He achieved the first and made advances toward the second. By the time he died, however, the second South African (Boer) War had already exposed the weaknesses of imperialism. His notions of the superiority of the Anglo-Saxon race soon faded, and his real dreams were never fulfilled. His name survives in the Rhodes scholarships at the University of Oxford.

EB Inc



Rhodes.

Early struggles and financial successes. Rhodes was born on July 5, 1853, the son of the vicar of Bishop's Stortford in Hertfordshire. The family's roots were in the countryside, where Cecil Rhodes always felt at home: tree planting and agricultural improvement were among his lifelong passions, though his earliest ambition was to be a barrister or a clergyman. His father was prosperous enough to send one son to Eton College, another to Winchester College—both leading English private schools—and three into the army. Cecil, however, was kept at home because of a weakness of the lungs and was educated at the local grammar school. Poor health also debarred him from the professional career he planned. Instead of going to the university, he was sent to South Africa in 1870 to work on a cotton farm, where his brother Herbert was already established. Though hc was never close to his father, parting from his mother and home caused him intense sorrow.

Diamond prospecting

The cotton farm in Natal, now a province in the Republic of South Africa, was not a success. On his arrival Rhodes found that his brother had already left for the diamond fields of Griqualand West, a former division of what is now the Republic of South Africa. Although Herbert returned to the farm, and the two brothers continued stubbornly trying to grow cotton for a year, the "diamond fever" eventually overcame them. In 1871 they moved to Kimberley, the centre of mining, where life was even harder than in Natal. Herbert was restless and stayed only till 1873, but Cecil's characteristic determination kept him at Kimberley off and on for years. until at last, after a harsh beginning, his luck turned.

Rhodes was early seen as a man apart. He kept his own company and dreamed his dreams. Shy himself, at the occasional dances he sought out the plainest girls, saying that he danced only for exercise, He was addressed only as "Mr. Rhodes," and his signature on letters, even to close friends, was always "C.J. Rhodes." Unlike other adventurers seeking their fortunes, he admired the pastoral

Boers—the descendants of the Dutch settlers in South Africa—and respected their love of their land. A severe illness in 1872 led him to make a journey of convalescence to the north, from which he returned with a fresh view of the potentialities for British expansion in Africa. Another illness in 1874 isolated him still more from the rough society of Kimberley; so did his habit of intellectual study, which led him to return home in 1873 to become an undergraduate at Oxford.

For eight years, until he took a belated degree in 1881, he divided his life between Kimberley and Oxford. Both societies found him odd, though he did his best to conform outwardly to the conventions. At Oxford he read the classics desultorily, attended occasional lectures, and became attracted to the ideas of social reform of the English essayist and art critic John Ruskin. Rhodes's eccentric habits, his falsetto giggle, his rambling monologues, and his unusual background intrigued the younger men around him. So did his philosophy of an almost mystical imperialism. "To be useful to my country" summed up his creed, which somehow he identified with Aristotle's theory of "the good life."

In 1877, while still an undergraduate, he drew up the first of his famous wills, leaving the fortune that he had not yet made to the colonial secretary and others for the purpose of founding a secret society, "the true aim and object whereof shall be the extension of British rule throughout the world." In various forms, all his seven wills had some such object in view, each revision representing a new development of his thought. Meanwhile, he returned annually to Kimberley to exploit his claims.

Formation of the De Beers Mining Company

He gradually advanced from being a speculative digger to the status of a man of substance with ambitious ideas on the future of the diamond industry. His first partnerships were with young men as impoverished as himself, such as C.D. Rudd, with whom he formed the De Beers Mining Company—so called after the De Beers mining claims, many of which he had acquired. Eventually, success brought new friends and also rivals. Alfred Beit, a German who knew the diamond market intimately, was his most valued friend. The most fateful was a young doctor, Leander Starr Jameson, who became his medical adviser in 1878. Outstanding among his rivals was Barney Barnato, from the slums of London, with whom he struggled for control of the diamond fields; it was always Rhodes's way to "square^{m}—buy out—a rival if he could, rather than fight him, and finally he did so in the case of Barnato, too.

With Beit's help, Rhodes expanded his claims until all the De Beers mines were under his control. In 1887 he set about acquiring the Kimberley mine, which was mainly controlled by Barnato. A furious competition to buy up shares ended in Rhodes's favour in 1888. He finally paid over £5,000,000 ($25,000,000)—a generous settlement—for Barnato's holding and celebrated by making his rival a member of the Kimberley Club, into which Barnato had never before even been admitted. Other lesser mines fell under Rhodes's control, until by 1891 his company, De Beers Consolidated Mines, owned 90 percent of the world's production of diamonds. He also acquired a large stake in the Transvaal gold mines, which had been discovered in 1885, and formed the Gold Fields of South Africa Company in 1887. Both Rhodes's major companies had terms in their articles of association allowing them to finance schemes of northward expansion.

Political involvement in Africa. Rhodes never regarded moneymaking as an end in itself. "Painting the map red," building a railway from the Cape to Cairo, reconciling the Boers and the British under the British flag, even recovering the American colonies for the British Empire, were all part of his dream. With these ideas in view, he first went into politics in 1881, offering himself for election to the parliament of the Cape Colony in a constituency in which he had to depend on Boer support. He held it for the rest of his life. Though unimpressive as a speaker and contemptuous of parliamentary procedure, he earned respect by his original views. He made friends with many Boer politicians, particularly Jan Hofmeyr, W.P. Schreiner, and in the Afrikaner Bond, a leading Boer organization in the Cape Colony; he espoused the cause of the natives in what were then Basutoland and Bechuanaland (now Lesotho and Botswana); and always he had his eyes fixed on the north. The north, however, included the Boer republic of the Transvaal, now in the Republic of South Africa, which successfully fought a brief war of independence with Britain in 1880–81; and its president, Paul Kruger, was one obstinate rival whom Rhodes could never "square."

His first intervention in native policy came in 1882, when he was appointed to a commission to pacify Basutoland after a minor rebellion. The rebellion had been put down by the former British governor of the Egyptian Sudan, Gen. Charles Gordon, acting for the Cape government. He succeeded not by force but by organizing discussion meetings with the tribal chiefs. Rhodes was impressed by the man and his methods, though less favourably by the contempt that Gordon showed for financial reward. Gordon too was impressed—so much so that when he returned to Khartoum two years later, he invited Rhodes to accompany him. Rhodes, however, refused; although the Sudan was included in the territory he wanted eventually to paint "all red" on the map, he intended to reach it by advancing from a secure base in the south.

His determination to keep open a road to the north involved him in many disputes. Other imperial powers—the Germans, Belgians, and Portuguese—were in competition for the uncharted interior of Africa, as were the Transvaal Boers. The missionaries were, in Rhodes's view, overly solicitous of native interests; the Cape government was weak; and the British government, which he called the "imperial factor," was too distant to understand his ideas. But he assiduously cultivated the government's representatives in Cape Town—particularly the high commissioner Sir Hercules Robinson—with profitable results.

Acquisition of Bechuanaland

The crucial area was Bechuanaland, through which ran the route used by the missionaries. Rhodes intended to use it to open up the northern territories of Mashonaland and Matabeleland (both now in Rhodesia). "If we get Mashonaland," he would say, "we shall get the balance of Africa." Mineral wealth, communications, and, eventually, white settlement were his objectives. All the boundaries were unsettled, however, and many intrusions had to be frustrated first. Boers from the Transvaal, trying to annex slices of Bechuanaland, proclaimed two small independent republics in Stellaland and Goshen. In 1882 a boundary commission, to which Rhodes again secured appointment, was sent to settle the boundaries of Griqualand West. Rhodes persuaded the commission to extend its mandate to the two small republics. In 1884, when the Germans in South West Africa declared a protectorate over two territories (which, along with Stellaland and Goshen, would have sealed off the Cape Colony from the north), he persuaded the high commissioner that the Imperial government must intervene. By the London Convention of 1884, the two republics were excluded from the Transvaal, and the Cape government agreed to help finance a protectorate over Bechuanaland. During the struggle, Rhodes briefly held office in 1884 as treasurer general of the Cape Colony, but the government soon fell when an outbreak of phylloxera, a plant louse that attacked the vineyards, was blamed on the government.

His settlement of the Bechuanaland question was also soon threatened, for the deputy commissioner in the new area, John Mackenzie, a missionary whom Rhodes contemptuously labelled a "negrophilist," antagonized the Boers. Rhodes insisted on his removal and was appointed in his place. He succeeded in conciliating the Boers of Stellaland but could not prevent Kruger from declaring a protectorate over Goshen, from which he withdrew only after an expeditionary force was sent up from the Cape. A conference to settle the matter was held in February 1885 on the River Vaal, where Rhodes and Kruger met for the first time. These two stubborn men, each determined to dominate Africa, each ever ready to quote Scripture for his purpose (Kruger always from the Old Testament, Rhodes chiefly from the New), naturally failed to achieve any meeting of minds.

Struggle with Kruger

Although Kruger was forced to give up Goshen, Rhodes did not get everything his own way. It was decided that southern Bechuanaland should become a crown colony and northern Bechuanaland a protectorate. Rhodes, who wanted both annexed by the Cape Colony, resigned in protest in March 1885 and thereafter devoted strenuous efforts, both in Cape Town and London, to securing the transfer of the coiony to the Cape government.

Two men still stood in the way of Rhodes's plans for developing the north. One was Kruger, with his policy of "Africa for the Afrikaners"—the Boers. By the Franchise Law of 1890, he denied political rights to the foreigners (Uitlanders) who had come to work the gold mines in the Transvaal. He also tried to extend Boer control to Mashonaland and Matabeleland. The ruler of the Matabele was King Lobengula, Rhodes's second obstacle. Kruger had approached him for a treaty and mining concessions in 1887, and so had many others. Lobengula, however, though uneducated, knew that once he let the white men in, he would never see their backs. The only white men he trusted were missionaries; and Rhodes duly found in John Moffat, the son of a famous missionary, a man to serve his purpose.

Once Moffat, as Assistant Commissioner for the crown colony of Bechuanaland, had, in February 1888, persuaded Lobengula to sign an exclusive treaty of friendship, Rhodes sent three of his trusted agents to obtain a mining concession based on the treaty. The concession was extracted from the reluctant Lobengula in October 1888: to the last, he hoped he had only allowed the white man to dig "a big hole." In fact, however, he had virtually signed away his kingdom, and Rhodes hastened to press the Imperial government, through the high commissioner, to grant a charter to a new company, the British South Africa Company, to develop the new territory. Difficulties had still to be overcome: other concession hunters had to be bought off; Lobegula's doubts had to be overcome by sending Doctor Jameson with a gift of weapons; the Aborigines Protection Society intervened to warn Lobengula. In October 1889, however, the charter was granted, and Lobengula, captivated by Jameson's charm, allowed the digging to begin.

By then Rhodes had powerful allies: bankers, such as Lord Rothschild; influential journalists; empire builders, such as Harry Johnston and Gen. Horatio Herbert Kitchener; politicians, such as Lord Rosebery. He also attracted suspicion and hostility: Lord Salisbury and Joseph Chamberlain—both powerful politicians--disliked him, as did the future Lord Lugard, the colonial administrator, and Henry Labouchere, a radical M.P. Rhodes's habit of seeking to "square" people who might oppose his plans by offering them large shares in his companies was open to misunderstanding. So were his contributions to the Liberal Party on condition that it should not give up British control of Egypt and to the Irish Nationalists on condition that they should accept Home Rule under a federal parliament. He was useful, however, to a parsimonious government, because he undertook to develop vast territories at his own expense. Queen Victoria found his imperialism attractive, no less than his courtly rebuttal of the accusation of being a woman hater: "How could I dislike a sex to which your Majesty belongs?" The upshot of his successful propaganda was that the charter granted by the Imperial government went far beyond what Lobengula had conceded. There was no northern limit on it; and Rhodes intended to extend the chartered company's control to Northern Rhodesia (now Zambia) and Nyasaland (now Malawi), as well as to the Bechuanaland Protectorate (now in Botswana).

Founding of Salisbury

In 1890 Rhodes's Pioneers began their hazardous march into Matabeleland and thence to Mashonaland, where they established a fort in September, to be called Salisbury, after the British prime minister. The size of the expedition, clearly intended for permanent occupation, disturbed Lobengula, but Jameson was at his capital, Bulawayo, to put his mind at ease. In the following year Harry Johnston took over the administration of Nyasaland in a dual capacity, as commissioner of the Imperial government and an employee of the chartered company. Although eventually the protectorate reverted fully to the Imperial government, Rhodes's influence was felt both north and south of the River Zambezi, and soon the new territories were called by his name.

Policies as prime minister of Cape Colony. In the meantime, he had returned to office in 1890 in the only post big enough for him, as prime minister of Cape Colony. For five years he proved a successful and imaginative prime minister, though still little known to the world. He acquired a property called Groote Schuur, which he rebuilt in the Dutch colonial style and bequeathed as an official residence for future prime ministers of the Union of South Africa. There he lavishly entertained Dutch and British inhabitants of the Cape Colony and eminent visitors of all nationalities. Everything he undertook was on a massive scale. "I like the big and simple—barbaric if you like," he would say. Men found him both harsh and generous but always pertinacious and persuasive.

In parliament he cultivated the support of the Afrikaner Bond without losing the goodwill of British liberals. His agricultural policies were sensible and effective. In native policy he had to move cautiously. His Strap Bill (1891), which would have permitted the flogging of Africans in certain cases, had to be dropped. The Franchise and Ballot Act (1892) was passed, limiting the native vote by financial and educational qualifications. The Glen Grey Act (1894), assigning an area for exclusively African development, was introduced from the highest motives: "a Bill for Africa," as Rhodes proudly called it. His main aim was to prevent the Dutch and British quarrelling over such policies. To him that involved the risk of "mixing up the native question with the race question" (Rhodes used the word "race" to distinguish the Boers and the British).

Northern policy

He also sought to unite the two white races on his northern policy. The prospects were good because Kruger's obstinacy alienated the Cape Dutch. To ensure that commercial traffic did not have to reach the Transvaal through the Cape Colony, Kruger had built a railway to Delagoa Bay; then in 1894 he closed the "drifts," or fords, of the River Vaal to prevent transport of goods by wagon, besides imposing heavy duties on Cape produce. Rhodes went to the Transvaal capital to protest, but in vain. Kruger was compelled to yield only after a declaration by W.P. Schreiner, Rhodes's attorney general, that he was in breach of the London Convention, coupled with a threat by Joseph Chamberlain, who had become colonial secretary in 1895, to support a military expedition.

Rhodes's patience had begun to wear thin even earlier, partly because he knew his health was precarious, partly because he learned that the gold deposits of the Transvaal were enormous, whereas those of Mashonaland were proving poor. His northern policy was encountering unexpected frustrations. The chartered company was in financial difficulties, its resources being overstretched. Although Rhodes's agents secured some new territories for the company, including Manica (now in Mozambique [Portuguese East Africa]) and Barotseland (now in Zambia), elsewhere he was forestalled. An Anglo-German agreement of 1889 gave a strip of land to Germany, cutting off Bechuanaland from the north. The Belgian king Leopold anticipated Rhodes in laying claim to Katanga (1890), now a province in the Democratic Republic of Congo. The Anglo-Portuguese Convention of 1891 ended his hopes of eliminating Portugal from Africa. Harry Johnston proved uncooperative in administering Nyasaland. When Rhodes paid his first visit to Rhodesia in 1891, he found the pioneers in an angry mood; to pacify them, he helped them generously out of his own pocket.

Serious trouble broke out in 1893, when Lobengula tried to reassert his control over Mashonaland. Jameson, who had been appointed resident commissioner in 1891, resisted the Matabele raiders on his territory. A short, sharp war ended in the total defeat and death of Lobengula, after which Rhodes instructed Jameson to settle and administer Matabeleland. The objections of the "imperial factor" and the missionaries were gradually overcome, and in July ***1894*** the British government gave its sanction to his settlement. Rhodes was then at the pinnacle of his achievement, but still the wider union of southern Africa eluded him. He was growing petulant and impatient and was visibly aging. By ***1895*** he was determined to settle accounts with the last obstacle, President Kruger.

There was already talk of force to remedy the grievances of the Uitlanders in the Transvaal. In ***1894,*** the Cape Colony's new high commissioner, Sir Henry Loch, went up to Pretoria to reason with Kruger, who made it clear that he would never give way. The Uitlanders formed a National Union to support their cause, among its leaders being Rhodes's brother Frank. Kruger sought the support of Germany, and in ***1895*** he again closed the "drifts" across the Vaal. Again he was forced to give way, and by this time a conspiracy against him was under way. Rhodes knew about it and worked actively to foster it. The original intention was that a "spontaneous" revolt in Johannesburg should lead to a military incursion from the Bechuanaland protectorate, followed by a personal intervention by the high commissioner from Cape Town. To this end Rhodes used his influence to have Sir Hercules Robinson returned to the Cape Colony as high commissioner; and through his agents in London he secured Chamberlain's agreement to the immediate cession of a strip of the Bechuanaland protectorate, ostensibly for his railway to the north but in fact to provide a "jumping-off place" for an armed force under Jameson.

Effects of the Jameson Raid on Rhodes's career. Chamberlain was privy to the plan, but no one foresaw what actually resulted. The National Union in Johannesburg lost heart and decided not to act. Rhodes, Robinson, and Chamberlain all assumed that the plan had been called off; but Jameson recklessly decided to force the hand of the Uitlanders by invading the Transvaal on his own. He launched the famous raid on December ***29, 1895.*** It was a fiasco, his whole force being captured, apart from a few killed. Rhodes was compelled to resign all his offices, not only in the Cape government but also in the chartered company; but he refused to denounce Jameson. Official inquiries were held in Cape Town ***(1896)*** and London (1897), blaming Rhodes but tacitly acquitting Chamberlain. Rhodes protected Chamberlain by suppressing the so-called missing telegrams that had passed between them, his purpose being to put moral pressure on Chamberlain not to abrogate his company's charter.

The raid was an almost complete disaster for Rhodes. Jameson and his colleagues were sent to prison; Kruger's power was consolidated; the Dutch and British colonials were more deeply split than ever; Rhodesia and Bechuanaland were taken over by the Imperial government. Only the charter was preserved, and Rhodes spent the rest of his life promoting developments in the north. He even won public sympathy, partly because the German Kaiser had sent Kruger a foolish telegram of congratulations and partly because his own evidence to the inquiry displayed an inspiring conception of imperial destiny. He sought to embody that conception in his will, the seventh and final version of which was signed in ***1899.***

Early in ***1896,*** while Rhodes was in England, there was a serious revolt in Matableland. Rhodes returned by way of Egypt and took an active part in suppressing the revolt. He finally brought it to an end by the technique of holding a peace conference ***(indaba)*** that he had learned from Gordon. On this occasion Rhodes found the site in the Matopo Hills that he called the "View of the World" and chose it for his burial place. Knowing that his days were numbered, he compensated for political eclipse by private activities: developing a large farm at Inyanga, rebuilding Groote Schuur (which was burnt down in 1896), and sending out his young "apostles" to prospect for minerals in the north.

Not only did his old Boer allies in the Cape Colony, such as Hofmeyr and Schreiner, now distrust him but so also did the government in London and its new high commissioner in Cape Town, Sir Alfred Milner. At first Rhodes avoided Milner and resisted all pressures to reenter politics. He spent much time travelling, partly for his health, in Europe and Egypt, always promoting his plans for spanning Africa by railway and telegraph. Gradually, it became apparent that he was still indispensable in the Cape Colony. Milner began to assimilate Rhodes's views and to become even harsher in his judgment of Kruger. In ***1898*** Rhodes assumed the leadership of the Progressive Party, which had been founded to promote the British interest. Though he never held office again, he powerfully influenced successive Cape governments. As war with the Boers of the Transvaal approached, however, Rhodes kept out of the way, advising his friends to "leave it to Milner."

Boer War and Rhodes's death

He insisted that Kruger would "climb down" in the end, perhaps hoping that he was wrong. When the Boer War broke out in October ***1899,*** Rhodes spent the first few months under siege at Kimberley. After a spirited defense, during which he assumed the rank of colonel, he emerged early in ***1900*** confident that the war would soon be over. He spent five months in Rhodesia planning for the future and pressing on Milner schemes of British settlement and reconciliation with the Boers.

His last years were soured by an unfortunate relationship with an aristocratic adventuress, Princess Radziwiłł, who sought to manipulate Rhodes and Milner and even Lord Salisbury, the English prime minister, to promote her ideas of the British Empire. Rhodes was unused to scheming women, nor could the young bachelors surrounding him protect him from her. She forged letters and bills of exchange in his name and was finally sent to prison, but not before she had caused him much annoyance and scandal. In ***1901,*** while he was in Europe, he was recalled to Cape Town to give evidence at her trial. His last political act on his return was to support Milner in suspending the constitution of the colony until the war was over. He was, however, already dying of an incurable heart disease. Before either the Boer War or even Princess Radziwiłł's trial was over, he died on March ***26, 1902.*** Two weeks later he was buried in the Matopo Hills. His last journey through Africa in the funeral train was a triumphal procession.

Reputation

When Rhodes's will was read in April ***1902,*** his reputation immediately rose to new heights. The imaginative scheme of scholarships at Oxford for young men from the colonies and from the United States and Germany appealed to the public instinct for a more disinterested kind of imperialism. Most of his fortune (over £3,000,000 when the debts and legacies were paid) was devoted to this purpose. As the will forbade disqualification on grounds of race, many nonwhite students have benefitted from the scholarships, though it is doubtful if that was Rhodes's intention. He once defined his policy as "equal rights for every white man south of the Zambezi," and later, under liberal pressure, amended "white" to "civilized." But he probably regarded the possibility of native Africans becoming "civilized as so remote that the two expressions, in his mind, came to the same thing.

BIBLIOGRAPHY. J.G. LOCKHART and C.M. WOODHOUSE, ***Rhodes*** (1963), the most up-to-date biography, based on the Rhodes papers; S.G. MILLIN, ***Rhodes,*** rev, ed. (1952), a work, by a South African woman, that supersedes all earlier biographies; "VINDEX," ***Cecil Rhodes, His Political Life and Speeches,*** *1881–1900* (1900), essential to an understanding of Rhodes's style of thought and expression; J. VAN DER POEL, ***The Jameson Raid*** (1951), a definitive account of the crucial episode in Rhodes's career; BRIAN ROBERTS, ***Cecil Rhodes and the Princess*** (1969), interesting mainly for the account of the last few months of Rhodes's life.

(*C.M.*Wo.)

Rhodesia (Zimbabwe)

Rhodesia, since ***1980*** officially named Zimbabwe, is a republic in southern Africa. It achieved majority rule and internationally recognized independence in April ***1980,*** after a long armed struggle against the white minority regime that in November ***1965*** had made a unilateral and unconstitutional declaration of independence from Great Britain. Although political power passed at independence to the representatives of the country's ***7,200,000*** blacks, the economic role of its ***230,000*** whites remained dominant. The new government committed itself to the ideal of building a multiracial state. Many changes were expected in the early years of independence in the process of land resettlement and economic reconstruction after a profoundly disruptive civil war.

Zimbabwe is landlocked. It shares a 125-mile (200-kilometre) border on the south with South Africa, and is bounded on the southwest and west by Botswana, on the north by Zambia, and on the northeast and east by Mozambique. Its total area is 150,804 square miles (390,-580 square kilometres). Its capital, Salisbury, has a population of more than 700,000, swollen by the influx of wartime refugees from rural areas.

The nationalist struggle

The three-sided quarrel between the European Rhodesian leaders, successive British governments, and the African Zimbabwean nationalists has overshadowed most aspects of contemporary life in the country. While the history of this quarrel goes back to the occupation of the region in 1890 and to the rebellions of 1893–97 in both Mashonaland and Matabeleland, its immediate roots stem from the circumstances attending the dissolution of the Federation of Rhodesia and Nyasaland in 1963. Southern Rhodesia had been a self-governing colony, never directly administered by Great Britain, when it entered this federation in 1953. Its premier, Winston Field, and his successor, Ian Smith, demanded as the price of acquiescence in the dissolution of the federation by the British Conservative government that Southern Rhodesia achieve full independence (as Rhodesia) no later than Northern Rhodesia (as Zambia) or Nyasaland (as Malaŵi). The British government was not satisfied that the Rhodesian administration, with an electorate of 85,000 whites and only 12,000 blacks, was a sufficiently representative government. It countered with a statement of five principles on which an independence constitution should be based; the most important of these was "unimpeded progress to majority rule." Fruitless negotiations continued for nearly two years, until the government of Ian Smith made its unilateral declaration of independence (UDI) in 1965.

The quarrel was soon afterward enlarged to engage the United Nations. The United Kingdom, while arguing that UDI was an act of rebellion with which it alone retained authority to deal, asked UN member states not to recognize Rhodesian independence and to apply voluntary economic sanctions. It later persuaded the UN Security Council to impose first selective, and later comprehensive, mandatory sanctions.

Three separate attempts by the British government to reach a political settlement with the Smith regime on the basis of its five (later six) principles filled the first half of the period of UDI. The last attempt, in which a British commission in 1972 toured Rhodesia to judge the acceptability of its settlement proposals to the people as a whole, proved a turning point in the struggle. The overwhelming rejection of those proposals provided an impetus for a revived nationalist movement. At the same time, advances by the nationalist forces fighting against the Portuguese in Mozambique opened up terrain favourable for Zimbabwean guerrilla infiltration. Although the British government, helped actively by U.S. negotiators, continued to press political proposals, during the next seven years of UDI it was the increasing effectiveness of the guerrilla armies operating from bases in Mozambique and Zambia that drove the Smith regime to concede majority rule. This came at a cost of some 27,000 lives and more than 1,000,000 refugees. (For further coverage of historical aspects, see SOUTHERN AFRICA, HISTORY OF.)

THE LANDSCAPE

Relief. Zimbabwe lies almost entirely over 1,000 feet (300 metres) above sea level. Its principal physical feature is the broad ridge running 400 miles from southwest to northeast across the whole country, from Plumtree through Gwelo and Marandellas to Inyanga. About 50 miles wide, this ridge ranges in altitude from 4,000 to 5,000 feet, until it eventually rises to 8,503 feet at Mt. Inyangani, the highest point in Zimbabwe, in the eastern highlands. This ridge is known as the Highveld and comprises about 25 percent of the country's total area of 150,804 square miles. On each side of this central spine, sloping down northward to the Zambezi River and southward to the Limpopo River, lies the wider plateau of the Middle Veld, which, at an altitude between 3,000 and 4,000 feet, makes up about 40 percent of the area of Zimbabwe. Beyond this again, and mostly in the south, whrre the Sabi, Lundi, and Nuanetsi rivers drain from the plateau into the Limpopo, lies the Lowveld, which comprises about 23 percent of the total area. The lowest point in Zimbabwe lies at an altitude of 660 feet near Dumela, where the Limpopo Rows down toward Mozambique. There are no parts of Zimbabwe that can properly be called desert, although a sector northwest of Plumtree along the Botswana border and a lengthy belt across the Lowveld in the south are severely arid.

Geology

A characteristic of the landscape is its Precambrian rock, which is between 570,000,000 and 4,600,000,000 years old. The most ancient part of this rock formation, known as the Basement complex, covers the greater part of the country. About four-fifths of the Basement complex consists of granite; the Matopo Hills south of the city of Bulawayo are formed of a hard outcrop of granite and gneiss. These outcrops, known as balancing rocks, have been severely weathered by wind and water, leaving some blocks precariously balanced upon others. Elsewhere are found innumerable rounded hillocks known as kopjes and composed of ball granite. Belts of schist in the Basement complex contain the veins and lodes of most of the country's gold, silver, and other minerals.

The Great Dyke, which is up to eight miles wide and about 320 miles long, is another notable feature. The Alkali Ring complexes are igneous intrusions of volcanoes; they occur in the Sabi Valley and near Beitbridge. The Karroo System—a thick layer of sedimentary rocks consisting of shale, sandstone, and grit—covers the Zambezi Valley and the valleys of its tributaries from Wankie southward to Bulawayo and spreads across parts of the southern Lowveld from Tuli, near the southern border, to the Sabi River.

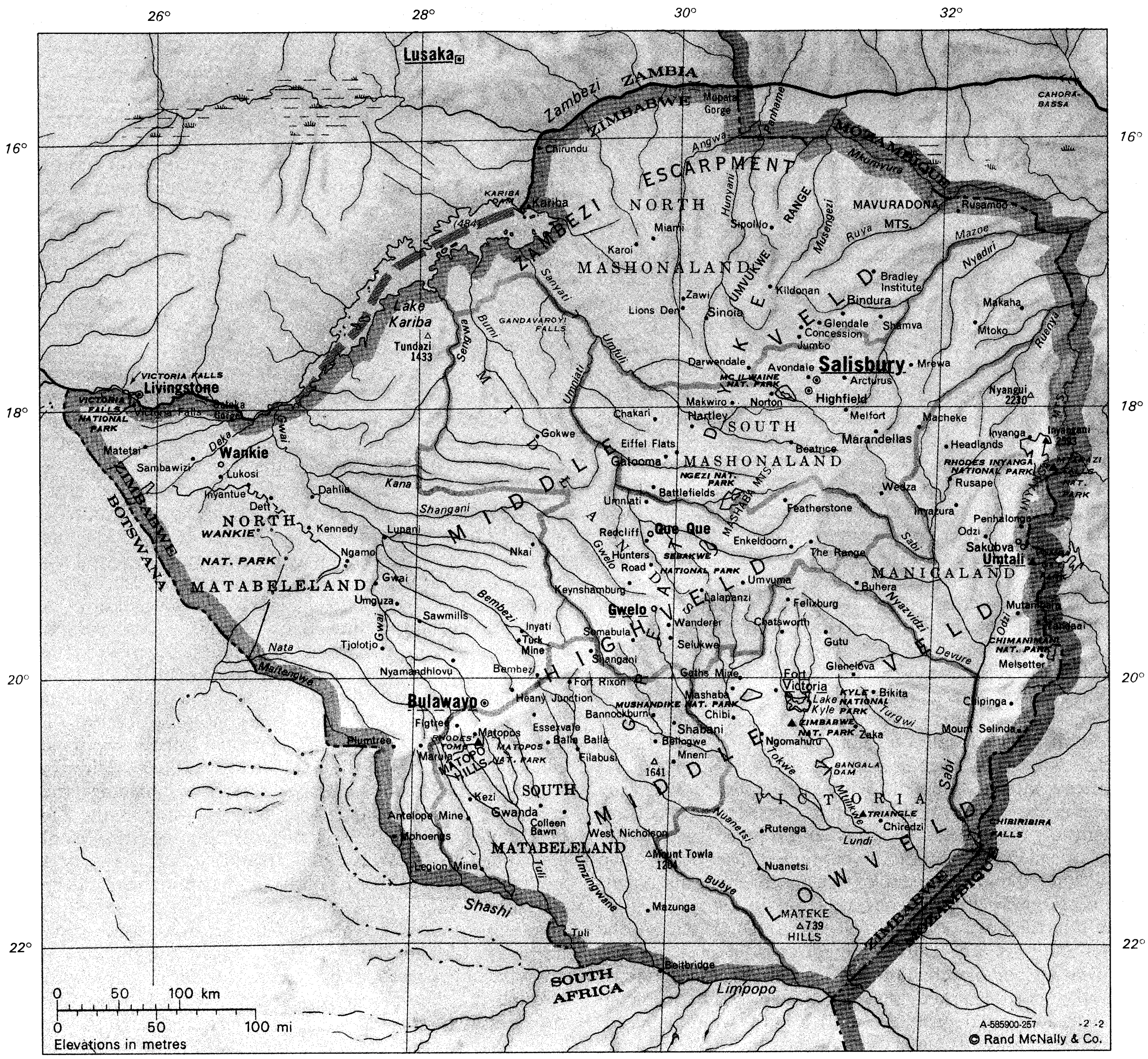
Drainage and soils. Major faulting from southwest to northeast formed the middle Zambezi trough, which is now partially flooded by Lake Kariba. Other faulting affected the depressions of the Sabi and Limpopo rivers. As a result of subsidiary rifting a few sizable rivers drain north and south from the Plumtree–Inyanga watershed.

Low fertility of soils

The light, sandy soils developed on granite rocks are highly weathered and leached, even in the areas of lower rainfall, and do not easily retain water. Since the bulk of the rain occurs in heavy showers during a few months of the year, problems of drainage and of retaining nutrient reserves also occur. The meagre mineral reserves in the soils imply an inherently low fertility; under cultivation, productivity drops rapidly. The difficulty of cultivating these lighter soils is greatest in the black farming areas, where population pressure no longer allows land to be temporarily abandoned after cultivation; black farmers, because of a lack of capital, are also less able than white farmers to maintain the mineral fertility with manure and chemical fertilizers. (For associated physical features, see VICTORIA FALLS; ZAMBEZI RIVER.)

Climate. Zimbabwe lies north of the Tropic of Capricorn but enjoys subtropical conditions because of its altitude. Toward the end of the hot, dry months, which last from August to October, monsoon winds that have crossed the Indian Ocean and Mozambique cause rainfall when they meet the rampart formed by the eastern highlands. The eastern districts consequently receive the heaviest rainfall and have a more prolonged rainy season (lasting from October into April) than the rest of Zimbabwe. The altitude of the broad plateau of western Zimbabwe helps to guarantee fine weather for most of the country during the cool, dry winter months of May to August.

June is generally the coolest month and October the warmest; temperature variations correspond closely to altitude. Inyanga, at about 5,500 feet in the eastern highlands, varies in temperature from a mean of 52" F (11° Cj in July to one of 65° F (18° C) in October. Salisbury, at about 4,800 feet, has temperatures varying from 57° F



ZIMBABWE 26 28'' 30^c 32▪

MAP INDEX

Political subdivisions

Manicaland	1930s	32·15e
Midlands	19·00s	29·45e
North Mashonaland	16·30s	30·00e
North Matabeleland	1930s	28·00e
South Mashonaland	18·15s	30·45e
South Matabeleland	2000s	28·15e
Victoria	20·54s	31·21e

Cities and towns

Antelope Mine	21·02s	28·27e
Arcturus	1747s	31·20e
Avondale	17·43s	30·58e
Balla Balla	20·26s	29·02e
Bannockburn	20·16s	29·51e
Battlefields	18·31s	29·52e
Beatrice	18·15s	30·55e
Beitbridge	22·13s	30·00e
Beiingwe	20·30s	29·53e
Bembezi	20·00s	28·56e
Bikita	20·06s	31·41e
Bindura	17·19s	31·20e
Bradley Institute	17·02s	31·27e
Buhera	19·18s	31·29e
Bulawayo	20·09s	28·36e
Chakari	18·05s	29·51e
Chatsworth	19·38s	31·13e
Chibi	20·19s	30·30e
Chipinga	2012s	32·38e
Chiredzi	21·03s	31·45e
Chirundu	1559s	28·54e
Colleen Bawn	21·00s	29·13e
Concession	17·22s	30·57e
Dahlia	18·35s	27·08e
Darwendale	1743s	30·33e
Dett	18·38s	26·50e
Eiffei Flats	18·15s	29·59e
Enkeldoorn	19·01s	30·53e
Essexvale	20·18s	28·56e
Featherstone	18·42s	30·49e
Felixburg	19·29s	30·51e
Figtree	20·24s	28·21e
Filabusi	20·34s	29·20e
Fort Rixon	20·01s	29·18e
Fort Victoria	20·05s	30·50e
Gaths Mine	20·00s	30·31e
Gatooma	18·21s	29·55e
Glenclova	19·59s	31·26e
Glendale	17·21s	31·04e
Gokwe	1807s	28·58e
Gutu	19·38s	31·10e
Gwai	1915s	27·42e
Gwanda	20·57s	29·01e
Gwelo	19·27s	29·49e
Hartlev	18·10s	30·14e
Headlands	1814s	32·03e
Heany Junction	20·06s	28·54e
Highfield	17·50s	30·59e
Hunters Road	19·09s	29·48e
Inyanaa	1813s	32·46e
Inyantue	18·32s	26·41e
Inyati	19·39s	28·54e
Inyazura	18·43s	32·10e
Jumbo	17·28s	30·55e
Kariba	16·30s	28·45e
Karoi	16·50s	29·40e
Kennedy	18·52s	27·10e
Keynshamburg	19·15s	29·39e
Kezi	20·58s	28·32e
Kiidonan	17·21s	30·37e
Lalapanzi	19·16s	30·15e
Legion Mine	21·23s	28·33e
Lions Den	17·16s	30·02e
Lukosi	18·30s	26·30e
Lupani	18·54s	27·44e
Macheke	18·05s	31·51e
Makaha	17·17s	32·37e
Makwiro	17·58s	30·28e
Marandelias	18·10s	31·36e
Marula	20·26s	28·06e
Mashaba	2002s	30·29e
Matetsi	18·16s	25·56e
Matopos	20·24s	28·28e
Mazunga	21·45s	29·52e
Melfort	17·59s	31·19e
Melsetter	19·48s	32·50e
Miami	16·40s	29·46e
Mneni	20·38s	30·03e
Mount Selinda	20·25s	32·43e
Mphoengs	21·10s	27·51e
Mrewa	17·39s	31·47e
Mtoko	17·24s	32·13e
Mutambara	19·36s	32·33e
Ngamo	19·08s	27·32e
Ngomahuru	20·26s	30·43e
Nkai	19·00s	28·54e
Norton	17·53s	30·42e
Nuanetsi	21·22s	30·45e
Nyamandhlovu	19·50s	28·16e
Odzi	18·58s	32·23e
Penhalonga	18·54s	32·40e
Plumtree	20·30s	27·50e
Que Que	18·55s	29·49e
Redcliff	19·02s	29·46e
Rusambo	16·35s	32·12e
Rusape	18·32s	32·07e
Rutenga	21·08s	30·45e
Sakubva	18·57s	32·39e
Salisbury	17·50s	31·03e
Sambawizi	18·21s	26·16e
Sawmills	19·31s	28·02e
Selukwe	19·40s	30·00e
Shabani	20·20s	30·02e
Shamva	17·18s	31·34e
Shangani	19·47s	29·22e
Sinoia	17·22s	30·12e
Sipoiilo	16·39s	30·42e
Somabula	19·41s	29·41e
Tandaai	19·36s	32·48e
The Range	19·00s	31·04e
Tjolotjo	19·47s	27·46e
Tuii	21·59s	29·15e
Turk Mine	19·45s	28·50e
Umguza	19·25s	27·51e
Umniati	18·39s	29·49e
Umtali	18·58s	32·40e
Umvuma	19·19s	30·35e
Victoria Falls	17·56s	25·50e
Wanderer	19·37s	29·59e
Wankie	1822s	26·29e
Wedza	18·35s	31·35e
West Nicholson	21·06s	29·25e
Zaka	2020s	31·29e
Zawi	17·13s	30·02e

Phvsical features and points of interest

Angwa, river	15·51s	30·35e
Bangala Dam	20·40s	31·15e
Batoka Gorge	17·56s	26.15e
Bembezi. river	18·57s	27·47e
Bubye. river	22·20s	31·07e
Bumi. river	16·50s	28·26e
Chibiribira Falls. waterfalls	21·14s	32·20e
Chimanimani National Park	20·00s	32·20e
Deka. river	18·03s	26·44e
Devure. river	1950s	31·45e
Gandavaroyi Falls waterfalls	1717s	29·07e
Great Dyke, hills	20.00s	30·00e
Gwai. river	1759s	26·52e
Gweio. river	1845s	28·36e
Highveld. physical region	19·30s	29·50e
Hunyani. river	16·00s	30·35e
Inyanga Mountains	18·00s	33·00e
Inyangani. mounfain	18·20s	32·50e
Kana. river	18·30s	27·22e
Kariba. Lake	1700s	28·00e
Kariba Dam	16·30s	28·50e
Kyle. Lake. reservoir	20·14s	31·00e
Kyle National Park	20·14s	31·00e
Limpopo, river	25·15s	33·30e
Lowveld, physical region	21·30s	31·30e
Lundi. river	21·43s	32·34e
Mcllwaine National Park	17·55s	30·50e

MAP INDEX (continued)

Maitengwe, *river*	19·57s	26·40e
Mashaha Mountains	18·45s	30·32e
Mateke Hills	21·48s	31·00e
Matopo Hills	20·36s	28·28e
Matopos National Park	20·33s	28·20e
Mavuradona Mountains	16·30s	31·20e
Mazoe, river	16·32s	33·25e
Middle Veld, *physical* region	18·30s	28·30e
	20·30s	30·30e
Mkumvura, river	15·55s	31·07e
Mtarazi Falls National Park	18·28s	32·47e
Mtilikwe, river	21·09s	31·30e
Mupata Gorge	15·38s	30·05e
Musengezi, river	15·55s	31·05e
Mushandike National Park	20·10s	30·38e
Nata, river	20·14s	26·10e
Ngezi National Park	18·40s	30·28e
Nuanetsi, river	22·40s	31·50e
Nyadiri, river	16·44s	32·33e
Nyangui, mountain	17·53s	32·44e
Nyazvidzi, river	20·00s	32·17e
Odzi, river	19·45s	32·24e
Rhodes Inyanga National Park	18·12s	32·45e
Rhodes Tomb	20·30s	28·30e
Ruenya, river	17·15s	32·58e
Ruya, river	16·35s	33·13e
Sabi, river	21·00s	35·02e
Sanyati, river	16·49s	28·45e
Sebakwe National Park	19·00s	30·14e
Sengwa, river	17·07s	28·05e
Shangani, river	18·41s	27·10e
Shashi, river	22·14s	29·20e
Tokwe, *river*	21·09s	31·30e
Towla, Mount, *mountain*	21·22s	29·52e
Triangle, ranch	21·00s	31·26e
Tuli, river	21·48s	29·04e
Tundazi, mountain	17·33s	28·05e
Turnwi, river	20·27s	32·17e
Umfuli, river	17·30s	29·23e
Umniati, river	17·30s	29·23e
Umvukwe Ranae	1720s	30·40e
Umzingwane, river	22·12s	29·56e
Victoria Falls, waterfalls	17·55s	25.21e
Victoria Falls National Park	17·55s	25·40e
Vumba National Park	19·05s	32·45e
Wankie National Park	19·00s	26·35e
Zambezi, river	18·50s	36·17e
Zarnbezi Escarpment	16·20s	30·00e
Zimbabwe Nationai Park	20·17s	30·57e

(14" C) to 70° F (21° C), and Bulawayo, at 4,400 feet, has temperatures varying from 58° F (14° C) to 73° F (23° C). On a 20-year average, Salisbury and Bulawayo have had a year-round daily mean of eight hours of sunshine, and this average does not drop below six hours during the rainy season. (For rainfall patterns, see below, Traditional regions.)

Savanna country

Vegetation. Zimbabwe is predominantly savanna (tropical grassland) country, with a generous tree growth encouraged by the wet summers. The only true forests, however, are the evergreen forests of the eastern border and the savanna woodland, which includes teak, northwest of Bulawayo. Various species of Brachysregia (a hardwood tree up to 90 feet high with pale reddish-brown wood) are dominant in the Middle Veld and Highveld. Other common varieties include the mohobohobo (a medium-sized tree with large spadelike leaves) and the thorn tree. In the valleys of the Zambezi and Limpopo rivers the mopani, which resembles the mohobohobo, is common, together with baobab and the knobby thorn tree. Australasian eucalyptus trees have been widely introduced, predominantly on white-owned farms, where they are used as windbreaks and for fuel; Australian wattle has been planted in the eastern districts as a source of tannin. Pure grassland is uncommon but occurs particularly along the eastern border around Melsetter. Swamps are widespread on the Botswana border.

Wankie National Park

Animal life. Cultivation of the land has resulted in the disappearance of many forms of animal life from large areas. Wankie National Park has an area of more than 5,000 square miles and stretches from the Bulawayo–Victoria Falls railway line westward to the Botswana border. Among the flesh-eating animals found there, and occasionally elsewhere, are the lion, leopard, cheetah, serval, civet, aardvark, spotted and brown hyena, black-backed and side-striped jackal, zorilla, ratel, bat-eared fox, ant bear, and scaly anteater. Elephants are found in the northern region and giraffes in the western bushland; hippopotamuses and crocodiles live in the larger rivers. Among a great variety of hoofed and horned ruminant animals are the eland (which is immune to the deadly tsetse fly), greater kudu, blue duiker, impala, klipspringer, steenbok and grysbok, and sable and roan antelope. Snakes include mambas, boomslangs, and the black-necked cobra. Baboons, which are the bane of farmers whose crops they damage, include the Rhodesian and yellow species, as well as the chacma, the largest known baboon species. Notable among the birdlife are the martial eagle, the bateleur eagle, and the little hammerhead, which builds enormous nests and is revered as a bird of omen.

Traditional regions. Zimbabwe may be divided into six-different regions, with the amount of rainfall constituting the determining factor in land use. Of the nation's 96,500,000 acres. some 1,500,000 acres (607,000 hectares) in the eastern highlands, with more than 25 inches of rainfall annually, are suitable for diversified farming with cattle and plantation and orchard crops. A further 18.000~000acres sweeping west along the centra! spine past Salisbury and to the midlands receive 20 to 25 inchcs of rain and are used for intensive farming of corn (maize) and tobacco and the raising of livestock. An almost equal area to the southwest, enclosing Bulawayo, receives 16 to 20 inches of rain a year; it is suitable for mixed farming and for raising livestock on a semi-intensive scale. One-third of the country, lying farther outward from the spinc of Zimbabwe, mostly to the south, and receiving 14 to 18 inches of rain a year, is used for semi-extensive farming, while 25,000,000 acres in the Lowveld toward the Limpopo and Zambezi rivers, receiving less than 16 inches a year, are fit only for ranching. Finally, some 3,000,000 acres, mostly toward the Zanibezi River, are unsuitable for either agriculture or forestry.

Land ownership

Racial segregation of land ownership persisted in prac-·~k cif not in law, from the 1890s into the early years of independence. Of the 35 percent of the total land area considered suitable for intensive or semi-intensive farming. three-fifths was in the hands of white farmers and the remaining two-fifths belonged to blacks. Of the 59 percent that was fit for semi-extensive or extensive farming, each racial group had nearly a half-sharc.

The landscape under human settlement. At independence, settlement patterns in Zimbabwe reflected the segregation of land ownership first established after 1890 and formalized in the Land Apportionment Act of 1930 and the Land Tenure Act of 1969. Thc latter law apportioned the nation's lands almost equally in size between whites and blacks and set aside a national area for parks and game reserves. Of the 44,962,000 acres (18,191,000 hectares) of land assigned to the black population, 39,922,100 acres were designated as Tribal Trust Lands and 3,670,400 acres as African Purchase Areas.

The black people of Zimbabwe are still predominantly rural. In 1975, the last year before the civil war brought widespread disruption in rural areas, 83 percent lived either in their own village communities or in compounds on white-owned farms. Housing and living conditions in the Tribal Trust Lands greatly worsened during the 1970s, as the population grew to 4,500,000 on some 40,000,000 acres of such poor fertility that they could not adequately support more than 1,000,000 people. Less than 2 percent of all black rural households had electricity, and four houses out of five were still built of pole and dakha (wattle and mud) material. Access to safe water was rare because most families in the Tribal Trust Lands drew their supplies from shallow wells or unprotected surface sources. A large proportion of these families survived with the help of remittances from relatives in the agricultural cash economy or workers on white-owned farms or in towns; they could not aspire to buy land in the 66 African Purchase Areas, where hundreds of the 8,000 farms remained vacant. The purchase areas were themselves producing only about one-quarter of what was cropped from the same amount of white-owned land.

The Rhodesian Front came to power in 1962 with the promise to maintain land segregation, despite the evidence gathered in 1960 by a Rhodesian parliamentary committee that the black reserves could not support all the families with land rights under the Native Land Husbandry Act of 1951. The Front put aside the implementation of this act and attempted to establish racial division "for all time" with the Land Tenure Act of 1969. In consequence, the nationalist struggle focussed sharply upon the issue of land ownership, and a major concern for the Zimbabwe government after independence was to carry through land reform and launch large-scale settlement of black families on former white farms.

The Land Apportionment Act made no provision for blacks who chose an urban life, for towns were designated as white areas. As a result, though urban blacks now outnumber whites by four to one, blacks mostly live in rented homes in townships located some miles from city centres. The cities of Salisbury and Bulawayo therefore constitute studies in contrast, with impressive office buildings and quiet white suburbs partially ringed by crowded black townships where there is an average of two people to each room. The Land Tenure Act was amended, while the civil war was still being fought, to allow blacks to purchase white farms and urban property, and after the end of hostilities residential segregation began to be significantly breached.

PEOPLE AND POPULATION

Linguistic, ethnic, and religious groups. The majority of the population of Zimbabwe speaks Shona; these are nearly four times the number of those who speak Ndebele as their first language. Both Shona and Ndebele are Bantu languages; from the time of their great southward migration, Bantu-speaking tribes have populated what is now Zimbabwe for more than 10 centuries. Those who speak Ndebele are concentrated in a circle radiating from Bulawayo, with Shona-speaking peoples beyond them on all sides—the Kalanga to the southwest, the Karanga to the east around Fort Victoria, the Zezuru to the northeast, and the Rozwi and Tonga to the north. Generations of intermarriage have to a degree blurred the linguistic division between the Shona and Ndebele peoples.

The Shona and Ndebele

Among the 230,000 whites in Zimbabwe at independence were the descendants of the country's first European immigrants. Only about one-quarter of the adult white population, however, was born in Zimbabwe. Since World War II the white population has been trebled by heavy immigration, and more than two-thirds of the present white population have their origins in Europe, the great majority from Britain. The rest have come from South Africa.

The white population is, nevertheless, outnumbered by Zimbabwe's 7,200,000 blacks by a ratio of one white for every 31 blacks. There are nearly 11,000 Asians, forming a community that is predominantly concerned with trade. There are also 24,000 Zimbabweans of mixed race, in Zimbabwe called Coloureds, who are mainly skilled and semiskilled workers. More than 80 percent of the whites live in towns, whereas only 17 percent of the black population do so. Of the whites living in rural areas, about one-quarter are Afrikaners. English is the language of government; teaching in schools is also conducted in English, except for the instruction of the youngest children in black schools.

Traditional religion

The great majority of the black population adheres to traditional religion based on reverence for ancestors. The Shona have preserved their ancient reputation for prophecy, divination, and rainmaking; they believe in Mwari, a supreme being. The stone ruins of Great Zimbabwe are regarded as a shrine of deep religious significance, as also are parts of the Matopo Hills. In the last 50 years mission schools have exercised much influence, and most of the members of the first Cabinet of independent Zimbabwe were graduates of these schools. The strongest influence has been that of the Roman Catholic Church, with some 800 schools and more than 500,000 black members and 37,000 white members. In comparison, the Anglican Church has some 150,000 black and 80,000 white adherents. The Methodist, Presbyterian, Baptist, and Dutch Reformed churches are also represented. Because the Catholic Church supported nationalist aspirations, it held a position of influence in the post-independence period.

Demography. *Birth and mortality rates.* The annual rate of population increase among Zimbabwean blacks is one of the world's highest, at 35 persons per 1,000. The death rate of 13 per 1,000 in 1978 had declined only fractionally in a decade, while the birth rate of 48 per 1,000 remained steady. At these rates, the black population would double in 20 years. Approximately 20 percent of the black population is under five years of age, and 51 percent is under 15. The average black adult in Zimbabwe has to support more than three times as many children as his counterpart in a developed country.

Among whites, there is about one child for every two adults. The death rate among whites has been eight per 1,000, and their life expectancy is 72 years, compared with 50 years for their black compatriots. The infant mortality rate among whites is 16.8 per 1,000, compared with about 32 per 1,000 among urban blacks and about 160 per 1,000 among rural blacks.

Immigration and emigration. Migration has been the most important factor influencing the size and composition of the white population. Net migration figures have fluctuated in reaction to political events, but most striking is the extremely high turnover of the 1960s and 1970s. In the years immediately preceding the breakup of the Federation of Rhodesia and Nyasaland, there was a net emigration of 13,000 whites; this was followed during the first 10 years of UDI by a net immigration of 40,000 (with 111,270 immigrants and 71,330 emigrants). As warfare spread after 1976, the pendulum swung again from a peak white population of 260,000 to fewer than 200,000 after independence. These net figures obscure, however, the gross turnover during 1965–79 of 132,560 immigrants and 133,864 emigrants. Even when allowance is made for the subsequent return of some emigrants, it is probable that at least half of the country's adult whites were newcomers after UDI.

White migration patterns

About 300,000 blacks in Zimbabwe were born outside the country, mostly in Malaŵi and Mozambique. Men from those countries came to Rhodesia to work in the mines and on white-owned farms. Since UDI there has been a net emigration of nearly 91,000 persons categorized in official statistics as "foreign African adult males."

Distribution of population. About 20 percent of the total population lives in Zimbabwe's 14 urban centres, nearly three-fourths of them in either Salisbury or Bulawayo. This total urban population includes 80 percent of the whites, most of the Asians and Coloureds, but only 17 percent of the blacks. Of the other 83 percent of the nation's blacks, three-fourths live in the Tribal Trust Lands or African Purchase Areas and the rest on white-owned farms. The population density of the areas designated as "European" is thus only about one-fourth of what it is in the Tribal Trust Lands. Among urban blacks, there is a disproportionately high number of males of working age, leaving an excess of older people, women, and children in rural areas. At least half of the black households are partly or wholly dependent on incomes earned in the wage economy—almost always from white employers.

Official government figures of December 1979 for the population of the main urban centres apparently did not include rural refugees. These figures were: Salisbury, 641,000; Bulawayo, 375,000; Gwelo, 72,000; Umtali, 65,000; and Que Que 51,000.

Demographic trends. At the present rate of natural increase of 3.2 percent a year, the black population will have grown by another 7,000,000 to more than 14,000,000 by the year 2000. The renewal of family-planning programs in the 1980s and the raising of living standards may considerably affect the birth rate. The trend in the white population will be obscure for several years after independence; the uncertainties that independence has posed for the younger whites make any forecast particularly vulnerable to error.

Zimbabwe, Area and Population

	area		population	
	sq mi	sq km	1969 census	1980 estimate
Provinces*				
Manicaland	13,845	35,859	773,500	. . .
Midlands	21,176	54,845	743,800	. . .
North Mashonaland	26,787	69,378	740,500	. . .
North Mataheleland	30,016	77,741	602,100	. . .
South Mashonaland	16,491	42,711	1,135,200	. . .
South Matabeleland	20,983	54,346	367,100	. . .
Victoria	21,506	55,700	734,100	. . .
Total Zimbabwe	150,804	390,580	5,099,300†	7,360,000

"Although the seven provinces listed as first-order subdivisions have no place in the structure of the government, they occasionally serve administrative functions and are useful divisions for presenting statistical data. †Includes 3,000 travellers.
Source: Official government figures.

THE NATIONAL ECONOMY

By the index of gross national product (GNP) per capita, Zimbabwe is among the world's middle-income countries. According to the 1980 economic survey prepared by the Ministry of Finance, the GNP in 1979 was Z$2,584,000,000; this represents a GNP per capita figure of about U S. $510. The only African countries south of the Sahara with a higher or comparable per capita GNP are Gabon, South Africa, South West Africa (Namibia), the Ivory Coast, Nigeria, and Zambia. As far as Zimbabwe is concerned, however, such a comparison should be treated with caution for two reasons. First, Zimbabwe still contains what are essentially two economies—a white one and a black one—and a combined figure distorts the reality of this dualism. Second, the country's economy was under siege for 14 years of United Nations sanctions, and these siege conditions led to the creation, in turn, of a protective wall behind which import-substitution industries thrived in an abnormally favourable climate. It will be some years before it is clear whether the Zimbabwean economy will benefit from the foreign investment and access to foreign markets that came with the lifting of sanctions, or if it will suffer from greater import competition. Further uncertainties surround the patterns of employment and wages in the mining and manufacturing industries and in the farming sector, as the pressures following independence for higher minimum wages find a response, and as the government makes plain how strongly it wishes to retain skilled white labour.

Sources of national income. ***Agriculture, forestry, and fisheries.*** Agriculture contributed nearly one-fifth of Zimbabwe's national income until about 1973, when the war caused its contribution to begin a decline to about 12 percent by 1979. Even in years of drought, as in 1979–80, agriculture still provided one-third of the country's total foreign exchange and about 35 percent of total employment in the wage economy for some 344,000 workers. Neither forestry nor fisheries had become significant to the economy by 1980.

Tobacco production

Economic sanctions over the 14 years of UDI had a broad effect on patterns of commercial agriculture. Tobacco, primarily the Virginia variety that is flue-cured and grown on white-owned farms, accounted for more than half of the gross value of output in 1965 but fell below 20 percent during UDI. There were 1,750 white farmers still growing tobacco at the time of independence, however, and the flue-cured production in 1979 of 111,700 metric tons was the highest since 1965. Tobacco faced an uncertain future because of the loss of purchasers during UDI, and in 1980 sales were sought in new markets.

Livestock—mainly beef and milk production—became more valuable than tobacco, but the cattle industry was hurt in the later stages of the war. Disease spread because of breakdown in dipping services, and the national beef breeding herd was reduced by more than 30 percent. Recovery to a planned level of some 6,000,000 head was hampered by cattle rustling.

Soya bean and cotton production have increased rapidly, and cotton is likely to become a major cash crop for smallholders in the Middle Veld.

Maize (corn) production was high enough in the mid-1970s to permit export but slumped so low in the late 1970s that the country imported maize from South Africa. Although the collapse of tribal agriculture during the war, one reason for the slump, may be remedied in the early 1980s, commercial farmers found maize to be an unprofitable crop and cut their acreage. Because of this and the high rate of population increase, Zimbabwe is unlikely to be much more than self-sufficient in this staple for some time. Wheat was a minor crop in 1965, but by the late 1970s the country was able to export to neighbouring territories. Sugar output in the Lowveld toward the Limpopo was expected to increase with the lifting of sanctions and the rise in world prices. Other crops that thrived during the diversification prompted by the sanctions were coffee and various vegetables.

The biggest question for agriculture was double-barrelled: how many white farmers were likely to leave the country in the aftermath of independence, and how long would it take black farmers to recover higher production in tribal lands and to move in significant numbers into the commercial sector. The country is dependent upon a small group of white farmers. During UDI, their numbers decreased by 900, and at independence only 5,400 whites were actively farming, In comparison, there were some 700,000 black farming families.

Mining and quarrying. Although mining accounted for only 8 percent of the gross domestic product (GDP) and provided work for only 6 percent of the employed labour force in 1979, its significance in the economy was considerable as a major earner of foreign exchange. Direct mineral exports accounted in 1979 for one-third of total value. The mining sector was expected to attract large-scale foreign investment in the 1980s.

It was the prospect of great mineral wealth—comparable to the gold deposits of the Witwatersrand in neighbouring South Africa—that attracted the first permanent European settlers in the 1890s. These great expectations faded for many years after the peak of gold production was reached in 1915. By the 1950s, however, production of the chrome mines along the Great Dyke was significant, as was that of asbestos and copper. During UDI, the value of mining output increased fivefold to Z$193,000,000 in 1979. The rise in gold prices in the 1970s revived gold as the country's leading export and led to the reopening in 1979–80 of more than 100 dormant mines. Nickel mining, which began in the late 1960s, was carried out in 1980 at five mines; the Great Dyke may indeed contain more nickel than chrome ore. The more than 20 known coalfields in Zimbabwe contain proved reserves of 500,000,000 tons of salable coal.

Manufacturing. Manufacturing has been the fastest growing sector of the economy. From 1954 to 1963, Southern Rhodesia was able to rely on the resources and larger market of the Federation of Rhodesia and Nyasaland for a 150 percent increase in manufacturing output. Then, after UDI in 1965, hundreds of new manufacturing projects were begun in the effort to defeat economic sanctions by import substitution. By 1979, manufacturing contributed one-quarter of the gross national product, leaving little scope for any immediate expansion.

Energy. Electricity production contributes only about 3 percent of the gross national product. Its principal users are industries, mines, and farms. Electrification of the railways was begun in 1980, and there has also been considerable electrification of low-cost housing in urban townships. Less than half the black homes in Bulawayo and Salisbury, however, had their own electricity at the time of independence.

During the 14 years of UDI, the consumption of power trebled, and the increase was covered by drawing heavily upon the Central African Power Corporation grid (capacity: 11,500,000 kilowatt-hours per year), with its twin stations on the Zimbabwean–Zambian shores of Kariba Gorge. The manufacturing, transportation, and construction industries consumed nearly 50 percent of the power distributed in the grid, and the mining industry about 20 percent. Following independence, plans for the expansion of power resources included a two-stage Wankie thermal plant fuelled by coal, an extension of Kariba South power station, and investigation of downstream sites on the Zambezi at Mpata Gorge and Batoka Gorge.

Financial services. Banking and insurance services grew during the 1970s and by 1979 contributed 5 percent of the gross domestic product. In the later stages of the civil war, the demand for commercial bank loans fell and, after 1974, the total assets of financial institutions began to decline. This decline reflected a fall in demand for hire-purchase and lease-hire facilities and a reluctance by the general public and the business sector to commit themselves to long-term credit. The principal cause of growth in the money supply has been government borrowing for increased expenditure.

Foreign trade. Economic sanctions, which had been imposed by stages from 1966 to 1968 on both Rhodesia's imports and exports, were lifted in December 1979. They had been widely breached, particularly in mineral exports and in the supply of petroleum, but they nevertheless strongly affected certain commodities, such as tobacco exports. Although the country's trade surplus was diminished in 1979 by the rise in oil prices, the value in 1980 of exports (Z$812,200,000) still outpaced that of imports (Z$733,600,000). As a landlocked country, Zimbabwe faces heavy freight and travel costs, as well as outflows on investment account and transfers such as pensions and migrants' remittances; and these payments turn the trade surplus into a small deficit.

Private enterprise and socialism

Management of the economy. ***Private and public sectors.*** The government of independent Zimbabwe moved cautiously to alter the pattern of management that it inherited from the white minority regime. The first budget of July 1980 was described by the finance minister as "conservative [with] a mild and pragmatic application of socialism." But the whites had passed on government machinery that included many levers of economic power. While the whites by inclination were wedded to a system of private enterprise, they had evolved a system of government intervention to support infant industries and maintain agricultural prices through marketing boards. The need to cushion the blows dealt by economic sanctions during UDI brought acceptance of the imposition of exchange and import controls.

Prime Minister Robert Mugabe was swift to use this apparatus in 1980 to guarantee farmers a pre-planting price for maize more than 40 percent higher than the price of 1979–80, while subsidizing the price to consumers of this staple food in preference to continuing the import of 10,000 tons of the grain a month. The country's infrastructure of roads, power generation, transport, and communications was run down during the civil war, and heavy government expenditure in these sectors was planned. To benefit foreign investors, foreign exchange controls were gently applied in the 1980 budget, 50 percent of after-tax profits were allowed to be remitted to non-resident shareholders, and new foreign investment was to be freely repatriated after two years. The government machinery by which the Smith regime controlled the economic affairs of rural blacks was altered after independence by the establishment of district councils that took on the major task of grass roots economic development.

Taxation. The government raises nearly one-third of its revenue from personal and corporate income taxes which, since 1966, have been collected on a pay-as-you-earn system. About one-fifth of government revenue comes from customs and excise duties and sales taxes, and much of the rest from government borrowing and, since independence, international aid.

The independent Zimbabwe government removed sales taxes on the staple items of food and fuel for the poorest people and extended sales taxes to travel, hotel accommodations, taxis, telecommunications, and other services. It continued the former rates of personal income tax—under which the maximum rate of 45 cents on the dollar becomes payable only when the taxable amount exceeds Z$17,000—and corporate income tax—a rate of 45 percent—but it imposed a 10 percent surcharge on both taxes for 1980 and 1981. Married immigrants continued to receive a special tax allowance of Z$800.

Consolidation of trade unions

Trade unions and employers' associations. The evolution of the trade-union movement was some two years behind the pattern of political change by 1980. The new government dealt with immediate labour problems, such as strikes for a higher minimum wage, rather than institute a thorough revision of the basic Industrial Conciliation Act of 1959. The government seemed to favour the strengthening, by mergers or amalgamation, of small unions in the same industry; the strengthening of the whole movement by the formation of a single trade-union congress from the five or six existing confederations of unions; and an arm's-length relationship of government with such a congress. There were in 1980 some 50 recognized unions, with railway workers' and mine workers' unions prominent among them, but the largest sections of the labour force—the agricultural workers and domestic servants—remained outside the system.

Employers' groups, such as the Associated Chambers of Commerce of Zimbabwe and the Association of Rhodesian Industries, remained influential.

Economic policies. The Zimbabwe government of Prime Minister Mugabe and the Zimbabwe African National Union (ZANU), while deriving economic theory from years as a liberation movement based in the professedly Marxist state of Mozambique, was in practice making a range of compromises with the capitalist policies of the previous regime. It faced the fact that about 70 percent of the capital stock in the country is under mainly British and South African control. A further constraint is the shortage of black artisans, technicians, and engineers, which would become even more critical should there be an early exodus of the thousands of white technicians. A first initiative was the establishment of workers' committees in many industries, although with ill-defined powers. A major concern has been the creation of job opportunities for up to 100,000 workers in the first year of independence, above the 990,000 persons in formal employment in 1980. This was to be followed by the building up of industrial and commercial centres based upon the resettlement of black farmers on formerly white-owned lands.

Transportation. The main road system, which is excellent, generally follows the line of white settlement along the spine of the country, with two branches north to Victoria Falls and Kariba and a network fanning out from Fort Victoria, close to the Great Zimbabwe ruins. It consists of about 3,100 miles of paved road. Wartime operations brought an improvement in certain areas, including the construction of strategic roads in the eastern highlands and near the Zambian border. The 14,000 miles of roads in white farming areas and the 30,000 miles of gravel and earth roads in the Tribal Trust Lands received barely adequate maintenance, however.

Railways

The railway closely follows the main road network; its 1,568 miles of single track have a gauge of three feet six inches. Economic sanctions, followed by the closure of the Zambian border in 1973, drastically cut rail traffic and transit revenues from Zambia's copper exports. The southern routes, however, became the Smith regime's lifeline after the British Navy imposed an oil blockade on the Mozambique port of Beira and effectively cut the Beira–Umtali pipeline supply in 1966. Postwar reconstruction was needed on the Maputo–Malvernia line, but more important to this landlocked country was the bottleneck created at the ports of Mozambique and South Africa.

Air Zimbabwe replaced Air Rhodesia, a government-backed company that had operated only within Rhodesia and to and from South Africa. Air Zimbabwe immediately added a weekly flight to Great Britain. The international airport at Salisbury has one of the longest civil runways in the world. Five other airports—at Bulawayo, Kariba, Fort Victoria, and Victoria Falls—can accommodate medium-sized jet aircraft.

ADMINISTRATION AND SOCIAL CONDITIONS

The 1979 constitution

The constitutional framework. The independence constitution of Zimbabwe, which was written in London during September–December 1979, secured majority rule for Zimbabweans. Under the 1979 constitution, white voters, registered on a separate roll, elect 20 of the 100 members of the National Assembly. Although these members no longer can veto constitutional amendments, a unanimous vote is required during the first 10 years to alter the Declaration of Rights, which stipulates (among other matters) that, if land is acquired for settlement schemes, there must be "prompt payment of adequate compensation . . . remittable within a reasonable time to any country outside Zimbabwe." The British insisted that there be a constitutional head of state, a president elected by the National Assembly, and an executive prime minister, and that citizenship of Zimbabwe be automatically available to anyone who was (or had the qualifications to be) a citizen of Rhodesia immediately before independence.

There is a Senate of 40 members, half of whom are either nominees of the white Assembly members or of the Council of Chiefs. The Senate has the power to delay ordinary legislation for 90 days and constitutional amendments for 180 days.

Regional, state, and local governments. There are no regional or state governments in Zimbabwe. At the time of independence, whites controlled the municipal councils, but legislation was soon introduced to amalgamate each municipal council with the council of its surrounding township and, for the first time, black mayors were elected in 1981. Local government elections in rural areas replaced the old apparatus of district commissioners with a party-based council structure.

The political process. In the elections held in February 1980 under the 1979 constitution, Ian Smith's Rhodesian Front took all 20 seats on the white voters' roll. Of the nine parties that contested the 80 common roll seats, Bishop Abel Muzorewa's United African National Council (UANC) secured only three seats; Joshua Nkomo's Zimbabwe African People's Union (ZAPU) won 20 seats, representing nearly all in Matabeleland, and Mugabe's Zimbabwe African National Union (ZANU) won 57 seats, nearly all in Mashonaland. A total of 2,649,529 valid votes were cast, and the Commonwealth Observer Group drawn from 11 nations declared their satisfaction that the elections had been "free and fair."

Justice. The Ministry of Law and Order of Smith's period was renamed the Ministry of Justice by the new independent government. The government voted to extend the state of emergency, first delcared at UDI, because of unsettled conditions, mainly in rural areas.

Under the constitution of 1979, a four-member Judicial Service Commission advises the president on the appointment of judges to the High Court. High Court judges may not be removed from office except for misconduct or incapacity. The pre-independence chief justice voluntarily resigned, however, accepting that his acquiescence in the Smith regime's claim to constitutionality made his continuance in office difficult. He was succeeded by another white judge, and the first black lawyer was appointed a High Court judge later in 1980. In addition to magistrates who preside over criminal and civil litigation, other courts adjudicate on matters of African law and custom.

Integration of the armed forces

Armed forces. The integration of three rival and politicized armies—the Rhodesian Army and the armies loyal to ZAPU and ZANU—was one of the most urgent tasks facing the new government. A team of British military advisers worked in late 1980 to form nine infantry battalions, each 1,000 strong, of integrated troops after about six months of retraining. Many of the 38,000 nationalist troops who had been gathered at rural assembly points during the pre-election cease-fire period became restless with their continued confinement. How many of them eventually would be retrained for the national army or given other training was not immediately clear; it seemed likely that several thousand more than the size or economy of Zimbabwe warranted would remain in uniform, probably as reservists. The Rhodesian Air Force maintained antiquated squadrons of Hunter fighters and Canberra light bombers, having spent reequipment funds on the purchase of counter-insurgency helicopters. The paramilitary British South Africa Police had been responsible for internal security from 1890; it was renamed the Zimbabwe Republic Police and underwent retraining.

Educational services. The dismantling of Rhodesia's segregated system of schooling began less than two years before independence, and the system's effects will be long felt in Zimbabwe. The minority government had concentrated upon providing compulsory (and virtually free) education to white children between the ages of five and 15, and had left the schooling of black children in the hands of missionaries. In 1950 there were only 12 government schools for blacks, compared with 2,230 mission and independent schools. The disparities continued through the 1970s, when there were about 800,000 blacks enrolled in primary schools and less than 50,000 blacks in secondary schools. Facilities were sharply reduced after the lower primary stage, and the civil war cut opportunities for blacks even further; from 1978, at least one-quarter of all black primary schools were closed. Enrollment fell by 300 at the University of Zimbabwe, where blacks had become a majority of the 1,480-strong student body by 1979. In the 1980–81 budget, priority was given to reopening and reequipping schools and to providing new schools in the drive toward free primary education. Because of the many opportunities in higher education offered Rhodesians by Commonwealth and other countries during UDI, Zimbabwe moved to independence with one of the best educated leaderships any African country had known.

Health services and housing. Health services were biased toward curative medicine in central hospitals, two of which—Harari Hospital in Salisbury and Mpilo Hospital in Bulawayo—are known for their service to black patients. Before 1980, missionaries had the major responsibility for running rural clinics and small hospitals. In the budget of 1980–81, health allocations were increased by 55 percent. A major effort was planned to reequip rural clinics and provide free health services to the poor, concentrating on preventive care.

A severe housing shortage in the main urban centres was aggravated by the influx of rural refugees and by the trebling of the cost of building materials during UDI. The 1980–81 budget provided for the construction of 27,000 homes, and the government appealed to private contractors and individual employers to aid in the deveiopmenr of housing for lower paid workers.

Social conditions. ***Wages and the cost of living.*** Statistics are no longer broken down by racial groups in Zimbabwe, but as late as 1977 the average wage of R$590 a year for the 900,000 blacks in cash employment was less than one-tenth of the average wage paid to non-blacks. The wage average for black workers was drawn down by the low wages paid to those in agriculture and domestic service, who numbered nearly 450,000 and earned a minimum Z$30.80 a month at independence. Mine workers received a minimum Z$44.66 a month, with food and housing allowances and a pension plan. For the higher income group, the rise in the cost of living after 1964 was greatest in transport, foodstuffs, drink, and servants' wages; for the lower income group, the greatest increase was in household stores. For both groups the consumer price index rose by less than 20 percent in the first seven years of UDI and then jumped 85 to 100 percent in the second seven-year period, while average earnings doubled.

Infant mortality

Health conditions. As in other Third World countries, the burden of disease is heaviest on Zimbabwe's youngest children. The infant mortality rate for the black population in malarial parts of the Zambezi Valley has been as high as 300 per 1,000, and the rate is thought to lie between 120 and 220 per 1,000 for the black population as a whole. Measles and pneumonia are major causes of death; a tuberculosis control scheme of the mid-1970s was relatively effective. Improved nutrition is increasingly seen as the most important health need.

CULTURAL LIFE AND INSTITUTIONS

The year-round temperate climate of the Highveld has combined with the natural inclinations of the white population to produce an outdoor society. Tennis—whether on farms or at urban clubs—and bowling have many more white-clothed followers than any ballet group, while a standing joke against local artists has been that so many have drawn their inspiration solely from the mauve-flowering jacaranda and from Zimbabwe's famous balancing rocks. The civil war stimulated among whites patriotic songs and poetry of doubtful artistic quality. Happily for the cause of reconciliation, the first sport heroes after independence were the members of the all-white team that was awarded the first gold medal for women's field hockey in Olympic history at Moscow in 1980. The most famous of Rhodesian-bred writers, Doris Lessing, settled in England in 1949.

Cultural renaissance

In some contrast, the nationalist struggle prompted a renaissance of Shona culture. A forerunner of this renaissance (and a victim of the liberation struggle) was Herbert Chitepo, both as abstract painter and epic poet. Stanlake Samkange's novels reconstruct the Shona and Ndebele world of the 1890s, while those of the much younger Charles Mungoshi explore the clash of Shona and Western cultures in both the Shona and English languages.

Folk traditions have survived in dance and pottery. The revival of sculpture has drawn on tribal religion and totems to produce some remarkable works, particularly those of Takawira and the Tengenenge school of craftsmen who sculpt in hard serpentine.

Newspapers and radio, although lively at times, have not contributed much that is significant to the country's cultural life. The major newspapers, the ***Herald*** (formerly, the ***Rhodesia Herald)*** of Salisbury and the ***Chronicle*** of Bulawayo, were owned by the Johannesburg-based Argus Corporation; in 1981, the government took a controlling interest. Newspapers that appeared in the Shona and Ndebele languages were short-lived, and the most successful multilingual paper, ***Moto,*** published by the Roman Catholic Mambo Press, was banned from 1974 until 1980. Wartime censorship had a deadening effect, most noticeably upon radio and television broadcasting.

PROSPECTS FOR THE FUTURE

The potential for Zimbabwe is remarkable. Its wide range of minerals and considerable energy resources provide a solid base for economic growth and social development. Cultivable areas are large and fertile enough not only to feed the country's own rapidly increasing population but to export grain crops to neighbouring states as well. Zimbabwe has among its white population a rich store of managerial and administrative experience, peppered by pioneer inventiveness, and probably the best educated leadership of any African country at independence. If these elements can be combined productively, if a transfer of land ownership can be achieved speedily and without rancour, and if the well-noted dangers of a white exodus and tribal fighting are largely avoided, Zimbabwe can become not only a prosperous state itself but also the focal point for development in southern Africa.

BIBLIOGRAPHY. PHILIP MASON, ***The Birth of a Dilemma: The Conquest and Settlement of Rhodesia*** (1958), the best account of the early days (up to 1918) of white settlement and race relations; T.O. RANGER, ***Revolt in Southern Rhodesia, 1896–97*** (1967), a full-length study, drawing from African sources, of the risings against white rule in 1896–97, with significance in terms of the modern liberation movement. LAWRENCE VAMBE, ***An Ill-Fated People: Zimbabwe Before and After Rhodes*** (1972), a portrayal with family history and humour of the sadness of occupation; COLIN LEYS, ***European Politics in Southern Rhodesia*** (1959), an analysis of the flow of immigrants and of the continuously rightward trend in party politics before federation; CLYDE SANGER, ***Central African Emergency*** (1960), describes the first years of federation and the growth of the nationalist movement in territories that have since become Zimbabwe, Malaŵi, and Zambia; NDABANINGI SITHOLE, ***African Nationalism*** (1959) and ***Letters from Salisbury Prison*** (1976), a balanced political analysis and an exhortation of a father's concern for six children scattered by the civil war; NATHAN SHAMUYARIRA, ***Crisis in Rhodesia*** (1965), a broad description of the racial disparities and political collisions that culminated in the unilateral declaration of independence; ROBERT BLAKE, ***A History of Rhodesia*** (1977), includes a commentary sympathetic to the white Rhodesian leaders; MARTIN MEREDITH, ***The Past Is Another Country*** (1979), a detailed and objective account of the political moves inside Rhodesia from 1965 to 1979; MARTIN BAILEY, ***Oilgate: The Sanctions Scandal*** (1979); ROGER RIDDELL, ***et. al., From Rhodesia to Zimbabwe,*** 8 vol. (1977–79), a thorough investigation of the major problem areas for change; D. MARTIN and P. KOHNSON, ***The Struggle for Zimbabwe*** (1981), an authoritative account of the liberation movement. DIANA MITCHELL and ROBERT CARY, ***African Nationalist Leaders in Rhodesia*** (1977), and DIANA MITCHELL, ***Who's Who 1980,*** providing invaluable background on the careers of the new leaders.

Of official reports published in Salisbury, the following retain more than historical interest. ***Second Report of the Select Committee on the Resettlement of Natives*** (1960), covering the heart of the dispute over landholdings and important recommendations on land apportionment reform: ***Final Report of the 1962 Census of Africans in Southern Rhodesia*** (1964), a basic source of information on socioeconomic conditions; ***Census of Population, 1969*** (1971); ***Report of the (Phillips) Advisory Committee on the Development of the Economic Resources of Southern Rhodesia, with Particular Reference to the Role of African Agriculture*** (1962), containing details of the country's resources and potentialities. Three reports published in London cover more recent historical landmarks. ***Report of the Commission on Rhodesian Opinion*** (1972), known as the Pearce Report, a critically important state document; ***Report on the Supply of Petroleum and Petroleum Products to Rhodesia*** (1978), known as the Bingham Report, a painstaking but incomplete investigation of circumventions of sanctions by oil companies and connivance by British officials: ***Southern Rhodesian Elections*** (1980), the report of the Commonwealth Observer Group, an illuminating picture of how the main actors behaved under strain in the final scenes before independence.

(C.W.S.)

Rhône River

The RhBne is a historic river of Switzerland and France and the only significant river of Europe flowing directly to the Mediterranean Sea. It is thoroughly Alpine in character, and in this respect it differs markedly from its northern neighbour, the Rhine, which leaves all of its Alpine characteristics behind when it leaves Switzerland. The scenic and often wild course of the RhBne, the characteristics of the water flowing in it, and the way it has been used by mankind have all been shaped by the influences of the mountains, right down to the river mouth, where sediments marking the Rhône's birth in an Alpine glacier are carried into the warmer waters of the Mediterranean.

The course of the river can be divided into three sectors lying, respectively, in the Alps; between the Alps and the Jura Mountains and through the latter; and finally in the topographical furrow of Alpine origin running from the city of Lyon to the sea.

The Alpine sector

The course. The RhBne originates in the Swiss Alps, upstream from Lake Geneva (Lake Léman). It comes into being at an altitude of 6,004 feet (1,830 metres), emerging from the glacier of the same name, which descends the south flank of the Dammastock, an 11,909-foot (3,630-metre) peak. The river then traverses the Gletsch Basin (5,778 feet; 1,761 metres), from which it escapes through a gorge, and flows along the floor of the Val de Conches (Goms) at an altitude between 4,600 and 4,000 feet. It next enters another gorge before reaching the plain of the Valais, which extends between the towns of Brig and Martigny, at an altitude between 2,300 and 1,600 feet. In crossing this high and rugged mountain area, the river makes successive use of two structural troughs. The first runs between the ancient crystalline rock massifs of the Aare and of the Gothard; further downstream, the second runs between the arched rock mass of the Bernese Oberland and, on the south, the massive rock face of the Pennine Alps. From Brig onward, the landscape changes. During the last ice age, a large glacier, fed by several small ones, plowed down the valley floor of the Valais, and, except for some harder rock obstacles found near the town of Sion, succeeded in widening and deepening the narrow valley floor. As it did so, it held back both the upper RhBne and those of its tributaries that come down from the Pennine Alps. When the ice sheets retreated, both the tributaries—the Viege (Vispa), Navigenze, Borgne, and Drance—and the RhBne cut new, deep gorges to connect their lower courses to the new valley floor. These have created considerable difficulty in terms of modern transportation, necessitating a whole series of hairpin-bend road links.

After Martigny, where the valley floor is wider and liable to flooding, the youthful RhBne thrusts northward at a right angle, cutting across the Alps through a transverse valley. At first, near the town of Saint-Maurice, this is no more than a spectacular gorge, but it soon becomes wider and flatter. Here, too, the river route has been assisted by structural factors, specifically by a dip in the crystalline rock massifs running from Mont Blanc to the Aare and by the discontinuity between the limestone masses of the Dents du Midi and of the Dent de Morcles. Across the mountain barrier, the muddy waters of the RhBne enter another wide plain and then plunge into the clearer, stiller waters of Lake Geneva, forming an enlarging delta.

The second sector of the Rhône's course commences with Lake Geneva, large (224 square miles in area) and deep (1,000 feet) and lying between Switzerland and France in a basin hollowed out of the less resistant terrain by the former RhBne glacier. It is possible that the

RhBne, before the last glaciation, headed originally toward the present-day lakes of Neuchâtel or Morat, and, as the Aar does now, reached the valley of the Rhine. Upon leaving Lake Geneva, which has turned the course of the river to the southwest and decanted the sediment from its waters, the RhBne very quickly regains in full the milky colour so characteristic of Alpine rivers. Just below the city of Geneva, it receives its powerful tributary the Arve, which rushes down from the glaciers of Mont Blanc.

The middle course

From its juncture with the Arve to the city of Lyon, the Rhône has to cross a difficult obstacle, the undulating series of mountains forming the Jura Range. It does this partly by cutting through narrow cross valleys and partly by using existing structural downfolds to skirt the obstacles. It follows a complicated zigzag course. In the valleys, it surges through narrow clefts known as *cluses.* At the town of Bellegarde, the river is joined from the north by the Valserire, and, swinging south, plunges into a deep gorge now submerged in the 14-mile-long Génissiat Reservoir. In the wider sections of its course in this region, the RhBne runs through glacier-excavated basins that its own deposits have barely filled, causing intermittent swamps. It is also joined by the Ain, from the north, and, on the left bank, by the Usses, Fier, and Guiers. The river next widens, and the terrain becomes less hilly and, at Le Parc (some 95 miles above Lyon) becomes officially "navigable," although the average depth is no more than three feet.

The lower sector

The RhBne then enters its third sector, marked by the great north–south alpine furrow that is also drained by its principal tributary, the Saône. The latter lies in the basins that the Ice Age glaciers hollowed out between the Jura Range to the east and, further west, the eastern edge of the Paris Basin and the uplands of the Massif Central. It forms an important commercial link to the industrialized regions of northern France. From the city of Lyon onward, the river occupies the trough lying between the Massif Central and the Alps, a channel up which the sea of the Pliocene Epoch, 10,000,000 years ago, ascended for a great distance. Over the last 70,000,000 years, powerful deposits brought down from the Alps have hammered out a very deep channel against the Massif Central. This is, in fact, so deep that at Vienne and again at Tain the RhBne cuts right through hard crystalline outcrops. The valley consequently takes the form of a series of gorges and basins, the latter often having a series of terraces corresponding to variations in the levels of ice and of river. Although the tributaries—notably the Ardtche—rushing down into the Rhône from the Massif Central are formidable when in flood, the great Alpine rivers, the Istre, and the Durance, joining the left bank, are most important in their effect on riverbed deposits and on the volume of water. Below Mondragon the RhBne Valley becomes wider and more marshy, liable to flood if unimproved, and the river course itself becomes more braided, divided into a network of interlaced streams. The river delta extends from near Arles, about 25 miles from the sea, where the twin channels of the Grand and Petit RhBne separate the intervening wild, marshy landscape of the Camargue from the surrounding coarse pastureland known as La Crau.

Flood volumes

Hydrology. The flow regime of the RhBne owes its remarkable mean volume to the influence of the Alps. At Beaucaire the flow amounts to 64,300 cubic feet per second; at Lyon 22,600 cubic feet per second (the Saône alone contributes 14,100 cubic feet per second); the Isère adds another 12,400 cubic feet per second. The melting of the Alpine snows gives the highest mean flows in May, while the Saône attains its maximum in January. The flood volumes of spring and autumn are formidable, reaching 460,000 cubic feet per second for the RhBne at Beaucaire, 152,000 for the Saône, 102,000 for the Isbre, and even up to 124,000 for the little Ardèche, as a result of the intensity of the Mediterranean rains flung against the eastern skirt of the Massif Central. The RhBne has an abundant flow but maintains a strong gradient almost to its mouth. At Lyon, for example, its altitude is 560 feet at 205 miles from the sea; the Rhine, by way of comparison, has an altitude of 456 feet at Strasbourg, more than 360 miles from the sea. As the size of the delta region testifies, the river transports large amounts of alluvial deposits and is also powerful enough to cut through a variety of rock masses. As a result, the RhBne of today is well adapted to the production of electricity but has always been difficult to navigate.

Human geography. The utilization of the Rhône region by man has required a long historical struggle, which only entered a decisive phase in the third quarter of the 20th century. The main aspects of this struggle have been in the areas of fishing, the mastery and working of the soil, navigation, and hydroelectric power production.

Fishing is practiced in all the water courses of the basin but attains a modest local commercial importance only on the lakes fed by the RhBne, Lake Geneva and the Lac du Bourget.

Agriculture in the RhBne Valley has largely covered the low areas, plains, and islands, which the river and its tributaries threaten with periodic flooding. In the Valais Valley, in the century following 1850, the Rhône was diked and narrowed, having raised its bed and made the draining of the surrounding plain necessary. These improvements, together with the protective agricultural policies of the Swiss, have made a fine development of farms, orchards, and vineyards possible. Comparable works have been carried out in France, notably on the Isbre, at Combe de Savoie and Grésivaudan, and on the Basse Durance, where the Comtat Plain is entirely given to vegetables. The Camargue region of the delta has become a vast rice field, while the Serre-Ponçon Dam on the Durance assures reliable irrigation.

Navigation

A bad outlet to the sea, the strong gradient, and gravel shoals form obstacles to navigation. Navigation has, however, always been carried on, particularly between Lyon and the sea. At first, horses and then tugs pulled the boats upstream. In spite of the navigational improvements made to the river, steam- and, later, gas-engine navigation, subject as it is to competition from the railways, has known only small-scale success. By the 1970s, attempts were made to stimulate navigation by a complete modernization of the engineering facilities. This was to be made possible, at least on the French part of the RhBne, by hydroelectric power projects. These would enable a comprehensive solution: the elimination of shoals, which would be submerged under reservoirs or avoided through bypass canals; the replacement of the original gradient by a succession of level reaches and locks; and a series of large generating stations resulting in an immediate profitability for the project. Between 1949 and 1970 the Compagnie Nationale du RhBne had already completed an extensive series of related projects: at Génissiat (electricity only), Donzère-Mondragon, Montélimar, Baix-le-Logis Neuf, Beauchastel and Pierre-Bénite. Together these projects harnessed more than half of the entire potential hydroelectric power of the RhBne, and the work is continuing. The production of energy alone has sufficed to justify these projects. It is hoped that navigation, too, will take on new life, but this cannot be positively predicted because of the uncertain competitive role of roads and pipelines.

Even with the faults of parts of its waterway, the RhBne Basin nevertheless constitutes one of the great economic regions of Switzerland and of France, draining rich plains, as well as an important part of the Alps. Great cities attest the antiquity and the strength of man's interest in it. In addition to Lausanne and Geneva on Lake Geneva, Lyon stands at one of the great European crossways. Grenoble is an important city at the confluence of the Isère and the Drac. Avignon, Vienne, and Arles were magnificent Gallo-Roman cities, and the river was one of the spearheads for the penetration of Mediterranean cultures and peoples into northern Europe. While the rivers of the basin are not everywhere navigable, their valleys have successively accommodated paths, roads, railways, and modern highways; whatever the means of travel, the RhBne Basin has always been a noble thoroughfare.

BIBLIOGRAPHY. DANIEL FAUCHER, *L'Homme et le Rhône* (1968), is a recent, comprehensive work on the Rhône, which also contains a copious bibliography. To supplement Faucher, see the exhaustive works of MAURICE PARDE, a hydrologist of world reputation: *Le Régime du Rhône*: *Étude hydrologique*, 2 vol. (1925), and *Quelques nouveautés sur le régime du Rhône* (1942), adds some corrections and new numerical data to the *1925* work. For the RhBne Valley complex in the Valais, see PAUL and GERMAINE VEYRET, *Au coeur de l'Europe: les Alpes* (1967), useful also for all alpinistic problems of the RhBne Basin.

(P.V.)

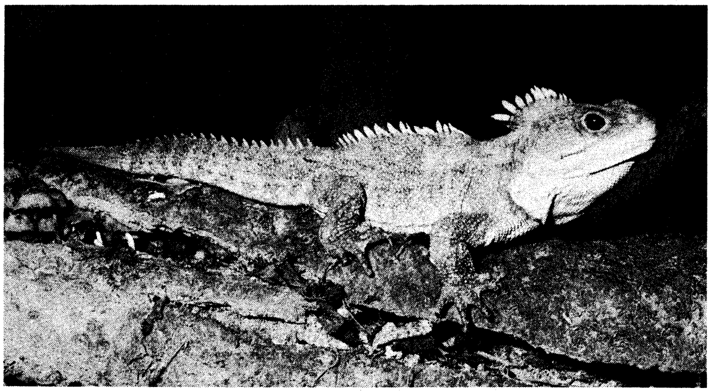
Rhynchocephalia

The Rhynchocephalia constitute one of the four orders of living reptiles; the only surviving representative of the group is the tuatara, or sphenodon (*Sphenodon punctatus*). Structurally, the tuatara is not much different from related forms, also assigned to the order Rhynchocephalia, that may have appeared as early as the Lower Triassic Period (over 200,000,000 years ago).

Distribution

Until recently the tuatara lived on the two main islands of New Zealand. Today, it is found only on certain islets in Cook Strait between the main islands and on islets between East Cape and North Cape of the North Island of New Zealand.

W.H. Dawbin



Male tuatara with white spines erected, a behaviour characteristic of the tuatara when excited or hunting.

Natural history. Male tuataras may attain a length of 60 centimetres (about 24 inches) and a weight of 1,000 grams (about 2.2 pounds). Females grow to about 50 centimetres (20 inches) and sometimes weigh as much as 500 grams (1.1 pounds). Eight to 15 eggs are laid in a nesting burrow covered by several centimetres of soil. The young emerge about 13 months later, in early summer to midsummer. Adult size is not reached for 50 to 60 years, and the animal may live to the age of 100. Sexual maturity is believed to occur after about 20 years.

During mating, the more prominent crest of the male becomes turgid and erect as he stalks the female, approaching her in a slow, jerky fashion. He then grips her over the shoulders with his forelimbs. Unlike the lizards, the tuatara does not use the jaw for grasping the female. The length of time between mating and egg laying is not known.

The mode of locomotion is primitive. The animal moves in a sprawling fashion, the belly leaving the ground only momentarily. The pattern of limb movement is unusual in that the forelimb contacts and leaves the ground before the hindlimb on the opposite side. As it moves, the body is thrown into marked lateral, or sideward, bends.

Tuataras are essentially solitary, nocturnal, burrowing animals, seldom travelling more than a few metres from their burrows during the day. Feeding takes place mainly at night. The principal diet of the young is insects; adults eat, in addition to insects, snails, lizards, young seabird chicks, and eggs. They will drink water if it is available but can survive for months on water obtained from dew and solid food.

The islands inhabited by the tuatara have large numbers of ground insects, which are an important source of food. Bird burrows often serve as ready-made homes for the tuatara, but the animal is capable, from the time it hatches, of making its own burrow if none is already available. The burrow may be shared by a bird and a tuatara, as well as by the bird's chick or egg. (Occasionally the tuatara preys on the chick or adult sharing the burrow.)

Natural enemies

The largely nocturnal habits of the animal conceal it from many potential predators, but occasionally it is preyed upon by hawks, gulls, or kingfishers. Rats, introduced by man, are serious enemies of eggs and the young. The animal may be regarded as seriously threatened with extinction on islands where rats occur.

Form and function. The tuatara has two pairs of well-developed limbs, a strong tail, and a scaly crest down the neck and back. The scales, which cover the entire animal, vary in size. The tuatara also has a bony arch, low on the skull behind the eye, that is not found in lizards. This arch is formed by the presence of two large openings (temporal fossae) in the region of the temple. It has been used as evidence that the tuatara is a survivor of the otherwise extinct order Rhynchocephalia and is not a lizard. The teeth of the tuatara are acrodont—*i.e.*, attached to the rim of the jaw rather than inserted in sockets. The tuatara is also unique among reptiles in not possessing a male copulatory organ. As in birds, sperm transfer is effected during contact between the male and female cloacae.

The heartbeat and respiratory rate of the tuatara are relatively slow; oxygen consumption, accordingly, is low. Tuataras actively forage in temperatures as low as 6° C (43" F) and can briefly tolerate temperatures above 37° C (99" F). Tests indicate that the preferred temperature is about 22° C (72" F)—rather low for a reptile. The excretion of nitrogenous wastes, a by-product of the animal's metabolism, are in the form of uric-acid masses, or concretions. About 10 to 30 percent of this waste may occur as the substance urea in the urine; the urea concentration rises as the amount of protein increases in the food consumed.

Evolution and classification. The tuatara has changed little in skeletal features since the Jurassic Period (136,000,000–190,000,000 years ago), when the closely related *Homoeosaurus* occurred in Europe. There is evidence that the line of rhynchocevhalians began as early as the Lower Triassic. Most forks of that time were of moderate size and, except for one specialized family (Rhynchosauridae), were never abundant. Fossils are unknown in any region during Tertiary times (2,500,000–65,000,000 years ago). Remains from the Pleistocene Epoch (10,000–2,500,000 years ago) in New Zealand are structurally identical with the living tuatara.

The order Rhynchocephalia belongs to the subclass Lepidosauria of the class Reptilia. Three families (Sphenodontidae, Rhynchosauridae, and Sapheosauridae) are assigned to the order; two families (Claraziidae and Pleurosauridae) are tentatively assigned.

In addition to the living species, *Sphenodon punctatus,* the family Sphenodontidae is represented by European fossils of three genera from the Upper Triassic (200,000,000 years ago); by Asian, European, and North American fossils of genera from the Upper Jurassic Period; and by a southern African fossil of a genus from the Lower Triassic.

The family Rhynchosauridae is represented by fossils of eight genera, some tentatively assigned. All are from the Triassic Period and occurred in Europe, South America, south Asia, southern Africa, and East Africa.

The family Sapheosauridae is represented by one genus from the Upper Jurassic Period in Europe.

BIBLIOGRAPHY. W.H. DAWBIN, "The Tuatara in Its Natural Habitat," *Endeavour, 21:16–24* (1962), a review of distribution, life cycle, and ecology; A. GUNTHER, "Contribution to the Anatomy of *Hatteria (Rhynchocephalus,* Owen.)," *Phil. Trans. R. Soc.*, 157:595–629 (1867), the original detailed description of skeleton, muscles, externals, and classification; A.S. ROMER, *Osteology of the Reptiles* (1956), sections on musculature, the brain, and the circulatory system, as well as osteology and relationships; RICHARD SHARELL, *The Tuatara, Lizards and Frogs of New Zealand* (1966), a recent popular account of the natural history of the tuatara.

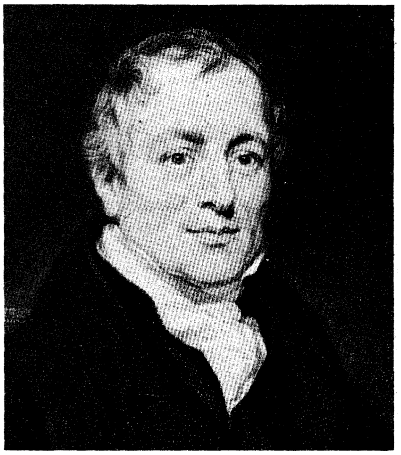
(W.H.D.)

Ricardo, David

Among the Englishmen who at the beginning of the 19th century systematized and gave classical form to the rising science of economics, the most famous and influential was David Ricardo. His theoretical work still commands the critical attention of economists.

Early life. Ricardo was born in London, Aprii 18 or 19, 1772, the third son of a Dutch Jew who had made a fortune on the London Stock Exchange. At the age of 14 he entered his father's business, for which he showed great aptitude. By the time he was 21, however, he had broken with his father over religion, become a Unitarian, and married a Quaker. This required him to set up on his own. He continued as a member of the stock exchange, where his talents and character won him the support of an eminent banking house. He did so well that in a few years he acquired a fortune and was then in a position to indulge his wide-ranging tastes in literature and science, particularly in the fields of mathematics, chemistry, and geology.

BY courtesy of the National Portrait Gallery, London



Ricardo, portrait by Thomas Phillips, 1821.
In the National Portrait Gallery, London.

Beginning of his interest in economics

His interest in economic questions arose in 1799 when he happened to read Adam Smith's *Wealth of Nations.* For ten years he studied economics, somewhat offhandedly at first and then with greater concentration. His first published work was *The High Price of Bullion, a Proof of the Depreciation of Bank Notes* (1810), an outgrowth of letters Ricardo had published in the *Morning Chronicle* the year before. His book gave a fresh stimulus to the controversy then raging over the policy of the Bank of England. The strains of the wars with France had caused the government to forbid the Bank of England to pay its notes in gold. Freed from the necessity of cash payment, both the Bank of England and the country banks had increased the amount of their note issues and the volume of their lending. The directors of the Bank of England maintained that the subsequent increase in prices and the depreciation of the pound had no relation to the increase in bank credit. Ricardo and other opponents of that view asserted that there indeed was a close connection between the volume of bank notes and the level of prices and that the level of prices in turn affected foreign exchange rates and the inflow or outflow of gold. It followed that the bank, as custodian of the central gold reserve of the country, must shape its lending policy according to general economic conditions and exercise control over the volume of money and credit. The controversy was thus of the highest importance in the development of central banking theory. A committee appointed by the House of Commons, known as the Bullion Committee, confirmed Ricardo's views and recommended the repeal of the Bank Restriction Act.

At this time Ricardo began to acquire some friends who were to have considerable influence on his further intellectual development. One of these was the philosopher and economist James Mill (father of John Stuart Mill), who became his political and editorial counsellor. Another friend was the utilitarian philosopher Jeremy Bentham. Still another was Thomas Malthus, best known for his theory that population tends to increase faster than the food supply—an idea that Ricardo accepted.

In 1815 another controversy arose over the Corn Laws, which regulated the import and export of grain. A decline in wheat prices had led Parliament to raise the tariff on imported wheat. This provoked a popular outcry and caused Ricardo to publish his *Essay on the Influence of a Low Price of Corn on the Profits of Stock* (1815), in which he argued that raising the tariff on grain imports tended to increase the rents of the country gentlemen while decreasing the profits of manufacturers.

Retirement from business. The year before his Corn Law essay, at the age of 42, he had retired from business and taken up residence in Gloucestershire, where he had extensive landholdings. In 1819 he purchased a seat in the House of Commons, as was done in those times, and entered Parliament as a member for Portarlington. He was not a frequent speaker, but so great was his reputation in economic affairs that his free-trade opinions were received with respect, although they did not yet represent the dominant thiiiking in the House.

Content of *Principles of Political Economy and Taxation*

In his *Principles of Political Economy and Taxation* (1817), Ricardo undertook to analyze the laws determining the distribution of the social product among the "three classes of the community," namely, the landlords, the workers, and the owners of capital. He applied his findings more widely, however, and elaborated various other economic principles. He found the relative domestic values of commodities to be dominated by the quantities of labour required in their production, rent being eliminated from the costs of production.

He concluded that profits vary inversely with wages, which move with the cost of necessaries, and that rent tends to increase as population grows, rising as the marginal costs of cultivation rise. He supposed that there was little tendency to unemployment; but he remained apprehensive lest population grow too rapidly, depress wages to the subsistence level, and, by extending the margin of cultivation, reduce profits and check capital formation. He also concluded that trade between countries was not dominated by relative costs of production and by differences in internal price structures that reflected the comparative advantages of the trading countries and made exchange desirable. He treated monetary questions and esyecially taxation at length. Although he built in part upon the work of Adam Smith, he defined the scope of economics more narrowly than had Smith and included little explicit social philosophy.

Illness forced Ricardo to retire from Parliament in 1823. He died that year, on September 11, at the age of 51.

Assessment. Despite his relatively short career and the fact that most of it was preoccupied with business affairs, Ricardo achieved a leading position among the economists of his time. His views won considerable support in England despite the abstract style in which he set them forth and in the face of heavy counterfire from his opponents. Although his ideas have long since been superseded or modified by other work and by new theoretical approaches, Ricardo retains his eminence as the thinker who first systematized economics. Writers of various persuasions drew heavily upon his ideas, including those who favoured laissez-faire capitalism and those, such as Karl Marx and Robert Owen, who opposed it.

BIBLIOGRAPHY. Ricardo's *Works and Correspondence,* 10 vol., were edited by PIERO SRAFFA and M.H. DOBB (1951–55). His *Principles of Political Economy and Taxation* have appeared in several editions, including Everyman's Library. A good introduction to Ricardo is J.H. HOLLANDER, *David Ricardo: A Centenary Estimate* (1910, reprinted 1960). A guide to Ricardo's opinions on economic matters is OSWALD ST. CLAIR, *A Key to Ricardo* (1957). For the serious student, two books by MARK BLAUG are to be recommended: *Ricardian Economics: A Historical Study* (1958), covering the rise and fall of Ricardian economics in the first half of the 19th century; and *Economic Theory in Retrospect,* rev. ed. (1968), which contains a treatment of Ricardo's system in modern analytical terms, as well as further bibliographical references.

Ricci, Matteo

Matteo Ricci, an Italian Jesuit missionary, introduced Christian teaching to the Chinese Empire in the 16th century. He became known in China as a man of refinement; thus he was able to direct the attention of the Chinese to Western culture, which they had largely ignored. When he arrived in China, he was called a "foreign devil"; later, his Chinese friends addressed him as the "Wise Man from the Great West." This contrast indicates the change both in his own social status and in the feeling toward him of Chinese scholars.

Early life and education

Ricci was born October 6, 1552, of a noble family in Macerata, in central Italy. His father, Giovanni Battista Ricci, a pharmacist by profession, dedicated most of his time to public affairs and for a time served as governor of the city. His mother, Giovanna Angiolelli, was a woman known for her simple piety. Matteo, their oldest child, after preliminary studies at home, entered the school that the Jesuit priests opened in 1561 in Macerata. After completing his classical studies, he set out at the age of 16 for Rome to study law. There he was attracted to the life of the Jesuits, and on August 15, 1571, he requested permission to join the order.

Approved by the Pope in 1540, the Society of Jesus (Jesuits) was already well known for its spirit of apostolic initiative. Its members were distinguishing themselves in scientific research as well as in their voyages to the new worlds. Stimulated by the examples of his seniors, Ricci dedicated himself to efforts in both fields. Shortly after beginning his study of science under the noted mathematician Christopher Clavius, he volunteered for work overseas in the Far East. In May 1577, he set off for Portugal, where he studied for a short time at the famous University of Coimbra while waiting for a ship. In the following year, on March 24, he embarked at Lisbon and arrived on September 13 at Goa, the Portuguese outpost on the central west coast of India. Ricci carried on his studies for the priesthood there but was ordained in 1580 at Cochin, on the Malabar Coast, where he had been sent for reasons of health. Returning to Goa, he was ordered, in April 1582, to proceed to China.

With its huge population, China was an area that Christian missionaries, especially the Jesuits, greatly wished to enter. St. Francis Xavier, one of the first companions of St. Ignatius of Loyola, died in 1552 on the tiny island of Shangchuan in sight of the tightly closed mainland. When Ricci arrived, China was still closed to outsiders; but the missionary strategy of the Jesuits had undergone modification. Great stress was put on the importance of learning the Chinese language and of acquiring knowledge of the culture. Previously, missionaries had attempted to impose Western customs and the use of the Latin language in religious rites. The new approach of adaptation to national customs was established by Alessandro Valignano, who had received Ricci into the Jesuits and was at this time visitor of the Jesuit missions in the Far East. (A visitor is the official responsible for making sure the religious and temporal affairs of all the houses of an institute in an area are properly followed.) First Michele Ruggieri and then Ricci were called to the Portuguese province of Macau to prepare to evangelize China; Ruggieri, however, returned to Italy in November 1588, leaving to his younger compatriot the burden and the honour of founding the church in China.

Mission to China

Ricci arrived at Macau, a small peninsula on the east coast of China, in August 1582, and began at once his study of Chinese. The following year he and Ruggieri were given permission to settle in Chao-ch'ing, then the capital of Kwangtung Province. In his *History of the Introduction of Christianity in China,* Ricci described their work as follows:

> So as not to occasion any suspicion about their work, the fathers [*i.e.*, the Jesuits] initially did not attempt to speak very clearly about our holy law. In the time that remained to them after visits, they rather tried to learn the language, literature, and etiquette of the Chinese, and to win their hearts and, by the example of their good lives, to move them in a way that they could not otherwise do because of insufficiency in speech and for lack of time.

Despite that caution, Ruggieri published the first Catholic catechism in Chinese, and Ricci produced the first edition of his remarkable map of the world, the "Great Map of Ten Thousand Countries," which showed the Chinese intelligentsia China's geographical relation to the rest of the world.

In 1589 Ricci moved from Chao-ch'ing to Shao-chou (Shiuhing), where he became a close friend of the Confucian scholar Ch'ii T'ai-su. Ricci taught him the rudiments of mathematics, receiving in return an introduction into the circles of the mandarins (high civil or military officials of the Chinese Empire) and of the Confucian scholars. Ch'ii, noting that Ricci wore the habit of a Buddhist monk (which he had adopted upon entering China), suggested that it would be better to dress as a Chinese scholar, a suggestion that Ricci followed immediately after leaving Kwangtung.

Feeling increasingly at home, Ricci decided to make an attempt to enter the imperial city of Peking. His effort in 1595, however, was not successful because a Sino-Japanese conflict in Korea had made all foreigners suspect. He had to return from Peking to stop first at Nan-ch'ang and then Nanking. During his stay at Nan-ch'ang, from 1595 to 1598, he became a friend of two princes of the royal blood. At the request of one of them, he wrote his first book in Chinese, On *Friendship.* At Nanking, where he settled in February 1599, he was engaged chiefly in astronomy and geography. In his *History,* he commented on the effects of this work:

> The fathers gave such clear and lucid explanations on all these matters which were so new to the Chinese, that many were unable to deny the truth of all that he said; and, for this reason, the information on this matter quickly spread among all the scholars of China. From this one can understand how much esteem was given to the Jesuits as well as to our land which thenceforth they did not dare to describe as barbarian, a word they were accustomed to use in describing countries other than China.

Encouraged by the reception he received at Nanking, Ricci made a second attempt to reach Peking. He entered the city in January 1601, accompanied by his Jesuit colleague, the young Spaniard Diego Pantoja. Although Ricci was not received by the Emperor, he was given permission to remain in the capital. From then on, he never left Peking, and he dedicated the rest of his life to its people, teaching them science and preaching the gospel. His efforts to attract and convert the Chinese intelligentsia brought him into contact with many outstanding personalities, among them Li Chih-tsao, Hsü Kuang-ch'i, and Yang T'ing-yiin (who became known as the "Three Pillars of the Early Catholic Church" in China and who assisted the missionaries, especially in their literary efforts) and Feng Ying-ching, a scholar and civic official who was imprisoned in Peking. During his years in Peking, Ricci wrote several books in Chinese: "The Secure Treatise on God" (1603), "The Twenty-five Words" (1605), "The First Six Books of Euclid" (1607), and "The Ten Paradoxes" (1608). He died on May 11, 1610, and was granted a place for burial by imperial order.

The secret of Ricci's success was his ability to go beyond cultural barriers and befriend men of another race and religion. His remark about his friend Feng Ying-ching brings out well the spirit of this great missionary: "He treated the affairs of our fathers as if they were his own and our fathers in turn treated his as if they were ours."

BIBLIOGRAPHY. VINCENT CRONIN, *The Wise Man from the West* (1955), a popular but authoritative account of the life of Ricci; LOUIS J. GALLAGHER (trans.), *China in the Sixteenth Century: The Journals of Matthew Ricci, 1583–1610* (1953; trans. from TRIGAULT'S *De Christiana Expeditione apud Simas Suscepta ab Societate Jesu,* 1615), the standard English source, contains both Ricci's description of Ming China and his account of the history of the early Jesuit missionary activities in China; GEORGE H. DUNNE, *Generation of Giants: The Story of the Jesuits in China in the Last Decades of the Ming Dynasty* (1962), an authoritative and sympathetic account of the Jesuits' activities in late Ming China, with an excellent bibliography on Ricci (pp. 371–379); KENNETH SCOTT LATOURETTE, *A History of Christian Missions in China* (1929), a standard reference; PASQUALE

M. D'ELIA (ed.), *Fonti Ricciane,* 3 vol. (1942–49), the principal sourcebook, contains the original text of Ricci's journals and extensive commentaries; PIETRO TACCHI VENTURI (ed.), *Opere storiche del P. Matfeo Ricci, 2* vol. (1911–13), volume 2 contains the letters of Ricci and his companions and is an indispensable source.

(J.H.-s.S.)

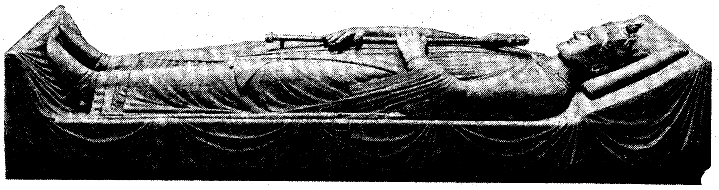
Richard I the Lion-Heart, of England

Although king of England from 1189 to 1199, Richard I, "Coeur de Lion," spent only six months of his reign in England and made little contribution to its governance. Yet his knightly manner and his prowess in the Third Crusade made him a popular king in his own time as well as the hero of countless romantic legends since. The third son of Henry II and Eleanor of Aquitaine, Richard was born on September 8, 1157. He was given the duchy of Aquitaine, his mother's inheritance, at the age of 11, and was enthroned as duke at Poitiers in 1172. Richard possessed precocious political and military ability, won fame for his knightly prowess, and quickly learned how to control the turbulent aristocracy of Poitou and Gascony. Like all Henry II's legitimate sons, he had little or no filial piety, foresight, or sense of responsibility. He joined his brothers in the great rebellion (1173–74) against their father, who invaded Aquitaine twice before Richard submitted and received pardon. Thereafter Richard was occupied with suppressing baronial revolts in his own duchy. His harshness infuriated the Gascons, who revolted in 1183 and called in the help of the "Young King" Henry and his brother Geoffrey of Brittany in an effort to drive Richard from his duchy altogether. Alarmed at the threatened disintegration of his empire, Henry II brought the feudal host of his continental lands to Richard's aid, but the younger Henry died suddenly (June 11, 1183) and the uprising collapsed.

Alliance with Philip II of France

Richard was now heir to England, and to Normandy and Anjou (which were regarded as inseparable), and his father wished him to yield Aquitaine to his youngest brother John. But Richard, a true southerner, would not surrender the duchy in which he had grown up, and even appealed, against Henry II, to his close friend, the young king of France, Philip II (Philip Augustus). In November 1188 he did homage to Philip for all the English holdings on French soil and in 1189 openly joined forces with Philip to drive Henry into abject submission. They chased him from Le Mans to Saumur, forced him to acknowledge Richard as his heir, and at last harried him to his death (July 6, 1189). Richard received Normandy on July 20 and the English throne on September 30. Richard, unlike his rival Philip of France, had only one ambition, to lead the crusade prompted by Saladin's capture of Jerusalem in 1187. He had no conception of planning for the future of the English monarchy and put up everything for sale to buy arms for the crusade. Yet he had not become king to preside over the dismemberment of the Angevin empire. He broke with Philip of France and did not neglect Angevin defenses on the Continent. Open war was averted only because Philip also took the cross. Richard dipped deep into his father's treasure, sold sheriffdoms and other offices. With all this he raised a formidable fleet and army and in 1190 departed for the Holy Land, travelling via Sicily.

Richard found the Sicilians hostile and took Messina by storm (October 4). To prevent the German emperor Henry VI from ruling their country, the Sicilians had elected the native Tancred of Lecce, who had imprisoned the late king's wife, Joan of England (Richard's sister), and denied her possession of her dower. By the Treaty of Messina Richard obtained for Joan her release and her dower, acknowledged Tancred as king of Sicily, declared Arthur of Brittany to be his own heir, and provided for Arthur to marry Tancred's daughter. This treaty infuriated the Germans, who were also taking part in the Third Crusade, and it incited Richard's brother John to treachery and rebellion. Richard joined the other crusaders at Acre on June 8, 1191, having conquered Cyprus on his way there (see CRUSADES). While at Limassol in Cyprus, Richard married (May 12) Berengaria of Navarre.


Richard I the Lion-Heart, tomb effigy in the abbey church of Fontevrault-l'Abbaye, France.
Giraudon

Results of Crusade

Acre fell in July 1191, and on September 7 Richard's brilliant victory at Arsūf put the crusaders in possession of Joppa. Twice Richard led his forces to within a few miles of Jerusalem. But the recapture of the city, which constituted the chief aim of the Third Crusade, eluded him. There were fierce quarrels among the French, German, and English contingents. Richard insulted Duke Leopold of Austria by tearing down his banner and quarrelled with Philip Augustus, who returned to France after the fall of Acre. Richard's candidate for the crown of Jerusalem was his vassal Guy de Lusignan, whom he supported against the German candidate, Conrad of Montferrat. It was rumoured, unjustly, that Richard connived at Conrad's murder. After a year's unproductive skirmishing, Richard (September 1192) made a truce for three years with Saladin that permitted the crusaders to hold Acre and a thin coastal strip and gave Christian pilgrims free access to the holy places.

Richard sailed home by way of the Adriatic, because of French hostility, and a storm drove his ship ashore near Venice. Because of the enmity of Duke Leopold of Austria he disguised himself, but was discovered at Vienna in December 1192 and imprisoned in the duke's castle at Diirrenstein on the Danube. Later, he was handed over to Henry VI, who kept him at various imperial castles. It was around Richard's captivity in a castle, whose identity was at first unknown in England, that the famous romance of Blondel was woven in the 13th century.

Under the threat of being handed over to Philip II, Richard agreed to the harsh terms imposed by Henry VI: a colossal ransom of 150,000 marks and the surrender of his kingdom to the emperor on condition that he receive it back as a fief. The raising of the ransom money was one of the most remarkable fiscal measures of the 12th century and gives striking proof of the prosperity of England. A very high proportion of the ransom was paid, and meanwhile (February 1194) Richard was released.

Return to England

He returned at once to England and was crowned for the second time on April 17, fearing that the independence of his kingship had been compromised. Within a month he went to Normandy, never to return. His last five years were spent in warfare against Philip II, interspersed with occasional truces. The king left England in the capable hands of Hubert Walter, justiciar and archbishop of Canterbury. It was Richard's impetuosity that brought him to his death at the early age of 42. The vicomte of Limoges refused to hand over a hoard of gold unearthed by a local peasant. Richard laid siege to his castle of Châlus, and in an unlucky moment was wounded. He died on April 6, 1199. He was buried in the abbey church of Fontevrault, where Henry II and Queen Eleanor are also buried, and his effigy is still preserved there.

Richard was a thoroughgoing Angevin, irresponsible and hot-tempered, possessed of tremendous energy, and capable of great cruelty. He was more accomplished than most of his family, a soldier of consummate ability, a skillful politician, and capable of inspiring loyal service. He was a lyric poet of considerable power and the hero of troubadours. In striking contrast with his father, and with King John, he was, there seems no doubt, a homosexual. He had no children by Queen Berengaria, with whom his relations seem to have been merely formal.

BIBLIOGRAPHY. K. NORGATE, *Richard the Lion Heart* (1924), a very full, somewhat old-fashioned narrative, strongly based on chronicle and other recorded sources; F.M. POWICKE, *The Loss of Normandy, 1189–1204,* 2nd ed. (1961), a brilliant survey of the Angevin Empire on the eve of its disintegration that illustrates Richard's strategic and tactical

skill; L. LANDON, ***Itinerary of King Richard I*** (1935), essential basic information, dating the King's movements, and listing his charters; S. RUNCIMAN, ***A History of the Crusades,*** vol. 3 (1954), a highly readable, reliable, mainly political narrative, beginning with a good account of the Third Crusade; AMY KELLY, ***Eleanor of Aquitaine and the Four Kings*** (1950), a readable and informative narrative on the Angevin Empire from the unusual viewpoint of Eleanor of Aquitaine, Richard's mother.

(G.W.S.B.)

Richardson, Henry Hobson

One of the leading architects in the United States following the Civil War, Henry Hobson Richardson designed buildings that were influential in leading to the development of an indigenous modem American architectural style. The profuse imitation of the most obvious details of his designs led to the Romanesque revival of the 1880s and 1890s. Their more essential qualities, their generally simple horizontal silhouettes, and the external colours and textures of natural materials, however, set Richardson's buildings apart from the bulk of American architecture in the last quarter of the 19th century. His works also had a stimulating effect upon the work of the internationally influential architects Louis H. Sullivan (1856–1924) and Frank Lloyd Wright (1867–1959), the American pioneers of modern architectural design.

Richardson was born on September 29, 1838, at Priestley Plantation on the Mississippi River in St. James Parish, Louisiana. He was the great-grandson of the 18th-century political and religious dissenter and discoverer of oxygen. Joseph Priestley (1733–1804). His distinguished pedigree and his own affability made his move from the South to Harvard University in 1855 as easy as it was eventually to be rewarding. Harvard then offered more in personal contacts than in intellectual stimulation, and Richardson's later clients, such as Henry Adams (1838–1918), were largely drawn from the Porcellian Club and other social circles that he entered with ease. He never returned to the South.

Studies in Paris

Sometime during his Harvard days Richardson decided to become an architect. In Boston he was surrounded by buildings of plain granite design that affected the best of his own later work, but for formal training he had to go abroad, for there were no schools of architecture in the United States before the Civil War. Fluent in French from his Louisiana childhood, he studied at the École des Beaux-Arts in Paris from 1860 to 1862, when the military events at home cut off his income. He then worked in the office of the French architect Theodore Labrouste (1799–1885). Richardson remained with Labrouste until he returned to the United States in October 1865. In Paris he mastered the analytical architectural planning that characterizes much of his mature work and that was formulated by his friend the architect and École professor Julien Guadet (1834–1908) in his ***Éléments et théorie de l'architecture,*** published in Paris in 1902.

Richardson returned to America with every expectation of quick success, for he was among the best trained architects in the country and had many important connections. In November 1866, he was awarded his first commission, the Church of the Unity in Springfield, Massachusetts (now demolished), where one of his former classmates was an influential member of the congregation. His career launched, Richardson married Julia Gorham Hayden of Boston on January 3, 1867. They moved into a house of his own design (now altered) on Staten Island, New York, where five of his six children were born. Richardson's neighbour was Frederick Law Olmsted (1822–1903), the journalist and renowned landscape architect with whom he later frequently collaborated.

H.H. **Richardson.**

By courtesy of The Prairie School Press, Palos Park, Illinois

Richardson lived and worked in New York City for the next eight years, forming, in October 1867, a partnership with the architect Charles D. Gambrill (1832–80) that lasted exactly eleven years, but was never more than one of administrative convenience. From his Manhattan office and the drafting board in his Staten Island home came the drawings for the early commissions in Springfield, the State Asylum for the Insane in Buffalo, New York, and the Brattle Square and Trinity churches in Boston. Designed for the renowned preacher Phillips Brooks (1835–1893), Trinity was one of the most important Episcopal churches in America. Richardson's Romanesque revival design won him a national reputation, many imitators, and so many New England commissions that it became desirable to move to the Boston area. In 1874 he bought an early-19th-century dwelling resembling the plantation houses of the South in suburban Brookline. He added to it his office and studio. His last twelve years were spent with his family and assistants in these busy surroundings so reminiscent of his childhood environment.

The Boston years

During these last years Richardson produced the buildings upon which his reputation principally rests. He designed houses, community libraries, suburban railroad stations, educational buildings, and commercial and civic structures. Instead of the splintered massing, narrow vertical proportions, and disparate Gothic features used by his contemporaries, he favoured horizontal lines, simple silhouettes, and uniform, large-scale details of Romanesque or Byzantine inspiration. Since his best commercial structure, the Marshall Field Wholesale Store in Chicago, and most of his railroad stations in the northeastern United States were demolished long ago, the development of Richardson's work in the last years of his life can now best be studied at Sever and Austin halls at Harvard University; at the Allegheny County Buildings in Pittsburgh, Pennsylvania; at the Glessner House in Chicago; or in the series of libraries in the small towns around Boston, from Woburn and North Easton to Quincy and Malden. The Crane Memorial in Quincy, with its tripartite layering of a rough-faced granite base beneath continuous clerestory windows topped with a tiled gable roof and its cavernous entrance arch, stands with the finest and most characteristic works of his maturity.

Character and personal life

Richardson was a sociable and colourful man. A figure of immense girth, he was known for his hearty appetite for food and drink. He enjoyed entertaining his friends, such as the American sculptor Augustus Saint-Gaudens, with large quantities of champagne and gourmet cuisine. On special occasions he served such delicacies as Chesapeake Bay oysters especially shipped from Baltimore, or would have a terrapin prepared by a Philadelphia restaurant and delivered in person by its chef. Such living hastened his end, however, for he suffered throughout his career from chronic nephritis, or Bright's disease, and worked at a strenuous pace, taking just one vacation during the 1880s: a whirlwind tour of Romanesque architecture in Europe in the summer of 1882. He died four years later, bloated and fatigued, on April 27, 1886, just 48 years old, at the top of his profession and with major buildings rising in Boston; Pittsburgh, Pennsylvania; Cincinnati, Ohio; Chicago; and St. Louis, Missouri. He left it to his successors, the Boston architectural firm of Shepley, Rutan, and Coolidge, to finish these, and to the Chicago architects Sullivan and Wright to carry on in the direction he had initiated.

MAJOR WORKS

Brattle Square Church, Boston (1870–72); State Asylum for the Insane, Buffalo, N.Y. (designed 1870–72); Hampden County Courthouse, Springfield, Mass. (1871–73); North Congregational Church, Springfield, Mass. (built 1872–73); Trinity Church, Boston (designed 1872, built 1873–77); Wm. Watts Sherman House, Newport, R.I. (1874–75); New York State Capitol, Albany, completion with Eidlitz (1875); Ames Memorial Library, North Easton, Mass. (1877–79); Winn Memorial Library, Woburn, Mass. (1877–78); Sever Hall, Harvard University (1878–80); Dr. James Bryant House, Cohasset, Mass. (1880); Crane Memorial Library, Quincy, Mass. (1880–83); Ames Gate Lodge, North Easton, Mass. (1880–81); City Hall, Albany, N.Y. (1880–82); Boston & Albany Railroad Station, Auburndale, Mass. (1881); Austin Hall, Harvard University (1881–83); M.F. Stoughton House, Cambridge, Mass. (1882–83); Allegheny County Jail, Pittsburgh, Pa. (1884–86); Allegheny County Courthouse, Pittsburgh, Pa. (1884–87); Marshall Field Wholesale Store, Chicago (1885–87); Glessner House, Chicago (1885–87).

BIBLIOGRAPHY. Richardson's surviving major works are in greater Boston, North Easton, Massachusetts, Albany, Buffalo, Pittsburgh, and Chicago. Surviving sketches and drawings are mainly at Houghton Library, Harvard University and Shepley, Bulfinch, Richardson, and Abbott, Boston. M.G. VAN RENSSELAER, *Henry Hobson Richardson and His Works* (1888), is a contemporary, and a fundamental work for the study of Richardson. HENRY-RUSSELL HITCHCOCK, *The Architecture of H.H. Richardson and His Times*, rev. ed. (1961; 1966), is the standard critical study. For Richardson's influence on Sullivan and Wright, see JAMES F. O'GORMAN, "Henry Hobson Richardson and Frank Lloyd Wright," *Art Quarterly*, 32:292–315 (1969).

(J.F.O'G.)

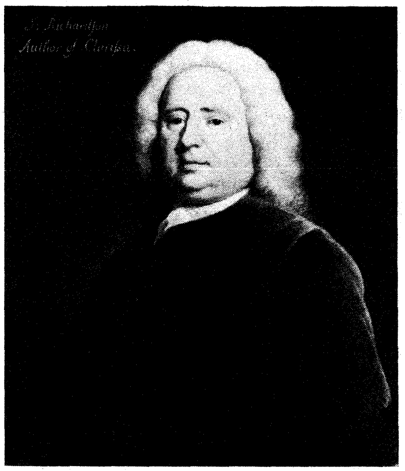
Richardson, Samuel

Samuel Richardson's *Pamela* is often credited with being the first English novel. Although the validity of this claim depends on the definition of the term novel, it is not disputed that Richardson was innovative in his concentration on a single action, in this case a courtship. By his use of the epistolary technique—the story is told in the form of letters—he provided if not the "stream" at least the flow of consciousness of his characters, and he pioneered in showing how his characters' sense of class differences and their awareness of the conflict between sexual instincts and the moral code created dilemmas that could not always be resolved. These characteristics reappear regularly in the subsequent history of the novel.

Richardson was 50 years old when he wrote *Pamela*, but about these 50 years little is known. His ancestors were of yeoman stock. His father, also Samuel, and his mother's father, Stephen Hall, became London tradesmen; and his father, after the death of his first wife, married Stephen's daughter, Elizabeth, in 1682. A temporary move of the Richardsons to Derbyshire accounts for the fact that the novelist was born in Mackworth, near the city of Derby, on August 19, 1689. They returned to London when Richardson was 10. He had at best what he called "only Common School-Learning," but at 13 he said he had gained the reputation among the girls in his London neighbourhood of being a good story-teller and ready to write love letters for them, marked with some of the same obliquity that later led Samuel Johnson to say of Clarissa that "there is always something which she prefers to truth."

Richardson was bound apprentice to a London printer, John Wilde. Sometime after completing his apprenticeship he became associated with the Leakes, a printing family whose presses he eventually took over when he set up in business for himself in 1721 and married Martha Wilde, the daughter of his master. Elizabeth Leake, the sister of a prosperous bookseller of Bath, became his second wife in 1733, two years after Martha's death. His domestic life was marked by tragedy. Of the six children of his first marriage, four died within a year of birth, and the others at the ages of two and three. By his second wife he had four daughters who survived him, but two other children (one of them the fourth son to be named after him) died in infancy. These and other bereavements contributed to the nervous ailments of his later life.

Domestic tragedy



Samuel Richardson, oil painting by J. Highmore (1692–1780). In the Nationai Portrait Gallery, London.
By courtesy of the National Portrait Gallery, London

In his professional life, Richardson was hardworking and successful. With the growth in prominence of his press went his steady increase in prestige as a member, an officer, and later as master, of the Stationers' Company (the guild for those in the book trade). During the 1730s his press became known as one of the three best in London, and with prosperity he moved to a more spacious London house and leased the first of three country houses in which he entertained an admiring group of young girls and a circle of friends that included Dr. Johnson, the painter William Hogarth, the actors Colley Cibber and David Garrick, Edward Young (many of whose poems he printed, including the celebrated Night Thoughts), and Arthur Onslow, speaker of the House of Commons, whose influence in 1733 helped to secure for Richardson lucrative contracts for government printing that included later the journals of the House.

In this same decade he began writing in a modest way, undertaking some editing and producing what he called "a few other little things of the Pamphlet kind." More importantly he was commissioned to write a collection of letters that might serve as models for "country readers," a volume that has become known as *Familiar* Letters on *Important Occasions.* Occasionally he hit upon continuing the same subject from one letter to another, and, after a letter from "a father to a daughter in service, on hearing of her master's attempting her virtue," he supplied the daughter's answer. This was the germ of his novel *Pamela*. With a method supplied by the letter writer and a plot by a story that he remembered of an actual serving maid who preserved her virtue and was rewarded by marriage, he began writing the work in November 1739 and published it as *Pamela:* or; Virtue Rewarded, a year later.

Publication of *Pamela*

On the death of Pamela's mistress, Richardson relates, her son, Mr. B, begins a series of mild stratagems designed to end in Pamela's seduction, but, these failing, he abducts her and renews his siege in earnest. Pamela preserves her virtue, and halfway through the novel Mr. B offers marriage. In the second half Richardson shows Pamela winning over those who had disapproved of the misalliance. Though the novel was immensely popular, Richardson was adversely criticized by those who thought his heroine a calculating minx or his own morality dubious. Actually his heroine is a blend of the artful and the artless. She is a sadly perplexed girl of 15, with a divided mind, who faces a real dilemma because she wants to preserve her virtue without losing the man with whom she has fallen in love. Since Richardson wrote the novel from Pamela's point of view, it is less clear that Mr. B's problem arises from his having fallen in love with a servant, who, traditionally, would have been merely a target for seduction. The author resolved the conflicts of both characters too facilely. perhaps, because he was firmly committed to the plot of the true story he had remembered. When the instantaneous popularity of *Pamela* led to a spurious continuation of her story, he wrote his own sequel, *Panzela in her Exalted Condition* (1742), a two-volume work that did little to enhance his reputation.

His later heroine Clarissa Harlowe

By 1744 Richardson seems to have completed a first draft of his second novel, *Clarissa.* **Or,** *The History of a Young Lady,* but he spent three years trying to bring it within the compass of the seven volumes in which it was published. He presents the heroine, Clarissa Harlowe, first when she is discovering the barely masked motives of her family, who would force her into a loveless marriage to improve their fortunes. Outside the orbit of the Harlowes stands Lovelace, nephew of Lord M and a romantic who held the code of the Harlowes in contempt. In her desperate straits, Clarissa appraises too highly the qualities that set Lovelace beyond the world of her family, and, when he offers protection, she runs off with him. She is physically attracted by if not actually in love with Lovelace and is responsive to the wider horizons of his world, but she is to discover that he wants her only on his own terms. In Lovelace's letters to his friend Belford, Richardson shows that what is driving him to conquest and to rape, finally, is really her superiority. In the correspondence of Clarissa and her friend Anna Howe, Richardson shows the distance that separates her from her confidant, who thinks her quixotic in not accepting a marriage, but marriage as a way out would have been a sacrifice to that same consciousness of human dignity that had led her to defy her family. As the novel comes to its long-drawn-out close, she is removed from the world of both the Harlowes and the Lovelaces, and dies, a child of heaven. In providing confidants for his central characters and in refusing to find a place in the social structure into which to fit his sorely beset heroine, Richardson made his greatest advances over *Pamela.*

In the years after writing *Pamela,* Richardson broadened his circle of correspondents, and his third novel was his bow to the requests for the hero as a good man, a counter-attraction to the errant hero of Henry Fielding's *Tom Jones* (1749). Fielding had been among those who thought Pamela a scheming minx, as he had shown in his parody, *An Apology for the life of Mrs. Shamela Andrews* (1741) and in the opening of his *Joseph Andrews* (1742). At the time Fielding had thought Cibber to be the author of the anonymously published *Panzela,* but in spite of Fielding's critical praise of *Clarissa* and the friendship that later developed between Richardson and Fielding's sister, Sarah, Richardson never forgave the author of what he stigmatized as "that vile Pamphlet Shamela." In *The History of Sir Charles Grandison* he provides a hero who is a model of benevolence. He faces little that a good heart cannot remedy and extricates himself from the nearest thing to a dilemma that he has to encounter: a "divided love" between an English woman, Harriet Byron, and an Italian, Signora Clementina. He is saved for Harriet by the last-minute refusal of the Roman Catholic Clementina to marry a firmly committed English churchman. The uneasy minds of Clementina and Harriet are explored with some penetration, but Sir Charles faces nothing in his society or within himself that requires much of a struggle. Furthermore his dilemma is not so central to the novel as were those of Pamela and Clarissa. He is surrounded with a large cast of characters who have their parts to play in social comedy that anticipates the novel of manners of the late 18th century.

Influence and reputation

Richardson had disciples when he died, at Parson's Green, near London, on July 4, 1761. Some of them show the influence of *Clarissa,* which seems to have been most responsible for the cult of Richardson that arose on the continent. It was *Grandison,* however, that set the tone of most of Richardson's English followers and for Jane Austen, who was said to have remembered "every circumstance" in this novel, everything "that was ever said or done." By the end of the 18th century, Richardson's reputation was on the wane both in England and abroad, but the 20th century has begun to awaken to the psychological subtleties of *Pamela* and *Clarissa.*

MAJOR WORKS

NOVELS: *Pamela: or, Virtue Rewarded,* 2 vol. (1740, dated 1741); *Clarissa. Or, the History of a Young Lady,* 7 vol. (1747–48, dated 1748); *The History of Sir Charles Grandison,* 7 vol. (1753–54, dated 1754).

OTHER WORKS: *Letters Written to and for particular friends, on the Most Important Occasions: Directing Not Only the Requisite Style and Forms To Be Observed in writing 'Familiar Letters', but How to Think and Act Justly and Prudently, itt the Common Concerns of Human Life* (1741), known as *Familiar Letters on Important Occasions,* a collection of model letters; issue no. 97 (1751) of the literary periodical, *The Rambler,* its subject being advice to unmarried ladies.

BIBLIOGRAPHY. WILLIAM M. SALE, *Samuel Richardson* (1936, reprinted 1969), is an analytical bibliography with historical notes; and ALAN D. MCKILLOP, *Samuel Richardson* (1936, reprinted 1960), a reliable and thorough account of Richardson's life as a novelist, contains a good bibliography of contemporary and later criticism and foreign translations. The largest collection of manuscript correspondence is in the Victoria and Albert Museum, London. The Shakespeare Head edition of the novels (18 vol., 1929–31) is standard. A selection of the correspondence, ed. by ANNA BARBAULD (6 vol., 1804, reprinted 1966) is valuable, but not always trustworthy. JOHN CARROLL (ed.), *Selected Letters* (1964), is excellent. T.C. DUNCAN EAVES and BEN D. KIMPEL, *Samuel Richardson: A Biography* (1971), is the most comprehensive life story and will be definitive for many years. It contains an excellent census of his correspondence. WILLIAM M. SALE, *Samuel Richardson: Master Printer* (1950), is an account of Richardson's professional career with a list of books from his press. BRIAN W. DOWNS, *Richardson* (1928, reprinted 1969), provides a good critical discussion of the novels as does an essay in ALAN D. MCKILLOP, *Early Masters* of *English Fiction* (1956). IAN P. WATT, *The Rise of the Novel* (1957), contains brilliant discussions of *Pamela* and *Clarissa.* The range of modern critical essays is excellently represented in JOHN CARROLL (comp.), *Samuel Richardson* (1969).

(W.M.S.)

Richelieu, Cardinal de

Both as statesman and churchman, Armand-Jean du Plessis, Cardinal de Richelieu, is the acknowledged architect of France's greatness in the 17th century and a contributor to the secularization of international politics during the Thirty Years' War. His success in bringing political order out of chaos and in bringing about the eclipse of Spain, France's paramount rival, made him a controversial figure to his contemporaries as well as to historians.

Heritage and youth. The family of du Plessis de Richelieu was of insignificant feudal origins but by intermarriage with the legal and administrative classes had risen to some prominence and had acquired the seigneury of Richelieu in Poitou. It was perhaps here, or more likely in Paris, where his father, François du Plessis, seigneur de Richelieu, was grand provost (chief magistrate)

Giraudon

Richelieu, portrait by Philippe de Champaigne (1602–74). In the Louvre, Paris.

to Henry III, that Armand-Jean du Plessis was born on September 9, 1585. His mother, Suzanne de la Porte, was the daughter of a councillor of the Parlement of Paris (the supreme judicial assembly, second only to the king's council), and in his intelligence, administrative competence, and instinct for hard work, he resembled his middle class ancestors.

He was five years old when his father died, leaving estates that had been ruined by inflation, extravagance, and mismanagement during the Wars of Religion, (1562–98), and he was conscious from his earliest years of the threat of penury. This inspired in him the ambition to restore the honour of his house and evoked in him the sense of grandeur he was to attribute vicariously to France. His provident mother, with three boys and two girls—all under the age of 13—set about reorganizing the family's precarious resources. The principal of these was the benefice of the bishopric of Luçon near La Rochelle, which had been granted by Henry III to the Richelieu family under the Concordat of 1516 (Concordat of Bologna). Unrest of the cathedral chapter threatened a revocation of the grant, and it became necessary for a member of the family to be consecrated bishop as soon as possible. Henri, the eldest son, was heir to the seigneury of Richelieu; and Alphonse, the second son, had become a Carthusian monk; so the obligation fell on Armand-Jean, who was a student at a cadet school for nobles, destined for the army and the court.

Early career* in *the church. The prospect of a career in the church was not displeasing to the thin, pale, and at times sickly boy, for he had an inclination toward learning, a facility for debate, and a relish for the prospect of governing the lives of others. So when he entered the Collège de Calvi to study theology he prepared himself well for a future that he did not doubt would encompass more than the bishopric of Luçon. Because he was below the canonical age for consecration upon the completion of his studies, he needed a papal dispensation. To gain it he went to Rome, where Paul V fell victim to the young man's skill as a charmer. Historians have been tempted to suppose that his personal experiences with the papal chancellery bred in Richelieu a distrust of Roman politics that was to lead him into persistent opposition to the Holy See in the days of his power, but there is no evidence of this. On April 17, 1607, at the age of 22, he was ordained priest and consecrated to the see of Luçon. He found on his arrival a diocese ruined by the Wars of Religion, a hostile chapter, and a demoralized clergy, but his opponents quickly succumbed to the unaccustomed authority that radiated from the episcopal palace.

Bishop of Luçon

The enemies of Richelieu in his own lifetime and some historians, bred on the legends propagated by them, have supposed that Richelieu set out to demonstrate his administrative capacity in the only situation available to him. That he was ambitious to an extraordinary degree, and ruthless in pursuit of his ambitions, is beyond dispute, but to suggest that he was motivated, either at Luçon or later, [illegible]l by [illegible] aggrandizement is to misunderstand him and th[illegible] age. For in [illegible] li[illegible] as in hi[illegible] contemporaries, ambition and piety were compatible. A son of the Counter-Reformation, he was religious; he believed intensely in God; and his instinct for certitude and authority made him particularly susceptible to the discipline of theology. He formed an association with the leaders of the mystical movement in the church, though he was too much a rationalist to be affected by their practices. He was the first bishop in France to implement in his diocese the reforms decreed by the Council of Trent, and he was also the first theologian to write in French and to establish the conventions of vernacular theological exposition.

Religious writings

In his own time his theological writings were esteemed. There were four of them—two published when he was in disfavour and administratively inactive, in 1618 and in 1621; the other two were written when he experienced an intensification of religious conviction in the dark days of the Thirty Years' War shortly before his death and were published posthumously. They reveal a mind of apologetic rather than mystical bent, concerned with truth, and confident that doctrinal error, once exposed and recognized, would be instantly repudiated. But he was also concerned with the perfection of the soul and with cultivation of the virtues. His catechism was used throughout the 17th century and was highly influential.

At least the earlier of his writings may have been prompted by the fact that Richelieu's energy and compulsion to influence events had to find an outlet when no avenue other than writing was open to him. Intellectual though he was, the pursuit of knowledge undirected to any immediate end was insufficient to his needs. When he studied contemporary theories of government of church and state it was with a view to the refutation of ideas that he felt were subversive of that natural order that could exist only in a society in which the crown was responsible only to God.

Inactivity, anxiety, and frustrated ambition produced in Richelieu recurrent periods of ill health. For long periods, both at Luçon and at the height of his power, he was in retreat because of, or on the pretext of, ill health. Although possessed of hypnotic power to charm when disposed lo dc so, he increasingly found the emotional cost of personal confrontation too great, and he preferred to argue his case on paper, The consequent remoteness made him appear to be a sinister figure in a web of espionage and intrigue that enmeshed everyone in France—a mandarin surrounded by his 14 cats, all-knowing, all-seeing, vindictive, unscrupulous, and ruthless.

This is the picture that the gossips of his own time, mostly involved in factions that he sought to suppress, transmitted to posterity. The reality was a hard-working, conscience-stricken man, combatting powerful forces that were dedicated to divisive political and social ends—a man obsessed with order as a superior moral end.

Rise to power. The France on which the Bishop of Luçon pondered gave every indication of falling again into the disorder of the Wars of Religion. The assassination of Henry IV in 1610 released separative forces that were endemic in the administrative system. The government of the queen mother, Marie de Médicis, as regent for Louis XIII, was inefficient and corrupt, and the magnates of the realm, satraps in their own provinces, motivated by personal and regional self-interest, struggled to control or combat it. Their disobedience was always accompanied by predatory expeditions of armed men and labyrinthine negotiations with the court; and on one of these occasions the Bishop of Luçon found himself an intermediary, which led to his being elected one of the representatives of the clergy of Poitou to the States General (assembly of the representatives of the French estates) of 1614. He put all his energy into persuading the assembly of his talents and the court of his support for royal authority. In a clash between the clergy and the Third Estate (consisting of the middle classes, artisans, and peasants) on the subject of the relationship between the crown and the papacy he played a conciliatory role, and he was prominent in moves of the clergy to persuade the Third Estate that the decrees of the Council of Trent should be promulgated. The outcome was that he was chosen to present the final address of the First Estate (the clergy) at the closing session on February 23, 1615.

Election to the Estates General

Some months later he was appointed chaplain to the new queen, Anne of Austria. This was an office of no political consequence, but it held the promise of eventual entry into the royal council, which, Richelieu had argued at the States General, should accord first place to prelates of distinction. He has been charged with basely promoting his own advancement through flattery of the inconstant Queen Mother and her favourites. In his time there was nothing unusual in this, for to dissemble was both a Baroque convention and a political necessity universally practiced, although rarely with Richelieu's skill. Clever negotiations with another disobedient faction led to his appointment as a secretary of state in 1616.

Secretary of state. Up to this time Richelieu had had no insight into international relations, and the regard for Spain with which he was credited was probably genuine because he had had no occasion to question Spain's ambitions. His year of office, however, coincided with war between Spain (ruled by a Habsburg dynasty) and Venice, which invoked its alliance with France. The resultant diplomatic involvement persuaded Richelieu of three interconnected factors: the vulnerability of France to Habsburg political and economic encirclement; the domestic ramifications of various European movements in the religious controversy between Catholics and Protestants; and the dependence of the small states in France's borderlands upon an equilibrium of power between France and Spain.

Cardinalate

Richelieu's tenure of office was abruptly terminated in April 1617 when a palace revolution overthrew the regency of Marie de Médicis and brought to power a new coalition of politicians under the direct government of Louis XIII. Richelieu was first banished to Luçon and then exiled to the papal city of Avignon, where he sought distraction from his melancholy in theological writing. A new rebellion of the princes, gravitating this time to Marie de Médicis as the focus of opposition to the royal council, led in 1619 to the King recalling Richelieu to his mother's entourage on the assumption that he would exercise a moderating influence upon her. The ascendency that he gained over her, however, did not lead to her submission but the contrary. There followed four years of intricate negotiation and even overt hostilities during which the King's nomination of Richelieu for a cardinal's hat became one of the issues involved in a settlement. A revolt of the Huguenots and the death of the King's favourite brought about Marie de Médicis' recall to the council and Richelieu's promotion, although for a time the King kept him at a distance.

First minister of France. In 1624 another crisis, over the Valtellina in northern Italy, led to a ministerial reconstruction and to the Cardinal's appointment as secretary of state for commerce and marine and chief of the royal council. Four years later the title of first minister was to be created for this office. The Valtellina controversy occurred when the Protestant Swiss canton of Grisons invoked a treaty of protection with France against Spanish ambitions in the valley of that name. The struggle had ramifications both within France and throughout Europe as the Protestants made common cause with Grisons and the Catholics with the Habsburgs. Richelieu recognized that vacillation in this crisis would threaten domestic stability, and so he struck in the Valtellina, expelling the papal troops from their fortresses. It was an action that created an enormous impression in Europe and gained for Richelieu an instant reputation for decision and ruthlessness.

It also disillusioned those who had seen in him a defender of Catholic interests and of a Franco-Spanish alliance and provoked the indignation of zealous Catholics, who were horrified at this assault on the Pope's army. The fractious and the pious now coalesced in yet another of the seditious movements that had long debilitated France, and they influenced Marie de Médicis against her erstwhile client who, now that he was invested with the responsibility of government, no longer deferred to her or permitted her whims and emotions to influence him.

The Valtellina affair gave Richelieu the opportunity to demonstrate his theory that diplomacy consists in patient and protracted negotiation from positions of strength and led to his intervening, through the medium of hired pamphleteers, in a paper war on the morality of statecraft. The use of force by princes, he argued, was morally justified as an essential step in the establishment of the peace of their subjects and the security of their states. In matters in which vital national interests are involved, less strict rules of behaviour are incumbent than in other matters. "Although in the ordinary course of affairs justice requires authentic proof," he wrote, "the same is not true of those which concern the State because in such case persuasive inference must sometimes be held to be sufficient, for parties and cabals which are formed against public security ordinarily act with such cunning and secrecy that there is never any evident proof save in the event, when the matter is beyond remedy." A statesman must not have "a cringing or scrupulous conscience" but must "pursue great things with ardour commensurate with the wisdom of the judgment" which he makes.

Conspiracies against Richelieu

From his first days in office Richelieu was the object of conspiracies to remove him, by assassination if other methods failed, and this continual threat was met with the courage that "requires that man must be exempt from weakness and fear." The success of Richelieu's security organization in ferreting out the disaffected, his implacable pursuit of his foes, his suspicion of everyone, and his manipulation of state trials made him misunderstood, feared, and detested. Yet according to the standards of the age his administration of justice did not depart from the moral principles that he believed to underlie all government; and while he pursued his objective of the integrity of France with tenacity he was never unaware of the problems of conscience, which he resolved by reference to the criterion of "probabilism"—the theory that in moral questions, where certainty is impossible, any course may be followed that is seen as solidly probable —which was current in Catholic theological exposition.

Measures for internal reform. The goals that Richelieu set himself were to counter Habsburg hegemony in Europe, which threatened France's independence of action, and "to make the King absolute in his kingdom in order to establish therein order and rule, to which his conscience obliges him." This implied not only the reduction to submission of the great nobles but also the subversion of the antiquated political institutions that fragmented France and weakened its government, particularly the provincial estates and parlements. But at no time was Richelieu sufficiently powerful to achieve his domestic ends by direct and overt measures. A respecter of law and history, he accepted the necessity of working within the traditional framework of administration. His sense of the feasible and his gift for seeing both sides of a question resulted in a pragmatism in practice that often contradicted his proclaimed theories, and he confused his critics by unexpected compromise and moderation and caused them to misinterpret his empiricism as opportunism and equivocation.

The extent to which Richelieu overcame the inertia of the governmental system is questionable. His use of commissions of justice to station in the provinces officers answerable to the central government was only the revival of an older expedient spasmodically employed in the past. They'were not regularly put to use until 1635—and then only as an extraordinary war measure—and were not systematized until the last year of his life, but they did eventually lead to government by intendants (provincial administrators appointed by the crown) in the reign of Louis XIV. At no time was he able overtly to challenge a provincial government, except in the case of Languedoc in 1632, when the Duc de Montmorency, encouraged by the provincial estates over an issue of tax collection by royal instead of provincial authorities, rebelled and lost his head.

Richelieu's great intellectual capacity enabled him to penetrate to the essence of events, and his tremendous will power, overcoming illness and hostility, drove him to incessant work and relentless exercise of his power. In his theory of politics he shared the rationalism of contemporary philosophers, believing that "the light of natural reason enables everyone to know that, since man is endowed with reason, he must do nothing except by reason." While he did not doubt the capacity of the mind to know what is naturally enjoined, he participated in the prevailing pessimism about man's will to act accordingly. Hence the refractory must be ruled so that they might not act "contrary to nature, and, as a result, contrary to Him Who is its Author." This twofold view of moral causes, the natural and divine, provided a philosophical axiom for state supervision of conduct in both the secular and the spiritual spheres. Sin and civil disobedience were, to Richelieu, but two aspects of disorder.

Struggle against the Huguenots. The gravest divisive factor in French society was religion. To Richelieu, the Huguenots constituted a state within a state, entrenched in 150 strongholds, with the civil government of major cities in their hands and considerable military force at their disposal. Yet Richelieu was prepared to tolerate this religious dissent so long as it did not amount to a political challenge. In this attempt to preserve social harmony at the expense of confessional difference he failed at first, for the Huguenot community was foolishly drawn into the intrigues of the Protestant magnates, who instigated England for the most trivial of reasons to war with France. Richelieu laid siege in 1628 to La Rochelle, the Huguenot centre, but it took a year to reduce the city, during which time Spain took advantage of the distraction to extend its hegemony in northern Italy at the expense of France's allies. While promising Richelieu help to combat the Protestants, Spain in fact subsidized their leaders in order to keep the French government preoccupied, and seized the strategic fortress of Casale in northern Italy. Again Richelieu acted with surprising vigour. The moment La Rochelle fell he led the army in winter over the Alps and checked the Spanish design. This reverse was countered by the Habsburgs with the introduction of imperial garrisons into parts of the duchy of Lorraine, which were claimed as fiefs of France. There followed intricate diplomatic manoeuvres, culminating in Richelieu's dramatic refusal to ratify the peace Treaty of Regensburg in 1630, and the Habsburgs' appeal to Pope Urban VIII to excommunicate Louis XIII for this supposed breach of faith.

This was Richelieu's moment of greatest political insecurity. His relationship with the King was distant and at times tense. The Catholic zealots did not share his distrust of Spain, and, disturbed at the implications of Richelieu's foreign policy, they provoked Marie de Médicis into a state of hysteria concerning the man who she believed had ungratefully deprived her of influence. On Richelieu's return from Italy in 1630, she struck, exerting all her emotional energy to influence her son to dismiss his minister. It was called the Day of Dupes because the court believed that Richelieu, like the train of his predecessors, had fallen, but it was not so. The King perceived that the issue was his own independence or his mother's domination and that there was no one but Richelieu who could relieve him of the responsibility of decisions at a moment of bewildering complications. After a day of suspense, he supported the Cardinal and thereafter did not waver in his support, even though Richelieu was able to survive his suspicions and dislike only by consummate tact and prudence.

Day of Dupes

The Day of Dupes had vast political consequences. Marie de Médicis and the King's brother Gaston fled to the Spanish Netherlands, there to constitute a focus of sedition that Spain promoted by conspiracy and finance and that Richelieu countered by a fatal involvement with the enemies of the Habsburgs. The central objective of his foreign policy was to restore the political equilibrium in the empire that Habsburg victories had disturbed, and, to this end, fan the suspicions that the states of the German Catholic League entertained about imperial intentions so that they might make peace with the Protestants and leave the latter free to counterbalance the Emperor. Although Bavaria was disposed to seek French protection, the Emperor's military successes and the Edict of Restitution occasioned a new mutual antagonism of Catholics and Protestants, which made neutrality of the Catholic League a practical impossibility.

Franco-Habsburg War. Richelieu's German policy fell into ruins as a result of his grant of subsidies to Gustavus II Adolphus of Sweden, then engaged in the conquest of Pomerania. He intended Gustavus Adolphus to distract imperial forces away from the French borderlands while he recovered the French fiefs in Lorraine, and he believed that by making him his pensioner he could harness him to this limited design while he promoted peace in the Rhine and Danube valleys. French subsidies, however, only liberated Gustavus Adolphus from constraint, and he fell on southern Germany, became embroiled with the armies of the Catholic League, and so consolidated the imperial and Catholic cause. The war spilled over the Rhine, and France's client states were by degrees drawn into the Habsburg orbit. The seizure by Spain in 1635 of the Archbishop of Trier, who was under French protection, at length led to France's alignment with the Protestant powers in the Thirty Years' War.

Assessment of Richelieu's role in the war. This involvement on behalf of the Protestants was regarded by many Catholics in his own time and subsequently as a betrayal of the church by one of its princes, and Richelieu has been criticized for intensifying a war whose horrors have rarely been equalled, apparently unnecessarily, in view of the imminent collapse of Spain as a military power. It was, in fact, the French struggle that revealed Spain's weaknesses, so that this is a judgment of hindsight. That Richelieu was drawn unwillingly by events into the vortex is clear, just as it is clear that the cost paid in social suffering and economic decline, leading to more frequent agrarian revolts, was high. Almost as soon as war broke out with Spain in 1635, Richelieu initiated secret peace negotiations and renewed them repeatedly while he lived. That he was motivated by the idea of extending France to its natural frontiers, the Rhine and the Alps, is unsupported by the evidence, although he took advantage of the predicament of its rulers to extend French protection into Alsace.

Richelieu's justification for war was the same as that for rigorous domestic discipline: only the statesman, furnished with all available information and equipped for judicious appraisal of events, is competent to judge the justice or otherwise of policy. Hence, "the subjects must blindly obey the prince; for often necessity constrains [his ministers] to adopt policies which cannot be supported by abstract reason alone, but can be justified only in the event."

Economic policies. In economic matters Richelieu was an amateur. He committed war expenditure with little regard for the difficulties of raising revenue, and he was given to economic improvisation that was often unsound. But he eschewed doctrinaire views and retained flexibility of mind. Whereas he was early influenced by the theories of the economist Antoine de Montchrestien, who thought of specie as the finite index of national wealth and who argued for economic self-sufficiency so as to conserve it, he was later persuaded that the drain of specie could be compensated for by trade. He accordingly promoted products and industries that could give France an export advantage and discouraged, mostly ineffectively, imports of luxury goods. Glass making, tapestry and silk, sugar, and the extractive industries attracted his interest. He planned canal systems and promoted overseas trading companies, in which he was a shareholder and which began the process of French colonization in Canada and the West Indies; and he gained economic and political footholds in Morocco and Persia.

His vast horizon reflected in part his concern with the French religious missions, which spread in every direction in Africa, the Middle East, and America and which extended French influence and created a vast intelligence network that fostered his political and economic designs. He laid the foundations for the French Navy by buying ships from the Dutch; and though he failed to have much influence on seapower, he developed shipping connections with the Baltic on which the French shipbuilding programs of his successors were to be founded. The legal reforms of his period were spasmodic and often frustrated by the Parlement, and how much of their content is due to him is questionable. The Code Michaud of 1629—which regulated primary industry and trade, companies, public offices, the church, and the army and also standardized weights and measures—was promulgated under his authority, although he may not have been its architect.

Later years in the church. In his last years Richelieu found himself involved in religious conflict, in opposition to the Pope, and in a struggle with the French church over the allocation of church revenues to the financing of the war. His relationship with Urban VIII became strained over diplomatic grievances, questions of church administration, and his own ambitions to extend French political influence by acquiring benefices for himself in the Holy Roman Empire. In spite of these conflicts, Richelieu remained strictly orthodox in his views on the relationship between church and state and thus resisted the Gallican challenge to the absolutism of papal authori-

Notion of papal authority

ty. He explained his position as follows: "While religiously obeying the Pope in spiritual matters, one can justly oppose him in his temporal designs." At the same time, his theory that "the King receives his crown and his temporal power from God alone" inevitably affected his view of the boundary between spiritual and temporal matters.

The theocratic concept of the state that resulted from this notion of kingship caused Richelieu to regard heresy as political dissidence, and he harried the apparently unorthodox, such as the first Jansenists, on the ground that they disturbed at once the spiritual and secular orders, just as he harried the recalcitrant nobles and stamped out duelling. Although there were canonical irregularities in his life, notably in the matter of pluralism, (the multiplication of ecclesiastical benefices), there is no evidence of a serious departure from the principles or practices of the church, and he maintained relationships with such religious figures as St. Vincent de Paul and St. John Eudes in their charitable activities. His accumulation of wealth was excessive even by the standards of the age, but it was largely dedicated to the public service and to patronage of the arts and of the University of Paris. Richelieu himself was a playwright and musician of moderate talent, and his establishment of the French Academy is one of the achievements for which France best remembers him.

Foundation of French Academy

His last months were agitated by the most dangerous of all the conspiracies against his life, that of the youthful royal favourite Cinq-Mars, who was exposed by Richelieu's secret service and died on the block. The Cardinal's health, bad for some years, had deteriorated, and it was virtually from his deathbed that he was compelled to dictate to the King five propositions respecting royal behaviour toward ministers that he considered essential for proper government. He died in the Palais Royal, which he left in his will to the crown, on December 4, 1642, with the last rites and after nominating Cardinal Mazarin as his successor, proclaiming, so it was said, that he had never had any intention "other than the good of religion and the State" and no enemies but those of the state. He was buried in the chapel of the Sorbonne, which he had financed, and his tomb, despite desecration during the Revolution, still stands.

Assessment. While in detail he was only moderately successful, Richelieu in substance attained his goals of orderly government under the royal authority and the defeat of Habsburg hegemony. Whether the centrifugal forces in Germany that he promoted—and which the Peace of Westphalia institutionalized—were advantageous to Europe in the long run is questionable, but the political fragmentation of the empire and the military eclipse of Spain made possible the grandeur of France that Richelieu foresaw and his successors realized. This mystical aspect of his designs is difficult to articulate but is essential to his greatness. The conspiracies that erupted under Mazarin failed as much because Richelieu had wrought a fundamental psychological change in favour of the moral ascendency of the crown as because, by the destruction of castles and city walls and the centralization of military authority, he had eliminated the power base of both aristocratic and religious dissent.

Written works. Apart from the religious works referred to, Richelieu left the following manuscripts: His *Journal* of the events of the Day of Dupes; his *Mémoires*, covering from 1610 to 1638, whose authenticity is controversial; his ***Testament politique***, which purports to be a guide to statecraft directed to the King and which was published first in 1688.

BIBLIOGRAPHY

Writings: ***Mémoires***, extracts were first published by M. PETITOT, ***Collection complète des mémoires relatifs à l'histoire de France, depuis le règne de Philippe Auguste,*** 130 vol. (1819–29). The period from 1610 to 1638 is covered in the 2nd series, vol. 21–30. The Société de l'Histoire de France published the text from the original ms. in 10 vol., ed. by BARON DE COURCEL and J. LAIR (1907–31). The same period is covered in a 4-vol. folio edition (1961). ***Testament Politique d'Armand du Plessis, Cardinal duc de Richelieu,*** 3rd ed. (1688). The standard edition is the ***ddition critique publike avec une introduction et des notes par Louis Andre' et une pre'face de Le'on Noel*** (1947). ***Journal de M. le cardinal duc de Richelieu qu'il a faict durant le grand orage de la court en l'année 1630 et 1631*** (1648; republished in ***Archives curieuses de l'histoire de France,*** 2nd series, vol. 5, 1838); ***Maximes d'État et fragments politiques,*** vol. 3 of ***Les documents inédits sur l'histoire de France,*** series 1 (1880); ***Les principaux poincts de la foy de l'Église catholique defendus contre l'escrit adressé au Roy par les quatre ministres de Charenton*** (1618; reprinted 1842; Eng. trans., ***The Principall Points of the Faith of the Catholike Church: Defended Against a Writing Sent to the King by 4 Ministers of Charenton,*** 1635); ***lnsfruction du chrestien*** (1621); ***Traitté de la perfection du chrestien*** (1646); ***Traittk qui contient la me'thode la plus facile et la plus asseurée pour convertir ceux qui se sont séparez de L'Église*** (1651).

Works about Richelieu: G. HANOTAUX and LE DUC DE LA FORCE, ***Histoire du cardinal de Richelieu,*** 6 vol. (1893–1947); D.P. O'CONNELL, ***Richelieu*** (1968); C.J. BURCKHARDT, ***Richelieu,*** 4 vol. (1935–67; Eng. trans, ***Richelieu and His Age: His Rise to Power,*** 2 vol., 1940–70); H.S.J. GRIFFET, ***Histoire de règne de Louis XIII,*** 3 vol. (1758), an important work that includes material that has since disappeared; D.L.M. AVENEL, ***Lettres, instructions diplomatiques et papiers d'État du cardinal de Richelieu,*** 8 vol. (1854–77), work that contains the most important of Richelieu's correspondence; G. FAGNIEZ, ***Le Père Joseph et Richelieu, 1577–1638,*** 2 vol. (1894).

(D.P.O'C.)

Richemont, Constablede

Arthur, comte de Richemont, was the third and last of the great Breton constables who fought for the French House of Valois during the Hundred Years' War between France and England. France's implacable struggle against England for independence furnished the backdrop for Richemont's talents as diplomat, politician, and, most importantly, as a military organizer. Richemont emerged in the late 1430s and early 1440s as the only figure in the court circle around the Valois king Charles VII who was capable of reconstructing the French Army into a loyal, modern, and efficient military force. In his capacity as chief of military operations, Richemont was responsible for the final French victories of 1449–53, which cleared his country of the English armies and were a direct preliminary to the emergence of France as the most powerful military and political force in late-15th-century Europe.

Early career and capture by English

Born in 1393 as a younger son of John IV, duke of Brittany, Arthur was given the English title of earl of Richmond (in French, comte de Richemont) by his older brother, Duke John V, in 1399. The marriage of their mother, Joanna, to Henry IV of England after her first husband's death, had re-established Brittany's connection with the English crown, but Richemont's primary interests remained in French affairs. In the bitter and divisive feud between the houses of Orléans and Burgundy—branches of the Valois dynasty—Richemont fought on the side of the former faction, shortly to be renamed Armagnac. During this same period, Arthur also became the intimate friend and partisan of the dauphin Louis, son of the French king Charles VI.

Richemont fought at Agincourt in 1415, where he was wounded and captured by the English victors, who, allied with the Burgundians, sought to unite France and England under the English crown. Richemont remained a prisoner in England until 1420, when he was released on parole and threw his support to the English side. He was now influential in persuading his brother, John, the duke of Brittany, to support the Treaty of Troyes under which Henry V of England became "Heir of France." Henry rewarded Richemont with the French county of Ivry. Richemont's connection with the Anglo-Burgundian faction was further sealed in 1423 by his marriage to Margaret of Burgundy, widow of the dauphin Louis, who had died young. This match made Richemont the brother-in-law of Philip, the duke of Burgundy, and John, the duke of Bedford, the English regent of France. Richemont was well on his way toward a high position in the ruling circles around Bedford and Burgundy when an unexplained quarrel broke out between him and the English regent. Richemont now deserted the English cause and

Return to French allegiance

returned to his initial French allegiance. Appointed constable of France by Charles VII in March 1425, he attempted to assume control of France's battered and unreliable military forces. He now totally supported the French cause, persuading his brother, John V of Brittany, to sign the Treaty of Saumur with France in October 1425.

The new constable quickly made himself unpopuiar by his rough manners and his grim insistence upon a vigorous prosecution of the war. His political power was therefore overshadowed by that of Charles VII's incompetent favourites, especially Georges de La Trtmoille. Richemont's influence at court was further weakened by Brittany's retum to the English cause. A treaty between John V and the regent Bedford in September 1427 caused the expulsion of the Constable from the French court. Richemont joined Joan of Arc at Orleans in 1429, fighting under her banner in several victorious engagements against the English until the influence of La Trtmoille forced him out of the army once again. Despite the favourite's power, Richemont was able to bring Brittany and Charles VII together once again in the Treaty of Rennes, but it was not until La Trémoille's final overthrow in 1432 that the Constable was able to retum to court.

Using his Burgundian connections, Richemont was able to arrange the Treaty of Arras (September 21, 1435), which ended the long quarrel between Duke Philip of Burgundy and the French king. Arras was the political and diplomatic turning point of the Hundred Years' War, as well as an important milestone in Richemont's own career.

But the military task of winning the war still remained. In April 1436 Richemont marched into Paris as the city rose against the English garrison, but the poorly organized French armies were unable to make much headway in the years that followed. Richemont now determined upon a total reform of the French Army, along with a reorganization of the financial structure of the French state in order to provide the revenues necessary for its support. Strongly supported by Charles VII and given a steady source of revenue by the taxes upon hearths and salt, Richemont now reorganized the French cavalry into the regular and highly professional ***gens d'armes d'ordonnance.*** These regular companies enabled Richemont to renew the war with overwhelming success after a brief truce had been concluded in 1444. In this final act of the long struggle, the Constable de Richemont played an active role, driving the English from the Cotentin Peninsula in September and October of 1449 and taking a decisive part in the climactic Battle of Formigny in April 1450. The conquest of Normandy followed in short order and that of Guyenne in the succeeding two years.

Reform of army and expulsion of English

France had finally won the Hundred Years' War, and Richemont's active career now drew to a close. Succeeding his nephew Peter II, he became duke of Brittany in September 1457. He died on December 26, 1458, leaving no legitimate children.

BIBLIOGRAPHY. The main, eyewitness source for Richemont's life is that written by his friend and companion, GUILLAUME GRUEL, *Chronique d'Arthur **de Richemont*** (1890). ***See*** also EUGENE COSNEAU's somewhat dated biography, ***Le Connétable de Richemont*** (1886); G. DU FRESNE DE BEAUCOURT's old but reliable ***Histoire de Charles VIZ, 6*** vol. (1881–91); and E. PERROY, ***La Guerre de cent ans*** (1945; Eng. trans., ***The Hundred Years War,*** 1951). No definitive modern work on Richemont presently exists.

(R.Br.)

Riding and Horsemanship

Before the advent of mechanized vehicles, riding on horseback was one of the chief means of transportation. As the horse never chose to be ridden, man of necessity evolved horsemanship, which is the art of riding with maximum discernment and a minimum of interference with the horse. Until recent years riding was a monopoly of the cavalry, cowboys, and others whose work required riding on horseback and of the wealthy, who rode for sport. A rapid change has taken place, however. Although hunting and polo tend to remain the sport of the wealthy and the role of the horse in battle has ended, especial value is now placed on horse shows of a high standard, in which the most popular event is undoubtedly show jumping. The finer aspects of horsemanship have remained a valued social asset and symbol of prestige, but the opening of many new riding clubs and stables has made riding and horsemanship accessible to a much larger segment of the population.

This article will treat the history of riding from the earliest known times to the development of horsemanship as an organized sport. The main emphasis here will be on modern equestrian competition, including the main events of both dressage and jumping, and on the art of horsemanship.

HISTORY

Origins **and** early history. From the 2nd millennium BC, and probably even earlier, the horse was employed as a riding animal by fierce nomadic peoples of central Asia. One of these peoples, the Scythians, were accomplished horsemen and used saddles. It is also likely that they realized the importance of a firm seat and were the first to devise a form of stirrup. A saddled horse with straps hanging at the side and looped at the lower end is portrayed on a vase of the 4th century BC found at Chertomlyk in the Soviet Union. This contrivance may have been used for mounting only, however, because of the danger of being unable to free the foot quickly in dismounting. The Greek historian Strabo says that the indocility of the Scythians' wild horses made gelding necessary, a practice until then unknown in the ancient world. The Sarmatians, superb horsemen who superseded the Scythians, rode bareback, controlling their horses with knee pressure and distribution of the rider's weight.

Among the earliest peoples to fight and hunt on horseback were the Hittites, the Assyrians, and the Babylonians; at the same time (about 1500 BC) the Hyksos, or Shepherd Kings, introduced horses into Egypt and rode them in all their wars. In the 8th and 7th centuries BC, the Scythians brought horses to Greece where the art of riding developed rapidly, at first only for pleasure. A frieze from the Parthenon in Athens shows Greeks riding bareback. Philip II of Macedon had a body of cavalry in his army, and his son Alexander's army had separate, organized horse units. In the 4th century BC another Greek historian, Xenophon, wrote his treatise ***Hippikē*** giving excellent advice on horsemanship. Many of his principles are still perfectly valid. He advocated the use of the mildest possible bits and disapproved of the use of force in training and in riding. The Roman mounted troops were normally barbarian archers who rode without stirrups and apparently without reins, leaving the hands free to use the bow and arrow.

Classical horsemanship

Early riding equipment. As a general rule almost every item of riding equipment originated among the horsemen of the Eurasian steppes and was adopted by the people of the lands they overran to the east, the south, and later the west.

Horseshoes of various types were used by migratory Eurasian tribes about the 2nd century BC, but the nailed iron horsehoe as used today first appeared in Europe about the 5th century AD, introduced by invaders from the East. One such, complete with nails, was found in the tomb of the Frankish king Childeric I (died 481/482) at Tournai, Belgium.

Attila is said to have brought the stirrup to Europe. Round or triangular iron stirrups were used by the Avars in the 6th century AD, and metal stirrups were used by the Byzantine cavalry. They were in use in China and Japan about 600 AD.

The principle of controlling a horse by exerting pressure on its mouth through a bit and reins was practiced from the earliest times, and bits made of bone and antlers have been found dating from before 1000 BC. The flexible mouthpiece with two links and its variations have been in use down the centuries, leading directly to the jointed snaffle bit of the present day.

Early, stumpy prickspurs have been found in Bohemia on 4th-century BC Celtic sites.

Military horsemanship. The importance of cavalry increased in the early Middle Ages, and in the 1,000 years that followed, mounted warriors became predominant in battle. Armour steadily became bulkier and heavier, forcing the breeding of more and more massive horses, until the combination rendered manoeuvrability nearly impossible.

Efforts to overcome this were made at a Naples riding academy in the early 16th century, when Federico Grisone and Giovanni Battista Pignatelli tried to combine classical Greek principles with the requirements of medieval mounted combat. After Xenophon, except for a 14th century treatise by Ibn Hudhayl, an Arab of Granada, Spain, apparently no literature on riding was produced until Grisone published his ***Gli Ordini di Cavalcare*** in 1550.

The development of firearms led to the shedding of armour, making it possible for some further modifications in methods and training under followers of the school of Pignatelli and Grisone, such as William Cavendish, Duke of Newcastle. In 1733 François Robichon de la Guérinière published ***Ecole de Cavalerie*** in which he explained how a horse can be trained without being forced into submission, the fundamental precept of modern dressage. Dressage, the French word for schooling, is the methodical training of a horse for any of a wide range of purposes, excluding only racing and cross-country riding.

Growth of modern technique

Meanwhile, the Imperial Spanish Riding School of Vienna and the French cavalry centre at Saumur aimed at perfecting the combined performance of horse and rider. Their technique and academic seat, a formal riding position or style in which the rider sits deep in the middle of the saddle holding his body erect, exerted considerable influence in Europe and America during the 18th and 19th centuries and are still used in modern dressage. The head riding master at Saumur, Comte Antoine d'Aure, however, promoted a bold, relaxed, and more natural, if less "correct," style of riding across country, in disagreement with his 19th-century contemporary, François Baucher, a horseman of great ability with formal *haute école* ("high school") ideas. Classical exercises in the manège, or school for riding, had to make way for simplified and more rational riding in war and the hunt. During this period hunters jumped obstacles with their feet forward, their torso back on the horse's haunches, and its head held up. The horse often leaped in terror.

At the turn of the 20th century, Captain Federico Caprilli, an Italian cavalry instructor, made a thorough study of the psychology and mechanics of lowmotion of the horse. He completely revolutionized the established system by innovating the forward seat, a position and style of riding in which the rider's weight is centred forward in the saddle, over the horse's withers. About the same time the American flat-racing jockey Tod Sloan had astounding success with his "monkey seat," an extreme forward riding position and the racing seat of today. Caprilli wrote very little, but his pupil, Piero Santini, popularized his master's fundamental principles. Except in dressage and showing, the forward seat is the one now most frequently used, especially for jumping.

THE ART OF HORSEMANSHIP

The basic principle of horsemanship is to obtain results in a humane way by a combination of balance, seat, hands, and legs.

Fundamentals. The horse's natural centre of gravity shifts with its every movement and change of gait. Considering that a mounted horse also carries a comparatively unstable burden approximately one-fifth of its own weight, it is up to the rider to conform with the movements of the horse as much as possible.

Before mounting, the horseman sees that the saddle fits him and the horse. In the saddle his position is such that he can stay on the horse and control it. The seat he adopts depends on the particular task at hand. A secure seat is essential, giving the rider complete independence and freedom to apply effectively the aids at his disposal. He does not overrule the horse, but, firmly and without inflicting pain, he persuades it to submit to his wishes.

The horse's movements. The natural gaits of the horse are the walk, the trot, the canter or slow gallop, and the gallop, although in dressage the canter and gallop are not usually differentiated. A riding horse is trained in each gait and in the change from one to another.

During the walk and the gallop the horse's head moves down and forward, then up and back (only at the trot is it still), and the rider follows these movements with his hands.

Walk. The walk is a slow, four-beat, rhythmical pace of distinct successive hoof beats in an order such as near (left) hind, near fore, off (right) hind, off fore. Alternately two or three feet may be touching the ground simultaneously. It may be a free, or ordinary, walk in which relaxed extended action allows the horse freedom of its head and neck, but contact with the mouth is maintained; or it may be a collected walk, a short-striding gait full of impulsion, or vigour; or it may be an extended walk of long, unhurried strides.

Trot. The trot is a two-beat gait, light and balanced, the fore and hind diagonal pairs of legs following each other almost simultaneously—near fore, off hind, off fore, and near hind. The rider can either sit in the saddle and be bumped as the horse springs from one diagonal to the other or he can rise to the trot, also termed posting, rising out of the saddle slightly and allowing more of his weight to bear on the stirrups when one or the other of the diagonal pairs of legs leaves the ground. Posting reduces the impact of the trot on both horse and rider.

Canter. As the horse moves faster, its gait changes into the canter, or ordinary gallop, in which the rider does not rise or bump. It is a three-beat gait, graceful and elegant, characterized by one or the other of the forelegs and both hindlegs leading—near hind, off hind, and near fore practically together, then off fore, followed briefly by complete suspension. Cantering can be on the near lead or the off, depending on which is the last foot to leave the ground. The rider's body is more forward than at the trot, his weight taken by the stirrups.

Gallop. An accelerated canter becomes the gallop, in which the rider's weight is brought sharply forward as the horse reaches speeds up to 30 miles (48 kilometres) an hour. The horse's movements are the same as in the canter.

Other gaits. There are a number of disconnected and intermediate gaits, some of which are done only by horses bred especially to perform them. One is the rack, a four-beat gait, with each beat evenly spaced in perfect cadence and rapid succession. The legs on either side move together, with the hindleg striking the ground slightly before the foreleg. The single foot is similar to the rack but slower. In the pace, the legs on either side move and strike the ground together, resulting in a two-beat gait. The fox trot and the amble are both four-beat gaits, the latter smoother and gliding.

The rack gait

Training. Depending on the abilities and inclinations of horse and trainer, training may include such elements as collection (controlled, precise, elevated style of movement) and extension (smooth, swift, reaching movement —the opposite of collection) at all paces; turns on the forehand (that part of the horse that is in front of the rider) and hindquarters; changing lead leg at the canter; change of speed; reining back, or moving backwards; lateral movements; and finally the refinements of dressage, jumping, and cross-country riding.

Communication with the horse is rendered possible by the use of the bit and the aids. The rider signals his intentions to the horse by a combination of recognized movements of his hands and legs, using several articles of equipment. By repetition the horse remembers this language, understands what is required, and obeys.

Bits. There are several types of bits, the most commonly used being the snaffle, the double bridle, and, to a lesser extent, the Pelham.

The simplest is the snaffle, also called the bridoon. It consists of a single straight or jointed mouthpiece with a ring at each end for the reins. The snaffle is used for racing and frequently for riding across country. It is appropriate for preliminary schooling, and, until the horse has accepted it and is thoroughly confident, a double bridle should not be used.

The double bridle

The double bridle is used for advanced schooling. It consists of a jointed snaffle and a straight bit placed together in the mouth, first the snaffle, then the bit, both functioning independently and attached to separate reins. The mouthpiece of the bit can have a port or indentation in its centre to give more control. The slightest pull on the bit rein exerts pressure on the mouth.

The Pelham is a snaffle with a straight mouthpiece; cheekpieces with rings at the lower ends for curb action; and a curb chain, with which pressure may be applied to the lower outside of the horse's mouth. The Pelham gives ample control with only slight discomfort and is popular for polo.

Aids. The principal features of a horse's mentality are acute powers of observation, innate timidity, and a good memory. To a certain extent it can also understand. Schooling is based on these faculties, and the rider's aids are applied accordingly. The natural aids are the voice, the hands through the reins and the bit, the legs and heels, and movement of the rider's weight. The whip, the spur, and devices such as martingales, special nosebands, and reins are artificial aids, so termed in theory, as the horse does not discriminate between natural and artificial.

Horses are easily startled. A good horseman will approach them quietly, speaking to them and patting them to give them confidence. Silence on the part of the rider can even cause disquiet to some horses, but they should not be shouted at. The rider's voice and its tone make a useful aid in teaching a horse in its early schooling to walk, trot, canter, and halt.

To keep the horse alert at all times, the rider's hands keep a light, continual contact with its mouth, even at the halt. The hands are employed together with the legs to maintain contact, to urge the horse forward, to turn, to rein back, and generally to control the forehand. The horse is said to be collected and light in hand when the action of the bit can cause it to flex, or relax, its jaw with its head bent at the poll, or top.

When pressed simultaneously against the flanks, immediately after the hands ease the reins, the legs induce the forward movement of the horse. They are of the greatest importance in creating and maintaining impulsion, in controlling the hindquarters, and for lateral movement.

The rider achieves unity of balance by means of the weight aid, that is, by moving his body in harmony with the movements of the horse, forward, backward, or to a side. Thus, in cantering to the left, he leans to the left; or when about to descend a steep slope, he stays erect while the horse is feeling for the edge with its forefeet, but as soon as the descent starts he leans forward, leaving the hindquarters free to act as a brake and to prevent scraping the back of the horse's rear legs on rough ground. Meanwhile the hands keep the horse headed straight to maintain its balance.

The use of the whip

The whip is used chiefly to reinforce the leg aid for control, to command attention, and to demand obedience, but it can be used as a punishment in cases of deliberate rebellion. A horse may show resistance by gnashing its teeth and swishing its tail. Striking should always be on the quarters, behind the saddle girth, and must be immediate since a horse can associate only nearly simultaneous events. This applies equally to rewards. A friendly tone of voice or a pat on the neck are types of reward.

Although normally the leg or the heel, or both, should be sufficient, spurs, which should always be blunt, assist the legs in directing the precision movements of advanced schooling. Their use must be correctly timed.

Martingales are straps used to control a horse's head carriage and are usually one of three types: running, standing, or Irish.

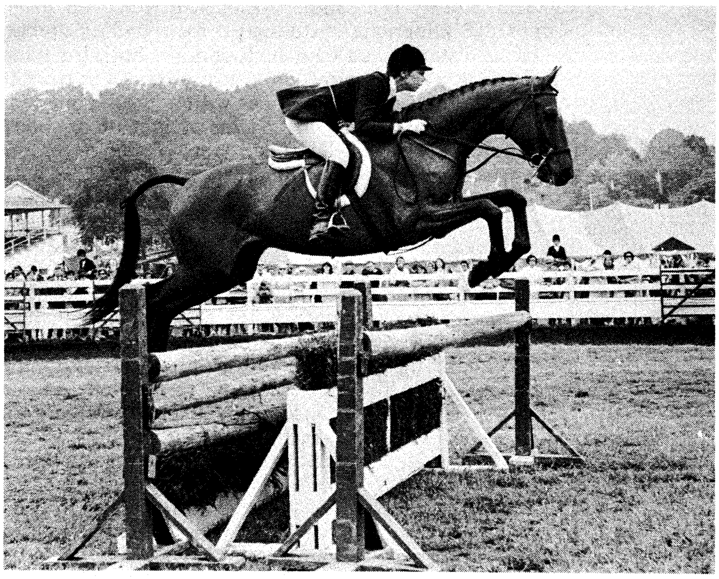
The running and standing martingales are attached to the saddle straps at one end and the bit reins or bridle at the other. The Irish martingale, a short strap below the horse's chin through which the reins pass, is used for racing and stops the horse from jerking the reins over its head. As the horse cannot see below a line from the eye to the nostril, it should not be allowed to toss its head back, particularly on approaching an obstacle, as it is liable to leap blindly. A martingale should not be necessary with a well-schooled horse, however.

The noseband, a strap of the bridle that encircles the horse's nose, may be either a cavesson, with a headpiece and rings for attaching a long training rein, or a noseband with a headstrap, only necessary if a standing martingale is used. A variety of other nosebands are intended for horses that pull, or bear, on the reins unnecessarily.

Seats. The saddle, the length of the stirrup, and the rider's seat, or style of riding, should suit the purpose for which the horse is ridden. The first use of the stirrup is to enable the rider to get on the horse, normally from the near (left) side. With his raised foot in the stirrup the rider should avoid digging the horse in the flank on springing up and should gradually slide into position without landing on the horse's kidneys with a bump. With an excitable horse, the rider may wait, resting on knees and stirrups, until the horse moves forward.

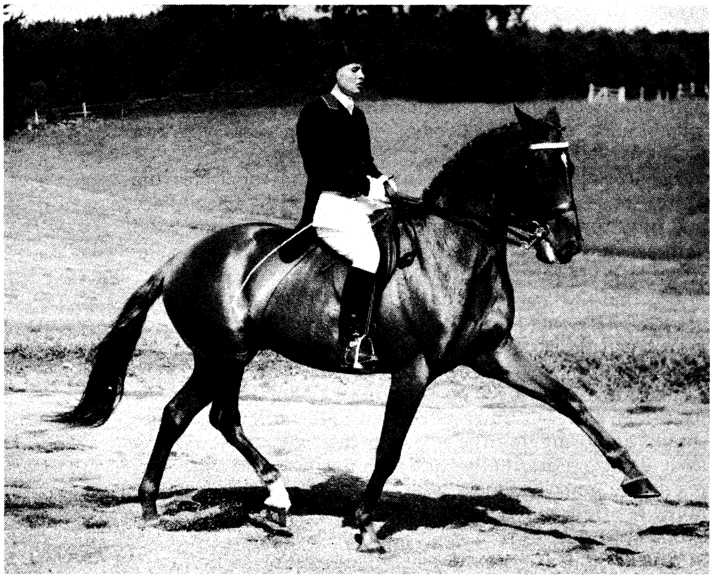
Forward. The forward seat, favoured for show jumping, hunting, and cross-country riding, is generally considered to conform with the natural action of the horse. The rider sits near the middle of the saddle, his torso a trifle forward, even at the halt. The saddle is shaped with the flaps forward, sometimes with knee rolls for added support in jumping. The length of the stirrup leather is such that, with continual lower thigh and knee grip, the arch of the foot can press on the tread of the iron with the heel well down. A wide and heavy stirrup iron allows easy release of the foot in case of accidents. The line along the forearm from the elbow to the hands and along the reins to the bit is held straight. As the horse moves forward, so does the rider with his hands, to suit the horse's comfort in all its movements.

Freudy Photos. N.Y.



Forward seat in jumping.

Dressage. In the show and dressage seat the rider sinks deep into the saddle, in a supple, relaxed but erect position above it. The saddle flaps are practically straight so as to show as much expanse of the horse's front as possible. The stirrup leather is of sufficient length for the rider's knee to bend at an angle of about 140 degrees and for the calf of his leg to make light contact with the horse's flank, the heel well down, and the toes or the ball of the foot resting on the tread of the stirrup iron. The rider keeps continual, light contact with the horse's mouth; and the intention is to convey an impression of graceful, collected action. In the past this type of saddle with its straight cut flaps was used for hunting and polo, for which the forward seat recently has become more popular.



Dressage seat—in the extended trot.
Freudy Photos, N.Y.

Stock saddle. The stock saddle seat is appropriate for ranchers but is also used at rodeos and by many pleasure and trail riders. The saddle, which can weigh up to 40 pounds (18 kilograms), is designed for rounding up cattle and is distinguished by a high pommel horn for tying a lariat. The rider employs long stirrups and a severe bit that he seldom uses since he rides with a loose rein, guiding his horse chiefly by shifting the weight of his body in the saddle. The gaucho roughriders of the Argentine Pampa have adopted a similar seat, using a saddle with a high pommel and cantle. Australian stockmen have used a saddle with a short flap, equipped with knee and thigh rolls, or props, which give an extremely secure seat.

Fiat racing. The flat-racing seat is a positively forward seat, with the jockey's knees on the withers and with very short stirrups. The rider stays on mainly by balancing himself above the point where the horse can carry the most weight, the flattened position of his body at the gallop offering less resistance to the air. The steeplechase seat is similar to the flat-racing seat, with slightly longer stirrups, and many riders now adopt a forward seat at the jumps.

Side saddle. Though now not so fashionable, the elegant and classical side-saddle seat is still favoured and considered correct by many horsewomen. On the near side the saddle has an upright pommel on which the rider's right leg rests. There is a lower, or leaping, pommel, against which the left leg can push upward when grip is required, and a single stirrup. Although the rider sits with both legs on one side of the saddle, forward action to suit the movement of the horse is feasible across country.

Bareback. Bareback means riding without saddle or blanket, the rider sitting in the hollow of the horse's back and staying there chiefly by balance. It is an uncomfortable seat but less so at the walk and the slow canter. When suffering from saddle galls horses are sometimes ridden bareback for exercise.

Dressage. Originally intended for military use, dressage training was begun early in the 16th century. Until recently manuals on the subject were written chiefly by cavalrymen.

The international rules for dressage are based on the traditions and practice of the best riding schools in the world. The following is an extract from these rules of the Fédération Équestre Internationale:

Object and general principles.

> The object of dressage is the harmonious development of the physique and ability of the horse. As a result, it makes the horse calm, supple and keen, thus achieving perfect understanding with its rider. These qualities are revealed by the freedom and regularity of the paces; the harmony, lightness, and ease of the movements; the lightening of the forehand, and the engagement of the hindquarters; the horse remaining absolutely straight in any movement along a straight line, and bending accordingly when moving on curved lines.
>
> The horse thus gives the impression of doing of his own account what is required of him. Confident and attentive, he submits generously to the control of his rider (Used with permission of the publisher).

Campagne is the term used for elementary but thorough training, including work on the longeing rein. This long rein, also used for training young or difficult horses, is attached to a headpiece with a noseband called a cavesson. The horse is bitted and saddled and is schooled in circles to left and right at the end of the rein. It is an accessory to training from the saddle, which is always best. Basic to *campagne* is collection, teaching the horse to arch its neck, shift its weight backward onto its hindquarters, and move in a showy, animated manner. Other important elements include riding in a straight line, turns, and lateral movements.

Haute école dressage

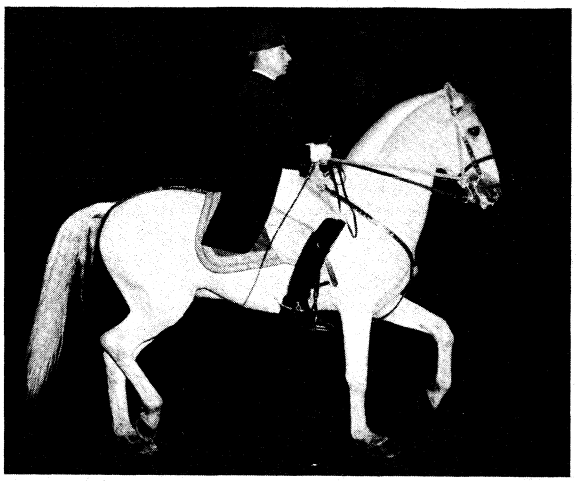
Haute école is the most elaborate and specialized form of dressage, reaching its ultimate development at the Vienna school in its traditional white Lippizaner horses. Some characteristic ***haute école*** airs, or movements, are the pirouettes, which are turns on the haunches at the walk and the canter; the piaffe, in which the horse trots without moving forward, backward, or sideways, the impulse being upward; the passage, high-stepping trot in which the impulse is more upward than forward; the levade, in which the horse stands balanced on its hindlegs, its forelegs drawn in; the courvet, which is a jump forward in the levade position; and the croupade, ballotade, and capriole, a variety of spectacular airs in which the horse jumps and lands again in the same spot.

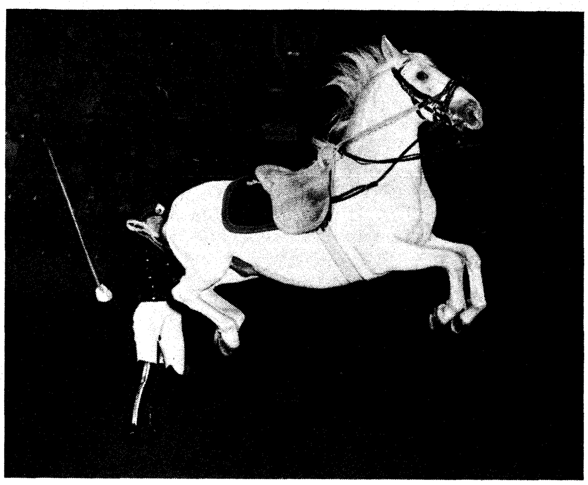
All of these movements are based, perhaps remotely in some instances, on those that the horse performs naturally.

Jumping. The most sensitive parts of the horse when ridden are the mouth and the loins, particularly in jumping. The rider's hands control the forehand while his legs act on the hindquarters. As speed is increased the seat is raised slightly from the saddle, with the back straight and the trunk and hands forward, the lower thighs and the knees taking the weight of the body and gripping the saddle, leaving the legs from the knees down free for impulsion. Contact with the mouth is maintained evenly and continually, the rider conforming with every movement of his mount as the horse's head goes forward after takeoff and as it is retracted on landing, the hands always moving in line with the horse's shoulder. In order to give complete freedom to the hindquarters and to the hocks, the rider does not sit back in the saddle until at least two strides after landing.

The horse is a natural jumper, but if ridden, schooling becomes necessary. Training is started in an enclosed level area by walking the horse, preferably in a snaffle, over a number of wooden bars or poles laid fiat on the ground. When the horse has become accustomed to this, its speed is increased. As the horse progresses, the series of obstacles is systematically raised, varied, and spaced irregularly. The object is to teach the horse: (1) to keep its head down; (2) to approach an obstacle at a quiet, collected, yet energetic pace; (3) to decide how and where to take off; and (4) after landing to proceed quietly to the next obstacle. The horse should be thoroughly confident over every jump before it is raised and should be familiarized with a variety of obstacles.

Only thoroughly trained riders and horses compete. Very strenuous effort is required of the horse, as well as of the rider, who does not by any of his actions give his mount the impression that something out of the ordinary is impending. If possible the horse is warmed up by at least a half-hour's walking and trotting before entering the ring. The horse is guided toward the exact centre of every obstacle, the rider looking straight ahead and not looking around after takeoff for any reason, as that might unbalance the horse. The broader the obstacle, the greater the speed of approach. Although a few experienced riders can adjust the horse's stride for a correct takeoff, this should not be necessary with a well-schooled horse.





Haute école figures: Lippizaner horses in (left) piaffe and (right) ballotade.
Freudy Photos

The rider is always made to conform with every action of the horse, the only assistance necessary being that of direction and increasing or decreasing speed according to the obstacle.

RIDING AND SHOWS

Racing on horseback probably originated soon after man first mastered the horse. By the 7th century BC organized mounted games were held at Olympia. The Romans held race meetings, and in medieval Europe tournaments, jousting, and horse fairs were frequent and popular events. Played in Persia for centuries, polo was brought to England from India about 1870. In North America, Western ranch riding produced the rodeo.

Associa-
tions
and clubs

Horse associations and pony clubs are today the mainstay of equine sport. They have improved the standards of riding instruction and the competitive activities of dressage, hunter trials, and show jumping. The latter has become an important event since 1869, when what was probably the first "competition for leaping horses" was included in the program of an Agricultural Hall Society horse show in London. National organizations such as the British Horse Society, the American Horse Shows Association (AHSA), the Federazione Italiana Sports Equestri, the National Equestrian Federation of Ireland, the Fédération Française des Sports Équestres, and similar groups from about 50 other nations are affiliated with the Fédération Équestre Internationale (FEI), founded in 1921 with headquarters at Brussels, the official international governing body and the authority on the requirements of equitation.

Horse shows. Horse shows are a popular institution that evolved from the horse sections of agricultural fairs. Originally they were informal displays intended to attract buyers and encourage the improvement of every type of horse. Now they are organized and conducted by committees of experts and by associations that enforce uniform rules, appoint judges, settle disputes, maintain records, and disseminate information. Riding contests included in the program, in which highly trained riders and horses compete in a number of events and classes, have become increasingly important.

Under the auspices of the Royal Dublin Society, an international horse show was first held at Dublin in 1864. It is an annual exhibition of every type of saddle horse, as well as broodmares and ponies. International jumping contests similar to Olympic competition, events for children, and auction sales are held during this five-day show.

The National Horse Show at New York, first held in 1883, is another great yearly event. Held at Madison Square Garden, it lasts several days and includes about 10 different events. Among the most important are the international jumping under FEI rules and the open jumping under AHSA rules. Other shows are held in many sections of the United States.

Horse and pony shows are held regularly in the United Kingdom, the most important being the Richmond Royal Horse Show, the Horse of the Year Show, and the Royal International Horse Show. The latter, an annual event first held in 1907, has flourished under royal patronage and includes international jumping, special items such as the visit of the Spanish Riding School with its Lippizaners in 1953, and a Supreme Riding Horse competition.

In Canada, the Royal Agricultural Winter Fair at Toronto, opened in 1922 and known in Canada as the "Royal," is a major event, and in Australia the Royal Agricultural Society organizes horse shows annually in every state. Further examples are the shows at Verona and at the Piazza di Siena in Rome; frequent horse shows in Belgium, France, Germany, and The Netherlands; the winter show in July in Buenos Aires; and the Exhibition of Economic Achievement in Moscow.

Olympic equestrian competition. The FEI organizes and controls the equestrian events at the Olympic Games. Included in each Olympics since the Games at Stockholm in 1912 (equestrian events were also held in 1900), these events are the occasion for keen rivalry and evoke high standards of horsemanship. They comprise a dressage grand prix, a three-day event, and a jumping grand prix, all open to team and individual competition.

The Grand Prix de Dressage involves performance of the walk, trot, canter, and collected paces and several conventional dressage figures and movements, as well as the correct rider's position. Scoring on each item is from a maximum of 10 for excellent down to 1 for very bad. A score of 0 means the rider has performed nothing that was required.

The Three-Day Event consists of tests in dressage, endurance or cross-country riding, and show jumping. Dressage is on the first day. On the second day there is an endurance test over a course 25 to 35 kilometres (16 to 22 miles) in length, covering swamp roads, tracks, steeplechase obstacles, and across country. Jumping tests, less strenuous than the Prix des Nations jumping event, are held on the third day.

The Prix des Nations jumping event is a competition involving 13 or 14 obstacles, heights varying between 1.30 and 1.60 metres (51 and 63 inches), and a water jump 4 metres (13 feet) across, over a course with 60 metres (200 feet) between obstacles. Penalties are scored for disobedience, knocking down or touching an obstacle, and for a fall. The rider with the lowest penalty score wins.

In addition to these competitions there is a riding section of the modem pentathlon, also conducted under FEI rules. Competitors must clear, riding a strange horse chosen by lot, 20 obstacles over a course of 1,000 metres (3,000 feet) (for lists of Olympic Games equestrian champions see ATHLETIC GAMES AND CONTESTS: The *Olympic* record).

BIBLIOGRAPHY

General works: R.S. SUMMERHAYS, ***Encyclopaedia for Horsemen,*** rev. ed. (1962), a useful reference work; C.E.G.

HOPE, *Riding,* rev. ed. (1968); V.S. LITTAUER, *Common Sense Horsemanship,* 2nd ed. (1963); H. WYNMALEN, *Equitation,* 2nd ed. (1971); E.R. FARSCHLER, *Riding and Training,* new ed. (1959, reissued 1972), contains a description of the gaits; J.S.-F. PAILLARD, *Understanding Equitation* (1974); C.E.G. HOPE, *The Horseman's Manual* (1972); J. KIDD, *Horsemanship in Europe* (1977).

History: V.S. LITTAUER, *The Development of Modern Riding,* ed. by EV. CONNETT (1962); C. CHENEVIX TRENCH, *A History of Horsemanship* (1970); G.R. VERNAM, *Man on Horseback* (1965), includes information on the origin and detail of equipment.

Horse shows: RS. SUMMERHAYS, *The Story of the International,* 1907–1957 (1957), and with C.E.G. HOPE, *Horse Shows: The Judges, Stewards, Organizers* (1969); *American Horse Shows Association Rule Book* (annual).

Rules: The rules for international competitions are given in publications of the FÉDÉRATION ÉQUESTRE INTERNATIONALE; in B. PHILLIPS (ed.), *Official Report of the Olympic Games* (1968); and in various publications of the British Horse Society.

Dressage: R.L. WATJEN, *Dressage Riding: A Guide for the Training of Horse and Rider,* trans. by V. SALOSCHIN (1958); BRITISH HORSE SOCIETY, "Notes on Dressage" (n.d.).

Jumping: L. GIANOLI, *Lo sport del cavallo* (1946), for a description of the change to the forward seat; FEDERICO CAPRILLI, *The Caprilli Papers: Principles of Outdoor Equitation,* trans, and ed. by P. SANTINI (1967).

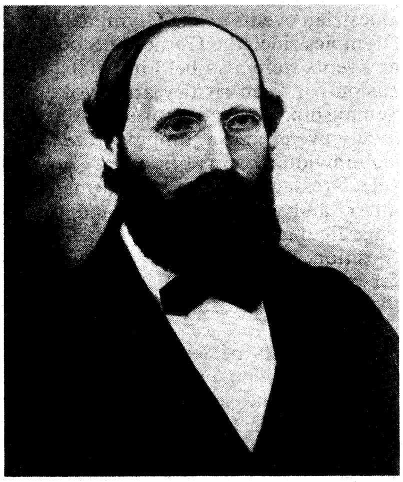
(C.E.C.)

Riemann, Bernhard

Georg Friedrich Bernhard Riemann was one of the most creative mathematicians of the 19th century. His relatively few published papers widely influenced geometry and analysis. Moreover, his bold ideas concerning the geometry of space had a profound effect on the development of modern theoretical physics. To a large extent, they provided the foundation for the concepts and methods used by Albert Einstein to develop his theory of relativity in the 20th century.

Born on September 17, 1826, in the village of Breselenz in Hannover, Germany, Riemann was the second of six children of a Lutheran pastor, who gave him his first instruction. He obtained a good education'with the encouragement of a happy and devout family. At the local Gymnasium (high school), he quickly progressed in mathematics beyond the guidance of his teachers, mastering calculus and the *Théorie des nombres* ("Theory of Numbers") of Adrien-Marie Legendre. In 1846–51 he studied at the universities of Gottingen and Berlin, where he was interested in problems concerning the theory of prime numbers, elliptic functions, and geometry. Following studies in experimental physics and Naturphilosophie, which sought to derive universal principles from all natural phenomena, he concluded that mathematical theory could secure a connection between magnetism, light, gravitation, and electricity; Riemann then suggested field theories, in which the space surrounding electrical charges may be mathematically described. Thus, during his student days he had already begun to develop some of the original ideas that later were to become important to modern mathematical physics.

Archiv fur Kunst und Geschichte



Riernann, lithograph after a portrait by an unknown artist, 1863.

Development of function theory

In 1851 he obtained the doctorate at Gottingen with a dissertation on the "Grundlagen fur eine allgemeine Theorie der Functionen einer veranderlichen complexen Grosse" ("Foundations for a General Theory of Functions of a Complex Variable"). Function theory, which treats the relations between varying complex numbers, is one of the major achievements of 19th-century mathematics. Riemann based his treatment on geometrical ideas rather than algebraic calculation alone. His work, which earned the rare praise of the renowned mathematician Carl Friedrich Gauss, led to the idea of the Reimann surface—a multilayered surface—on which a multivalued function of a complex variable can be interpreted as a single-valued function. This idea, in turn, contributed to methods in topology, which deals with position and place instead of measure and quantity. His probationary essay (Habilitationsschrift) for admission to the faculty in 1853 was "On the Represention of a Function by Means of a Tngonometrical Series."

Riemann's non-Euclidean geometry

While continuing to develop unifying mathematical themes in the laws of physics, Riemann also prepared in 1854 for his inaugural lecture at Gottingen, required for admission to the faculty as a *Privatdozent*, an unpaid lecturer dependent entirely on student fees. He listed three topics, from which Gauss, representing the faculty, chose "Über die Hypothesen, welche der Geometrie zu Grunde liegen" ("On the Hypotheses Which Form the Foundations of Geometry"). Gauss himself had devoted long, profound speculations to this difficult subject. In this lecture, one of the most celebrated in the history of mathematics, Riemann developed a comprehensive view of geometry. With a thorough understanding of the limitations of ordinary, Euclidean geometry, which is based on the postulate of parallels, he independently formulated a non-Euclidean geometry. In so doing, he was apparently unaware that Nikolay Lobachevsky and János Bolyai had already shown the possibility of devising a consistent geometry without this postulate. Riemann's non-Euclidean geometry was an alternative to theirs and to that formulated by Gauss. He postulated that, through a point outside a line, there are no parallels to that line, a physical example of which can be seen in the fact that two ships on a meridian must meet at a pole. He correctly perceived that his ideas would benefit physics, as indeed they did when Einstein later drew upon them to build his model of space–time in relativity theory (see PHYSICAL THEORIES, MATHEMATICAL ASPECTS OF: Relativity theory).

Beginning in 1855, Riemann received a small stipend that represented unusual academic progress at the time and removed him from the ranks of the hardship cases. In 1857 he became professor extraordinarius (associate professor) and in 1859 professor, succeeding the mathematician Peter Gustav Lejeune Dirichlet, who had succeeded Gauss four years earlier. Riemann was beset by overwork, deaths in his family, and his own faltering health. He continued, however, to produce original papers, which, though few in number—some were published posthumously—contained many rich ideas, such as his work on partial differential equations. A measure of his influence is the extensive list of methods, theorems, and concepts that bear his name: the Riemann approach to function theory, the Riemann–Roch theorem on algebraic functions, Riemann surfaces, the Riemann mapping theorem, the Riemann integral, the Riemann–Lebesgue lemma on trigonometrical integrals, the Riemann method in the theory of trigonometrical series, Riemannian geometry, Riemann curvature, Riemann matrices in the theory of Abelian functions, the Riemann zeta functions, the Riemann hypothesis, the Riemann method of solving hyperbolic partial differential equations, and Riemann–Liouville integrals of fractional order. In 1859 he wrote the paper "Über die Anzahle der Primzahlen unter einer gegebenen Grosse" ("On the Number of Primes in a Given Magnitude"), in which he partially described the asymptotic frequency of primes (positive integral numbers that have no other factors except one and themselves, as 2, 3, 5, . . .).

Riemann's growing reputation finally earned him a permanent post in 1859 at Gottingen as the second successor to Gauss. In 1862 he married Elise Koch, and, for a time, the conditions of his life improved. Then he fell ill with pleurisy, which was complicated by tuberculosis. His strength gradually ebbed, despite several visits to Italy for recuperation. He died, in the Lutheran faith of his childhood, on July 20, 1866, at Selasca, in northern Italy.

BIBLIOGRAPHY. Riemann's collected works (all in German), ed. by HEINRICH WEBER and RICHARD DEDEKIND, were published in one volume in 1876, including an extensive biography by Dedekind. A second edition was published in 1892, and a supplement, containing a large amount of previously unpublished material, appeared in 1902. A French translation of selections from the collected works appeared in 1898; although Dedekind's biography was among the items omitted, this volume did contain a French translation of a long address by FELIX KLEIN given in Vienna in 1894 on Riemann's influence on modern mathematics. The collected works were published in the United States in 1953; in addition to the 1892 German edition and the 1902 supplement, it contained an excellent article by HANS LEVY.

Rift Valleys

A rift valley is formed by subsidence between parallel, opposing scarps or cliffs produced by faulting (displacement of the Earth's crust). A rift valley, or simply rift, is therefore a tectonic feature, attributable to Earth movements, in contrast with river or glacier valleys, which are produced by erosional processes. The floors of rift valleys were initially contiguous with their bordering plateau highlands prior to faulting, although large-scale crustal separation, volcanism, and valley filling by sedimentation can complicate this simple picture.

The neologism the Great Rift Valley was applied in 1893 to the extensive system of tectonic valleys that traverses eastern Africa and the Middle East. The unity of this system, between Mozambique and the Dead Sea, had been recognized since the exploration of the chain of lakes between latitudes 15" S and 8° N in Africa in the late 1800s. The similarity of the valleys of eastern Africa to the Rhine Graben in Germany, which had been intensively studied, was particularly noted. The German term graben (trench or ditch) thus came to be synonymous with rift valley, though there is now a growing tendency to use "rift valley" for the gross structure and graben for smaller, faulted valleys that may occur within or diverge from the main rift valleys.

Continental and oceanic rifts

Rift valleys are sparsely and irregularly distributed in the continents of the world. The African–Middle East rifts are unique in magnitude and extent, but similar structures form the Rhine rift valleys (with possible extensions into France and Scandinavia), the Baikal Rift Valley of central Siberia, and the Imperial Valley of Southern California.

The post-World War II discovery of a very extensive system of faulted valleys on the ocean floors, in particular along the crests of large ridges running the length of the Atlantic and Indian oceans, has given rise to the term oceanic (or midoceanic) rifts. Though morphologically similar to continental rift valleys, oceanic rifts are formed by different processes: crustal separation and upwelling of molten material from beneath the Earth's crust, followed by a poorly understood process of faulting and uplift of the rift margins.

Oceanic rifts are a fundamental feature of global tectonics (Earth deformation). They marked the precise boundaries between pairs of crustal plates that are drifting apart in response to sea-floor spreading. As the plates move apart, the suture along the rift floor is healed with an upwelling of basaltic magma, which solidifies and accretes on the trailing edges of the separating plates. The results of this phenomenon can be closely observed in Iceland, where the Mid-Atlantic Rift is exposed above sea level. The direct connection of the oceanic rift system with the continental rift system of Africa, through the northwestern Indian Ocean and Gulf of Aden, suggests that the continental system is similar but complicated by the presence of thick continental crust that overlies the active zones of the upper mantle. This similarity has not yet been clearly established to the satisfaction of all.

Influence of the African system on man

The nature of the East African Rift System, with which this article is primarily concerned, has had far-reaching effects on man and his society, Volcanic eruptions and lake deposits, associated with development of the African rifts, have buried and preserved evidence of mankind's origins and earliest evolution. From those earliest times, man has found the climatic, pastoral, and agricultural conditions of the high plateaus of eastern Africa congenial to his needs. In Ethiopia, perhaps more than anywhere else in the world, it is possible to identify a nation's consciousness, historical development, and actual survival with intimate physiographic and geological determinants. It has been said that the rift valley lies like a natural moat in front of Abyssinia, which it has helped to render one of the most independent of African countries.

From *Bulletin of the Geophysical Observatory*, vol. 3, *no.* 1 (1962); Haile Selassie I University

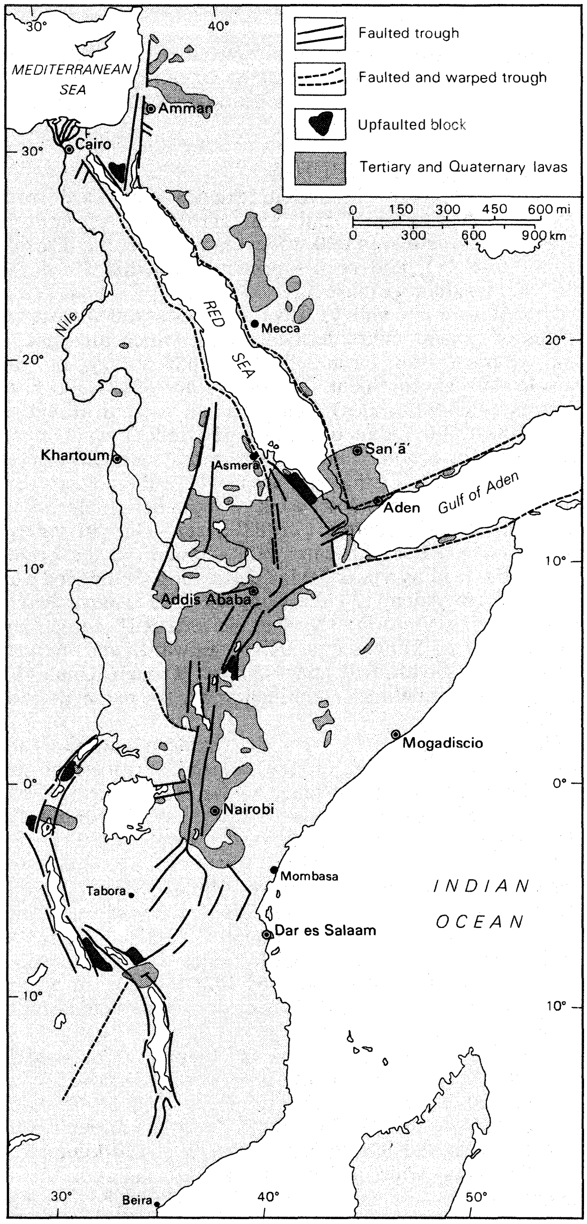
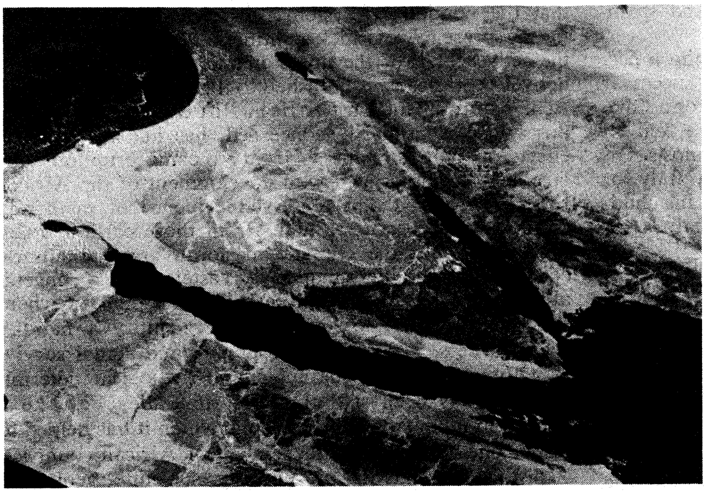


Figure 1: The Afro-Arabian rift system (see text).

This article treats the morphology, structure, geological history, and origin of rift valleys, principally by use of the most prominent terrestrial example, the East African Rift System. The nature and origin of rift systems is intimately associated with Earth history in general, and



Gemini 11 view of the Sinai Peninsula, showing the northern end of the East African Rift Belt extending through the Gulf of Suez (left) and the Gulf of Aqaba (right).
By courtesy of National Aeronautics and Space Administration

for further information the interested reader should see SEA-FLOOR SPREADING; CONTINENTAL DRIFT; MOUNTAIN-BUILDING PROCESSES; and ROCK MAGNETISM. See also PLATEAUS AND BASINS.

GEOMORPHIC CHARACTERISTICS

The Afro-Arabian rift system extends northward from Mozambique to the Dead Sea, across 50° of latitude (6,000 kilometres [4,000 miles], or 1/7 of the Earth's circumference), and occurs principally within 5° of the 35° E meridian (Figure 1).

The African rift valleys traverse broad, oval-shaped regions of crustal uplift, termed swells, which are part of the characteristic basin-and-swell morphology of the whole African continent. The swells are 600–1,500 kilometres (400–900 miles) long in a north–south direction and 300–1,000 kilometres (200–600 miles) wide in an east–west direction. The largest, most mountainous of these swells is the Ethiopian area, where the plateau surface that was once below sea level now lies at elevations as high as 3,300 metres (11,000 feet). The rift valleys that bisect the East African swells include (1) the Ethiopian rifts, leading via the Afar depression to the Red Sea and Gulf of Aden; (2) the Gregory Rift (Kenya Rift), east of Lake Victoria; (3) the Western Rift, containing lakes Rukwa, Tanganyika, Kivu, Edward, and Albert; and (4) the Nyasa Rift and Luangwa Trough. Units (1) and (2) are sometimes combined under the name Eastern Rift.

The swells associated with the Western and Gregory rifts are superimposed on the respective fringes of the broader East African plateau, whose shallow central sag is occupied by Lake Victoria (1,130 metres [3,706 feet]). The flat surfaces of the high swells have been dissected by erosion to form steep-sided canyons such as the 1,500 metre-deep (5,000-foot) Abay (Blue Nile) Canyon. They also have provided the base for several superimposed volcanic cones—Kilimanjaro (5,895 metres or 19,340 feet), Mt. Kenya (5,200 metres or 17,060 feet), and Ras Dejen (4,540 metres or 14,890 feet), for example. The rift valleys and their boundary scarps cut abruptly across this uplifted terrain.

Dimensions of the rifts

In Africa the width between the boundary fractures is remarkably constant: for the most part, it is 40–55 kilometres (25–34 miles) and within limits of 30–75 kilometres (19–47 miles). Similar widths occur in other continental rifts: the Rhine Rift Valley is 30-40 kilometres (19–25 miles) wide, and the Baikal Rift Valley is 55–70 kilometres (34–44 miles) wide. This similarity is quite meaningful, because structural studies show that rift dimensions partly reflect the thickness of the fractured crust. The Mid-Atlantic Rift, by way of contrast, is only 10–20 kilometres (6–12 miles) wide, and this reflects the fact that the Earth's crust thins markedly in oceanic areas.

The depths of rift valley floors below their uplifted margins are more variable than widths. Values vary from only a few hundred metres through 1,000 metres (3,300 feet) in the central sectors of the Ethiopian, Gregory, and Western rifts to 2,500 metres (8,200 feet) in the deep troughs occupied by Lake Nyasa and Lake Tanganyika. The floor of the Baikal Rift is as much as 3,000 metres (9,800 feet) below the bordering plateau surface. These large elevation differences are entirely attributable to depression of the rift floor, and not to any exceptional uplift of bordering plateaus.

Along the Eastern Rift of Africa, the elevations of the rift floor and the plateau margins vary in unison and increase from the fringes to the centre of each swell. Thus, the rift floor rises from 400 metres to 2,000 metres (1,300 to 6,600 feet) in central Kenya and from 500 metres to 1,700 metres (1,600 to 5,600 feet) in central Ethiopia. In these regions, however, there is a very thick infilling of the rift valleys with volcanic sediments that masks the presence of a basement floor as deep as that under Lake Tanganyika and Lake Baikal.

Rift valleys tend to be linear over distances of as much as several hundred kilometres, and any lateral displacements are usually accomplished in an *en echelon* pattern (parallel offsets, such as roofing shingles). Nevertheless, some strikingly arcuate trends mark the Western Rift and also the Baikal Rift. The Western Rift is an arc of about 900-kilometre (560-mile) radius, with a hypothetical focus near Kilimanjaro. Arcuate rift valleys appear to be unique to continents; oceanic rifts are characteristically linear.

Drainage and sedimentation

The plateaus bordering rift valleys are notably upwarped to produce "rift shoulders." This is clearly shown by the resulting drainage pattern; the rift valleys are left as narrow, isolated basins of internal drainage, without possible outflow.

Quaternary (occurring within the last 2,500,000 years) upwarping of the rift shoulders has visibly affected previously established drainage patterns. The classic example is the ponding of Lake Kyoga in northern Uganda from the upwarping of the eastern shoulder of the Western Rift at Lake Albert. The Victoria Nile is forced to flow up an old tributary valley, and thence circuitously into the northern end of Lake Albert, instead of following its old course directly westward into the southern end of that same lake.

Recent and continuing uplift of the East African swells has led to rapid downcutting by rivers and the formation of deep canyons. This dissection raises considerable problems for communications, especially in Ethiopia, where the relief is most violent and where, when convenient, the main roads tend to follow the watershed along the rift shoulders.

The floors of rift valleys are typically divided into a number of internal drainage basins, separated by broad transverse arches. In the Eastern Rift this results in a chain of relatively small lakes that, because of the dry climatic regime and very high insolation, (incoming solar radiation), are natural evaporating pans for the alkaline emanations of rift valley volcanoes. Lake Natron, in the Gregory Rift, takes its name from the salt that saturates its waters, natron (a hydrated sodium carbonate), which also occurs in abundance in Lake Shala, Ethiopia.

The small volume of drainage entering the rift valley lakes produces a relatively slow infilling by sediments, despite the violent erosion of rift escarpments by thunderstorm flash floods. The rate of sedimentary infilling may not keep pace with the rate of subsidence unless it is complemented by volcanism. The rate of rift-floor subsidence relative to elevation of the plateaus is geodetically measured to be a few millimetres per year in the Icelandic and Dead Sea rifts, and geological evidence indicates that a similar average rate has prevailed in the African rift valleys during the last 20,000,000 years.

Effectsof climate on configuration

The climate of eastern Africa has had an important effect on the morphology of the rift system. Neglecting the wetter glacial periods of the Pleistocene Epoch (10,000 to 2,500,000 years ago), the climate of the plateaus is essentially a long dry season terminated by a rainy season of a few months duration. The range of temperature rarely exceeds limits of 5° and 30° C (41" and 86" F). The rift valleys themselves are drier and hotter than the plateaus, and erosion is not of great significance within them. On the plateaus, however, geological durations of fluvial erosion have resulted in the formation of relatively flat erosion surfaces, with remnants of older surfaces preserved at successively higher levels (except where faulting has caused them to be moved down into the rift valley). Such surfaces have been subjected to uplift and warping at several times during their history. There have been several alternations of uplift and fluvial erosion since the Paleozoic Era (which ended 225,000,000 years ago) in Africa, and dating of the various surfaces by means of sediments and their contained fossils (or by radiometric dates, where available) provides a clear insight into the evolution of the swells and the subsiding rift valleys. (See further LANDFORM EVOLUTION for a discussion of erosion surfaces and the relation of climate and morphology.)

GEOLOGICAL HISTORY AND STRUCTURE

History of the East African Rift System. The East African Rift System traverses continental crust until it meets with oceanic crust at its junction with the Red Sea and Gulf of Aden. The foundation of the African continent is the Precambrian basement (older than 570,000,000 years), which consists of nuclei of rocks older than 2,000,000,000 years surrounded by interweaving and intersecting orogenic belts (zones of mountain building) of younger Precambrian age. The last episode of basement tectonism (Earth movements and deformation) affected primarily the border regions of the present African continent, during the 450,000,000–700,000,000-year period. Rocks of this age form the Mozambique Belt, which occupies the eastern fringe of Africa from Mozambique to Egypt. The structural trend of the belt is meridional, though some large departures do occur. The Mozambique Belt consists, at least in part, of older rocks that were altered under conditions of high temperature and pressure—*i.e.*, metamorphosed. The intensity of metamorphism and related deformation increases eastward, from the eastern edge of the ancient Precambrian nucleus of the Congo. Numerous plutons (deep-seated bodies of crystalline rock), dominantly granodiorite, were emplaced during and immediately after this orogenic episode.

The East African Rift System is situated almost entirely within the Mozambique Belt. The close relationship of rift valleys to the younger elements of basement structure, with a frequent parallelism of rift faults and basement foliation (grain of the rocks), has led to suggestions of a long-standing common cause. But several instances of regional nonparallelism occur, particularly in Ethiopia, and a genetic relationship between such differing structural features is most unlikely. The rift valley faults merely take advantage of the basement "grain" where the trend is suitable; they follow a path of least resistance. This is true also of the Baikal Rift.

The splitting up of Gondwanaland, the ancient continent of the Southern Hemisphere, in the Early Mesozoic Era (from 65,000,000 to 225,000,000 years ago), following a long period of erosion throughout the Paleozoic Era (from 225,000,000 to 570,000,000 years ago), led to marine invasion and sedimentation over eastern Africa. Fossiliferous limestones and sandstones of Jurassic–Cretaceous age (65,000,000 to 190,000,000 years old) were formed; today these rocks crop out on the high plateaus east and west of the Ethiopian Rift, bearing witness to the subsequent uplift phases of Tertiary (2,500,000 to 65,000,000 years ago) and Quaternary times.

Tertiary and Quaternary events

In the Early Tertiary, the present line of rift valleys was marked by an intermittent chain of troughs that became accentuated by uplift of the bordering plateaus (this uplift occurred in stages throughout the Tertiary and Quaternary). As the troughs deepened, they were partly filled by thick sedimentary deposits; in the Red Sea, these sediments were dominantly halite (sodium chloride, or common salt). In Ethiopia and Kenya, the fill was largely volcanic, with colossal outpourings of extruded lavas—fissure basalts and fissure phonolites (rocks rich in feldspar and feldspathoid minerals), respectively. Volcanism and sedimentation has continued through the Quaternary, especially in the northern half of the East African Rift System. During Pliocene–Pleistocene time (about 3,000,000 years ago), the most marked phase of plateau uplift occurred, and the rift troughs were finally faulted to their present form.

The surface rocks of the rift-valley floors are dominantly of Quaternary age. This has caused recent questioning of the concept of rift valleys as subsided blocks covered by younger sediments; some authorities have asked whether continental rifts may be somewhat like oceanic rifts, zones of crustal separation without a basement of ancient rocks. This will be discussed later, in the section treating theories of origin.

Surface structures. Rift valleys are zones of normal faulting—one block being dropped down relative to the other. Transcurrent faulting, involving primarily lateral movement, is important in the narrow Dead Sea Rift and possibly in the Baikal Rift, but in Africa the dominating movements at the surface have been vertical.

Kinds of tectonic features

The typical rift valley is formed by two parallel fault zones, of complementary direction and magnitude of vertical displacement, about 50 kilometres (30 miles) apart. Although faulting sometimes takes the form of a single, huge scarp (for example, the Nguruman Escarpment west of Lake Magadi and the Guraghe Escarpment south of Addis Ababa), several closely spaced, steplike faults are more commonly developed. The total displacement is typically in excess of 1,000 metres (3,300 feet), but accurate estimates are difficult to obtain where the thick fill of young rocks in the rift prevents the matching of displaced, older strata.

The classical, ditchlike rift valley, exemplified in western Uganda, central Kenya, and southern Ethiopia, is not developed everywhere along the East African Rift System. Almost equally common are asymmetric rifts, where one margin is typically faulted, but the opposing margin is a downwarp cut by small antithetic faults (which dip in the opposite direction); examples are the rifts at Lakes Natron, Magadi, Rukwa, Tanganyika, and Rudolf. In the downwarped margins of asymmetric rifts, the strata and erosion surfaces may incline as steeply as 30" into the rift. Another tectonic style is the occurrence of tilted blocks, without the occurrence of faulted or warped grabens. This type of structure occurs at the southern ends of the Gregory and Western rifts and in the western United States in the Basin and Range Province.

Crustal blocks that are uplifted between parallel faults are termed horsts. Large horsts composed of Precambrian rocks form the Mbeya, Kungwe, Ruwenzori, and Amaro mountains, and in some of these cases the horsts rise from the rift floor to higher elevations than the hordering plateaus. In the culminating example of Mt.

Ruwenzori (5,109 metres [16,758 feet]), upwarping has accompanied the raising of the horst, particularly on the eastern flank. They tend to be situated in the fork of rift-valley bifurcations.

Rift valleys and their boundary faults may be arcuate on a small scale (southern Gregory Rift) or on a large scale (Western Rift). Sometimes an abrupt change of trend results in a transition zone of intersecting faults, as at the northern end of the Aberdare Range, central Kenya. Very commonly, rift-valley faulting is transposed in an *en echelon* pattern; such a pattern is strongly developed at the northern end of the Ethiopian Rift. Thus, although east of Addis Ababa the rift faulting retains its north-northeastern trend, *en echelon* displacements of these faults to the right have resulted in a northeastern topographic trend to the rift valley itself.

The floors of the rift valleys are usually fairly flat. This is due to lacustrine sedimentation and volcanic deposition rather than a reflection of any underlying sunken plateau surface. The planar rift floors have been intensely fractured by a narrow belt of very recent faulting (with associated volcanic fissures and cones) that is generally median to the boundary faults of the rift. The median belt is five to ten kilometres (three to six miles) wide, and its closely spaced faults, termed grid faults, have produced numerous small horsts and grabens. Such fine fracturing of the rift floor is suggestive of strong crustal distension and thinning.

Association with alkaline lavas

Volcanism. Rift valleys are intimately associated with volcanism. Compared with the lavas of oceanic ridge rifts or of island arcs, rift-valley lavas and tuffs are enriched in sodium and potassium and depleted in silica and alumina. The extreme case is revealed in the carbonatite (carbonate rocks derived from the Earth's mantle) volcanoes and plutons of East and Central Africa, where the magma was extraordinarily enriched in carbonates at the expense of silicates. The active volcano Oldoinyo Lengai, at the southern end of the Gregory Rift, erupts ashes and lavas in which the chief constituent is sodium carbonate.

More abundant, however, are lavas ranging in composition from alkali olivine basalt, alkali trachyte to peralkaline rhyolite (comendite and pantellerite). Parallel to this series of rocks is a second series, poorer in silica, with feldspar minerals replaced by feldspathoids: this is the melaneyhelinite-nevhelinite-uhonoliteseries.

Volcanism shows an unmistakable yet capricious relationship to rift valleys in space and time. In Ethiopia, volcanism has been intense and is still continuing. Voluminous flood basalts commencing in the early Eocene (38,000,000 to 54,000,000 years ago) were followed by the building up of Hawaiian-like shield volcanoes (layered forms with a central vent) in the Miocene, but after that the volcanism abruptly became more silicic, with extensive ignimbrites (pyroclastic rocks deposited from hot volcanic clouds of debris) during the Pliocene (2,500,000 to 7,000,000 years ago). Quaternary volcanism has been restricted largely to the rift floor, and in particular to the median fault belt. Volcanoes are situated at *en echelon* displacements along the fault belt, and some of the faults have provided escape for fissure basalts.

In Kenya, the volcanism has been less voluminous, and commenced only in Miocene time (7,000,000 to 26,000,000 years ago). Its alkaline character is even more pronounced, however, and the extensive fissure phonolites of the Gregory Rift have no parallel in any other rift valley. The great volcanoes of Elgon, Kenya, Meru, and Kilimanjaro are all situated about 100 kilometres (60 miles) into the plateaus from the rift margins, but otherwise it is the rift floor that has been the main site for volcanism in the Pliocene to Quaternary time interval, and a chain of dormant trachyte calderas occurs along the median fault belt as in Ethiopia.

Volcanism in the Western Rift and Nyasa Rift has been sparse and is comparable to that in the Rhine and Baikal rifts. The volcanics of the Western Rift are in some cases strongly enriched in potassium, as compared with the soda-enriched volcanics of the Eastern Rift.

Geophysical data. Continental rifts, like oceanic rifts, are zones of shallow-focus earthquakes. The greatest recorded depth is about 60 kilometres (35 miles), but an average depth is about 20 kilometres (12 miles). Theoretical calculations indicate both vertical and occasionally lateral movements at the focuses. The largest recorded magnitude for an African rift earthquake is 6.75 on the Richter earthquake-intensity scale; by comparison, some of the largest earthquakes that have been recorded anywhere (*e.g.*, the Alaskan event of 1964) have a magnitude of about 8.5.

Earthquake, heat-flow, and gravity observations

The dispersion of earthquake epicentres outside the rifts in Africa is much greater than for the oceanic rifts. Regions of high seismicity occur at the northern end of Lake Tanganyika, the Ruwenzori–Lake Albert region. the southern end of the Gregory Rift, and the western margin of the Afar depression. By contrast, the northern part of the Gregory Rift is virtually aseismic, despite fresh faulting and active volcanism. These patterns of high and low seismicity appear to change over periods of a century or more, however.

Heat-flow measurements have so far been made only in the Nyasa Rift. There the values range from 0.5 to 0.7 microcalorie per square centimetre of surface per second (μ cal/cm^{2}sec) at the northern and southern ends of the lake to 2.3 μ cal/cm^{2}sec in the central region, and this abrupt change suggests a local heat source within a few tens of kilometres of the surface. Heat flow in the Eastern Rift is suspected to be much higher than this because abundant fumaroles, volcanic vents that emit steam and other gases, and hot springs occur; and in the Red Sea and Gulf of Aden rifts, values of 5 μ cal/cm^{2}sec or more compare with an average of 1.3 μ cal/cm^{2}sec for the Indian Ocean basins.

Gravity data confirm the attenuation of the crust under the rift valleys compared with the plateaus and also indicate the presence of lighter, modified mantle material beneath the swells. The latter represents a shallowing of the low-velocity layer of the upper mantle, and this conforms with the active magmatism of the rift zones and the failure of near-surface shear waves to be transmitted under these zones. Negative Bouguer gravity anomalies (indicating mass deficiencies relative to the expected mass of the Earth's crust) for the Gregory, Western, and Nyasa rifts are not attributable entirely to the presence of a light sedimentary infilling, and they indicate the accretion of unspecified light material, possibly syenite, at the base of the crust.

The median fault belt of the rift valleys is marked by a positive gravity anomaly in Kenya and Ethiopia. This suggests basaltic injection to shallow levels in a strongly distended crust. Where such injection reaches the surface, as in the Red Sea and the oceanic rifts along lines of sea-floor spreading, strong positive Bouguer anomalies occur.

ORIGIN OF RIFT VALLEYS

The "dropped keystone" theory

Early theories. In the early part of the 19th century, the Rhine Graben was first envisaged as a collapsed strip between the Vosges and Schwarzwald plateaus. Subsequently, the sinking of the graben was connected with an uparching of the adjacent plateaus, giving rise to the "dropped keystone of the arch" theory, a theory that was extended to include the East African Rift System in 1898.

The early explorers of eastern Africa realized that the long lake-filled valleys could not have been eroded by rivers. By the end of the 19th century it was clear that the rift valleys were of tectonic origin and that the nature of the faulting, especially the step faulting of the rift margins, signified tensional forces pulling at right angles to a massive line of weakness in the Earth's crust. The origin of the East African Rift System was related to the foundering and splitting of Gondwanaland, and the "dropped keystone" theory was widely accepted. The possible importance of crustal separation in rift genesis was also mooted in the 19th century, but this idea remained unpopular until the recent revival of the continental-drift hypothesis.

One of the fundamental problems in theories of rift genesis is to explain the uplift of the rift shoulders, and indeed the swell as a whole. Simple crustal tension could not explain this uplift, and in the early 1900s a compressional origin was suggested for rift valleys. The absence of exposed surfaces of thrusting along the rift margins, which would he expected if compression was involved, was explained in terms of subsequent downward sliding and faulting of the thrusted margins, which masked these supposed fundamental structures. Compression was held to account for the uplift of the swells, the upwarping of the rift shoulders, the exceptional depth of the rift in Lake Tanganyika (excessive for a dropped keystone), and the absence of volcanism from the deepest rifts. The deeper the rift, the stronger the presumed compression, and thus the more effective the bottling down of volcanic magmas; however, it is now suspected that, allowing for volcanic infill, the Eastern Rift is even deeper than Lake Tanganyika.

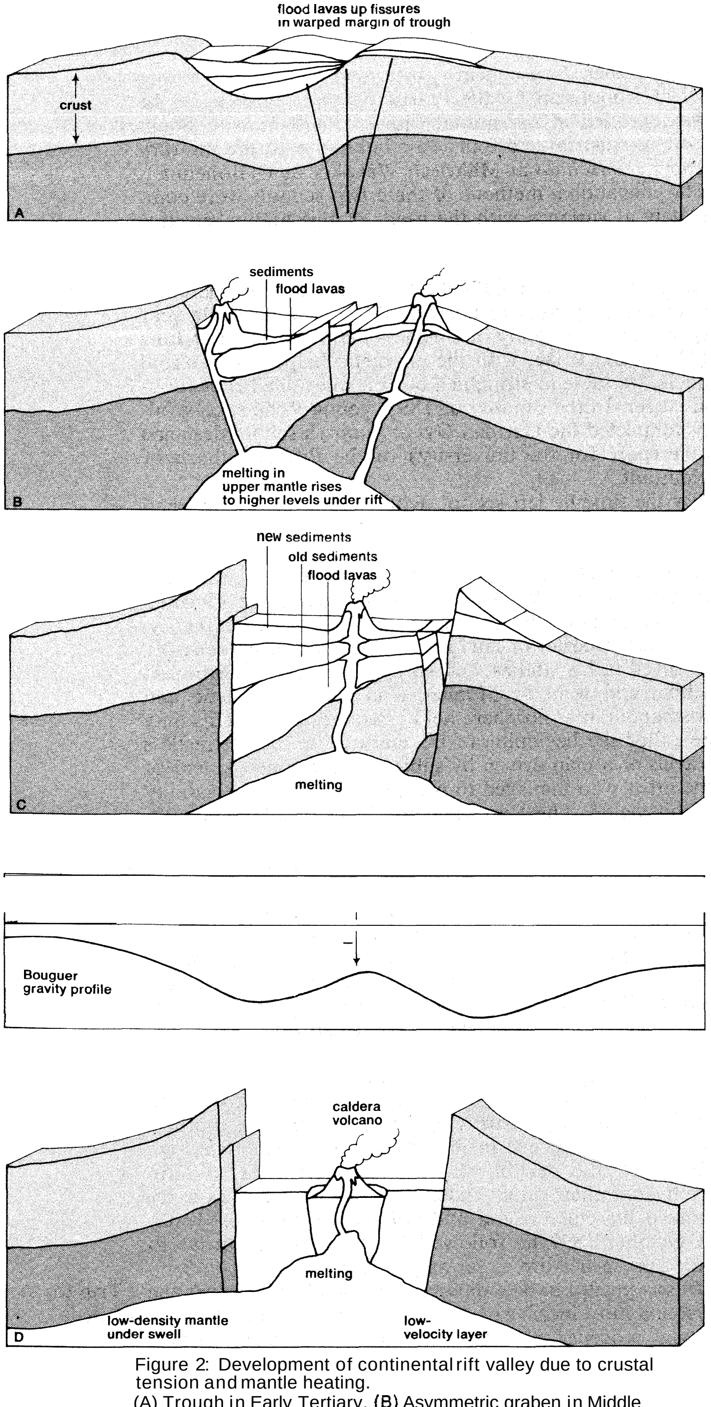


Figure 2: Development of continental rift valley due to crustal tension and mantle heating.
(A) Trough in Early Tertiary, (B) Asymmetric graben in Middle Tertiary, (C) Simple rift valley in Late Tertiary, (D) Graben-in-graben form (Quaternary) with corresponding Bouguer gravity profile.

The discovery in the 1930s that the East African rift valleys had strongly negative Bouguer gravity anomalies gave apparent confirmation to the compressional theory, the supposed thrust faults preventing the light rift block from rising to an equilibrium position. The anomalies are now explained as due to the infill of light sediments, however, and the accretion of silicic igneous rocks at the base of the crust under the rift. The absence of compressional folding, the confirmation that the main rift faults are, without exception, normal faults even in regions of bifurcation, and the focal mechanism of rift earthquakes all indicate tension as the basic process operating at rift valleys.

Sea-floor spreading and the tensional model

Modern theories. Since the early 1960s, the origin of rift valleys has been related to sea-floor-spreading processes, though the complicating presence of thick continental crust, compared to the thin crust of the oceans, has yet to be evaluated. The uplift of the swells, like the midoceanic ridges, is ultimately the result of high heat flow from an underlying hot zone in the mantle beneath the crust, which results in the transformation and expansion of rock mineral assemblages into lower pressure, lighter density forms. This thermal regime is also favourable for the generation of magmas of the alkaline type that are observed in rift valleys.

A model of the tensional processes thought to produce rift valleys demonstrates that a rigid crustal layer overlying a relatively mobile subcrust will fracture to produce a sunken graben block. The fracture planes are inclined at 60°–65° to the horizontal, and the graben width approximates the thickness of the fractured crust. Asymmetric graben can be produced if the graben block is free to sink into the mobile subcrust, a fractured margin being opposed by a warped margin. This simple tensional model accounts for some features of the African rift valleys, but not all.

As mentioned previously, the great depth of rift valleys, allowing for infilling in places, is incompatible with a "dropped keystone" model, and some absolute separation is also required, especially in the central regions of the swells. The grid faulting and caldera (*i.e.*, a greatly enlarged crater resulting from explosive activity and ultimate collapse) volcanism of the Eastern Rift floor suggests a thinning of the rift "block." And the gravity data show a concentration of low-density material under the rifts and adjacent plateau regions. These factors are compatible with crustal separation allied to high heat flow.

Oceanic rifts are entirely the result of crustal separation, the fissuring of the rift crust being healed by injection of basalt magma. This process is exposed in Iceland, where *gja* fissures—through which magma has failed to rise to the surface—flood basalts, and normal faults are intimately associated. *Gja* fissures recently have been identified in the Ethiopian Rift Valley. There are some major differences between the oceanic and African rifts, however. In the oceans, the ridge crust was generated at the rift and has since been transported outward and upward in the sea-floor-spreading process. The African plateaus are continental crust and have not been generated in the rift zone. Furthermore, continental rocks are present and exposed in some sectors of the East African Rift System; for example, at the southern end of the Ethiopian Rift.

A suggested sequence of events in the evolution of the African rift valleys is presented in Figure 2. This sequence shows that the rift valleys have been depressed regions from the beginning; that to the concept of a dropped keystone must be added downwarping of sufficient magnitude to produce the observed depths of rift valleys; and finally, that crustal separation has only begun to operate in recent geological times.

Plate tectonics and the Afar junction

The remarkable success of plate theory (the concept that the Earth's continents and ocean basins consist of six or more great plates that move about relative to each other) in unifying tectonics on a global scale has invited its application to such a major feature as the East African

Rift System. Plate geometry has been applied to the Afar triple junction, where the Ethiopian Rift meets with the proven sea-floor-spreading zones of the Red Sea and Gulf of Aden. The rates of spreading in these two zones have been calculated from a knowledge of magnetic lineations and fault orientations, as in the oceans; and these rates, together with their directions, uniquely determine the third conjoining spreading zone: the East African Rift System. This geometric computation is dependent on an absence of deformation outside the three spreading zones, which seismic and geological evidence indicates to be not altogether true.

The geometric result is that the East African Rift System has opened by 1.9" about a rotation pole located at 8.5" S, 31.0" E. This result fits the geological observations fairly well, especially when it is considered that this is the first application to a continental rift system. But many difficulties remain. The African rift valleys extend south beyond the rotation pole; they lack the linear offset pattern characteristic of the oceanic rifts and instead are arcuate and frequently bifurcate; and an opening of 1.9" is incompatible with the amount of continental crust exposed in some sectors of the rift valleys.

The model has been modified subsequently to better accord with observation, such that the indicated spreading rate in the African rift valleys is now estimated to be one millimetre per year or less, averaged over the last 20,000,000 years. This rate may well have accelerated in more recent times, but it fits with estimates for the Rhine and Baikal rift valleys, neither of which have yet been fitted into the global tectonic pattern. It seems certain that the presence of thick, light continental crust is modifying the spreading phenomenon observed in the oceans and that continental rifts are not necessarily potential oceanic basins.

BIBLIOGRAPHY. The following references are listed chronologically to better accord with the historical development of knowledge on rift valleys: E. SUESS, "Die Briiche des ostlichen Afrika," *Denkschr. Akad. Wiss., Wien,* 58:555–584 (1891), the first clear insight into East African tectonics; J.W. GREGORY, *The Great Rift Valley* (1896), the classic account of the East African rift valley; *Rift Valleys and Geology of East Africa* (1921), Gregory's final synthesis on the African rift valleys, of astonishingly wide compass; B. WILLIS, *East African Plateaus and Rift Valleys* (1936), a good historical review of ideas on rift valleys to date of publication; F. DIXEY, *The East African Rift System* (1956), a terse and thorough review of rift-valley geology, with emphasis on geomorphology rather than geophysics; A. HOLMES, "Plateaus and Rift Valleys," in *Principles of Physical Geology,* 2nd ed., ch. 29 (1965), a lucid summary of rift-valley structure and volcanism; B.H. BAKER, P.A. MOHR, and L.A.J. WILLIAMS, *Geology of the Eastern Rift System of Africa* (1971), a review of rift valleys in terms of geology, geophysics, volcanic petrogenesis, and plate tectonics.

(P.A.M.)

Rilke, Rainer Maria

Rainer (René) Maria Rilke was a major Austro-German poet whose contribution to 20th-century German literature has won worldwide recognition. Together with writers such as James Joyce, Marcel Proust, T.S. Eliot, and Franz Kafka, he stands as one of the founders and giants of modern literature.

Early life. Rilke was born on December 4, 1875, in Prague (then the capital of the Kingdom of Bohemia, one of the original members of the Austro-Hungarian monarchy), the only son of a not too happy marriage. His father, Josef, a minor civil servant, was a man frustrated in his professional career; his mother, the daughter of an upper middle class merchant and imperial councillor, was a difficult and pretentious woman, though not devoid of some talent, who felt that she had married beneath her station. She had a penchant for the nobility and high society and liked to dress in sombre black in the manner of aristocratic dowagers. She left her husband in 1884 and moved to Vienna so as to be close to the imperial court.

Education

The education afforded the young Rilke was ill planned and fragmentary. It had been decided that Rainer was to become an officer to assure him the social standing barred to his father. Consequently, after some years at a rather select school run by the Piarist brothers of Prague, he was enrolled in the military lower *Realschule* of Sankt Pölten (Austria) and four years later entered the military upper *Realschule* at Mahrisch-Weisskirchen (Bohemia). The educational methods of these two schools were completely at variance with the needs of this highly sensitive boy, and he finally was forced to leave the school prematurely because of poor health. In later life he called these years of his "misused" childhood a time of merciless affliction, a "primer of horror." After another futile year spent at the Academy of Business Administration at Linz (1891–92), Rilke, with the energetic help of a paternal uncle, was able to straighten out his misguided educational career. In the summer of 1895, young RenC successfully completed the German *Gymnasium* (a school designed to prepare for the university) of the Prague suburb of Neustadt.

Rilke, 1925.
By courtesy of the Swiss National Library

By the time he left school, Rilke had already published his first volume of poetry (1894), and there was little doubt in his mind that he would pursue a literary career. Matriculating at Prague's Charles University in the fall of 1895, he enrolled in courses in German literature and art history and also, to appease his family, read one semester of law. But he could not become really involved in his studies, and so in September 1896 he left school and went to Munich, a city whose artistic and cosmopolitan atmosphere held a strong appeal. This may be called the beginning of his mature life, of the restless travels of a man driven by subjective inner needs, and of the artist who managed to persuade others of the validity of his vision. The European continent in all its breadth and variety—Russia, France, Spain, Austria, Switzerland, and Italy, always Italy—was to be the physical setting of that life.

Mature life and works. In May 1897 Rilke met Lou Andreas-Salomé, who within a few weeks of their meeting became his mistress. Lou, then 36 years of age, was a native of St. Petersburg, the daughter of a Russian general of Huguenot descent and a German mother. In her youth she had been wooed by—and had turned down—the great philosopher Friedrich Nietzsche; 10 years before her meeting with Rilke she had married a German professor. Rilke's affair with Lou was a turning point in his life. More than mistress, she was surrogate mother, the leading influence in his *éducation sentimentale* and, above all, the person who introduced Russia to him. Even after their affair ended some years later, Lou remained his close friend and confidante until his death. In October 1897 he followed her to Berlin in order to take part in her life as far as possible.

Trip to Russia

Russia proved to be a milestone in Rilke's life. It was the first and most incisive of a series of "elective homelands," leaving a deeper mark than any of his subsequent discoveries, with the possible exception of Paris. He and Lou visited Russia twice at the turn of the century, first in the spring of 1899 and then in the summer of 1900. There he found an external reality that he saw as the ideal objective correlative, or symbol, of his feelings, his inner reality. Russia for him was imbued with an amorphous, elemental, almost religiously moving quality—a harmonious, powerful constellation of "God," "human community," and "nature"—the distillation (one is almost tempted to say) of the "cosmic" spirit of being.

Russia evoked in him a poetic response that he later said marked the true beginning of his serious work: a long three-part cycle of poems written between 1899 and 1903 and published in 1905 under the title ***Dns Studen-Buch,*** which is dedicated to Lou. Here the poetic "I" presents itself to the reader in the guise of a young monk who circles his god with swarms of prayers, a god conceived not as an extraterrestrial or superterrestrial "Thou" but as the incarnation of "life," as the numinous quality of the innerworldly diversity of "things." The language and motifs of the work are largely those of Europe of the 1890s. Art Nouveau, moods inspired by the dramas of Henrik Ibsen and Maurice Maeterlinck, the enthusiasm for art of John Ruskin and Walter Pater, and, above all, the emphasis on "life" of Nietzsche's philosophy—all of these appear in the ***Stunden-Buch.*** And yet, the self-celebratory fervour of these devotional exercises, with their rhythmic, suggestive power and flowing musicality, contained a completely new element. In them, a poet of unique stature had found his voice.

Soon after returning from his second trip to Russia, Rilke accepted an invitation to join the artists' colony of Worpswede, near Bremen, where he hoped to settle down in the attractive countryside among congenial artists experimenting with developing a new life-style. In April 1901 he married Clara Westhoff, a young sculptor from Bremen who had studied with Auguste Rodin. The couple set up housekeeping in a farm cottage in nearby Westerwede. There Rilke worked on the second part of the ***Stunden-Buch*** and also wrote a book about the Worpswede colony, which appeared in 1903. In December 1901 Clara gave birth to a daughter, and soon afterward the two decided on a friendly separation so as to be free to pursue their separate careers.

Move to Paris

Rilke, having been commissioned by a German publisher to write a book about Rodin, left for Paris (where the sculptor lived) in August 1902. For the next 12 years, until the outbreak of World War I, Paris was to be the geographic centre of Rilke's life. During those years he frequently left the city for short or lengthy visits to other cities and countries, beginning in the spring of 1903 when, to recover from what seemed to him the indifferent and cruel life of Paris, he went to Viareggio, Italy. There he wrote the third part of the ***Stunden-Buch.*** For a time he also lived and worked in Rome (1903–04), in Sweden (1904), and repeatedly in Capri (1906–08); he also made voyages of discovery to the south of France, to Spain, Tunisia, and Egypt and frequent visits to his friends in Germany and Austria. Yet Paris was his second elective home, no less important than Russia, both for its historic, human, "scenic" qualities and its intellectual challenge.

Rilke's Paris was not the ***belle époque*** capital steeped in luxury and eroticism; it was a city of abysmal, dehumanizing misery, of the faceless and the dispossessed, and of the aged, sick, and dying. It was the capital of fear, poverty, and death. This preoccupation with these phenomena combined with a second one: his growing awareness of new approaches to art and creativity, an awareness gained through his association with Rodin. Their friendship lasted about four years, until the spring of 1906. Rodin taught him his personal art ethic of *toujours travailler*, or unremitting work, which stood in sharp contrast to the traditional idea of artistic inspiration. Rodin's method was one of fanatical dedication to detail and nuance and of unswerving search for "form" in the sense of concentration and objectivization. Rodin also gave Rilke new insight into the treasures of the Louvre, the Cathedral of Chartres, and the forms and shapes of Paris. Of the literary models France had to offer, the poet Charles Baudelaire impressed and influenced him the most.

During those Paris years Rilke developed a new style of lyrical poetry, the so-called ***Ding-Gedicht*** ("object poem"), which attempts to capture the plastic essence of a physical object in language. Some of the most successful of these poems are imaginative verbal translations of certain works of the visual arts. Other poems deal with such subjects as landscapes, portraits, and biblical and mythological themes as a painter would depict them. These ***Neue Gedichte*** (1907–08) represented a departure from traditional German lyric poetry. Here Rilke forced his language to such extremes of subtlety and refinement that one is tempted to characterize it as a distinct art among other arts and a language distinct from existing languages. The worldly elegance of these poems cannot obscure their inherent emotional and moral engagement. When Rilke, in the letters about Paul Cézanne written in the autumn of 1907, defines the painter's method as a "using up of love in anonymous labour," he doubtless was also speaking of himself. In a letter to Lou Salomé written in July 1903, he had defined his method with this significant formulation: "making objects out of fear."

Die Aufzeichnungen ***des*** *Malte Laurids Brigge* (***The Notebook of Malte Laurids*** *Brigge*, 1930), on which he began work in Rome in 1904 and which was published in 1910, is a prose counterpart to the ***Neue Gedichte.*** The two works complement one another. That which hovered in the background in the poems, behind the perfection of style, is in the foreground of the prose work; the subjective, personal problems of the lonely occupant of a Paris hotel room, the "fear" that is the inspiration for the creation of "the objects." And if the poems seem like a glorious affirmation of the Symbolists' idea of "pure poetry," the ***Aufzeichnungen*** read like a brilliant early example of Existentialist writing. They are an artfully assembled suite of descriptive, reminiscent, and meditative parts, supposedly written by Malte, a young Danish expatriate in Paris (a member of the sophisticated, world-weary, and despairing generation of fin ***de siècle*** European decadents) who refuses to abide by the traditional chronology of narrative exposition but, instead, presents his themes as "simultaneous" occurrences set against a background of an all-encompassing "spatial time." Here are found all of Rilke's major themes: love, death, the fears of childhood, the idolization of woman, and, finally, the matter of "God," which is treated simply as a "tendency of the heart." The work in its entirety must be seen as the description of the disintegration of a soul—but a disintegration not devoid of a dialectic mental reservation: "Only a step," writes the narrator (Malte), "and my deepest misery could turn into bliss."

Themes of the *Aufzeichnungen*

The price Rilke had to pay for these two masterpieces was a writing block and depression so severe that it led him to toy with the idea of giving up writing altogether. Aside from a short and not first-rate poetry cycle, ***Das Marienleben,*** which appeared in 1913, he did not publish anything for 13 years. The first works in which he transcended even his ***Neue Gedichte*** were written in the early part of 1912—two long poems in the style of elegies. He did not undertake their immediate publication, however, because they promised to become part of a new cycle. He wrote these two poems while staying at Duino Castle, near Trieste, as the guest of Princess Marie von Thurn und Taxis, whom he had met and befriended in 1909.

The outbreak of World War I found him in Munich, and there for the time being he decided to remain. He spent most of the war years there. In December 1915 he was called up for military service with the Austrian army at Vienna, but by June 1916, he had returned to civilian life. The social climate of these years was inimical to his way of life and to his poetry, and by the time the war ended he felt almost completely paralyzed. He had only one relatively productive phase: the late fall of 1915, when, in addition to a series of superb new poems, he wrote the "Fourth Duino Elegy."

Late life and works. Rilke spent most of the remaining seven years of his life in Switzerland, the last of his series of elective homes. There he once more came into full command of his creative gifts. In the summer of 1921 he took up residence at the Château de Muzot, a small, old castle in the Rhône Valley, as the guest of a wealthy Swiss patron. In February 1922, within the space of a few days of obsessive productivity, he completed the Duino cycle begun years earlier. And unexpectedly and almost effortlessly, another superb cycle of 55 poems, in mood and theme closely related to the *Elegies*, came into being—his *Sonette an Orpheus* (*Sonnets to Orpheus*).

The myth-like character of the *Duino Elegies*

The *Duineser Elegien (Duinn Elegies)* are the culmination of the development of Rilke's poetry. That which in the *Stunden-Buch* had begun as a naïvely uncertain celebration of "life," as a devotional exercise of mystical worship of God, and which in *Malte* led him to assert that "this life suspended over an abyss is in fact impossible" in the *Elegies* sounds an affirmative note, in panegyric justification of life as an entity: "The affirmation of life and death prove to be identical in the *Elegies*," wrote Rilke in a letter (1925) in which he sought to define the meaning of his "expression of being." These poems can be seen as a new myth that reflects the condition of "modern" man, the condition of an emancipated, "disinherited" consciousness maintaining itself as a counterpart to the traditional cosmic image of Christianity. Like Nietzsche, Rilke opposes the Christian dualism of immanence and transcendence. Instead, he speaks out for an emphatic monism of the "cosmic inner space," gathering life and death, earth and space, and all dimensions of time into one undifferentiated, all-encompassing unity. This Rilkean myth is articulated in an image-laden cosmology that, analogous to medieval models, sees all of reality—from animal to "angel"—as a hierarchical order. This cosmology in turn results in a systematic, consistent doctrine of life and being in which man is assigned the task of transforming everything that is visible into the invisible through the power of his sensory perceptions: "We are the bees of the invisible." And this ultimate fate of man is concretized in the activity that alternately is called "saying," "singing," "extolling," or "praising." Thus it comes about that the poet is turned into the protagonist of humanity, its representative "before the Angel" (the pseudonym of God), as in the "Ninth Elegy," and even more strikingly in the *Sonnets* to *Orpheus*. This message of the late Rilke has been celebrated by some as a new religion of "life" and rejected by others as the expression of an unbridled aestheticism and an attempt on the part of the poet at "self-redemption" by virtue of his personal gift. It might be best to refrain from an overproblemizing of his message and instead look at the poetry that celebrates itself as a plea to value poetry "as poetry"—as a free play of the imagination, which, by the very act of raising problems of meaning, disposes of them.

Other late poems

The triumphant breakthrough of February 1922 was Rilke's last major contribution, yet both thematically and stylistically some of his late poems go beyond even the *Elegies* and the *Sonnets* in their experimentation with forms that no longer seem at all related to the nature of the poetic language of the 1920s. In addition to these late works he also wrote a number of simple, almost songlike poems, some short cycles, and four collections in French, in which he pays homage to the landscape of Valais, tendering his thanks.

Muzot remained his home until his death, but still he continued his travels, mostly within Switzerland, devoting himself to his friends and his vast, superbly articulate correspondence. In the early part of 1925 he once more went to Paris, with whose literary life he had remained in close touch. He was royally received by such old friends as André Gide and Paul Valéry as well as by new admirers; for the first and only time in his life he found himself at the centre of a literary season in a European metropolis. But the strain of this visit proved too much for his frail health. On August 18, abruptly and unannounced, he slipped out of Paris. He had been in ill health since the end of 1923, with frequent visits to sanatoriums and health resorts. The cause of his debility, a rare form of incurable leukemia, was not diagnosed until a few weeks before his death. He died on December 29, 1926, at Valmont, a sanatorium above Territet, on Lake Geneva. In accordance with his wishes, he was buried in the cemetery adjoining the church at Raron, a half-hour drive from Muzot upstream along the Rhône.

MAJOR WORKS

VERSE: *Leben und Lieder. Bilder und Tagebuchblatter* (1894); *Wegwarten* (1896); *Larenopfer* (1896); *Im Friihfrost* (1897); *Traumgekrönt* (1897); *Advent* (1898); *Mir zur Feier* (1899; republished with alterations as *Die friihen Gedichte*, 1909); *Das Buch der Bilder* (1902; 2nd ed., with 37 additional poems, 1906); *Das Stunden-Buch* (1905), consisting of *Das Buch vom monchischen Leben*, *Das Buch von der Pilgerschaft*, and *Das Buch von der Armut und vom Tode* (*The Rook of Hours*, trans. by A.L. Peck, 1961); *Neue Gedichte*, 2 vol. (1907–08; *New Poems*, trans. by J.B. Leishman, 1964); *Requiem* (1909; *Requiem and Other Poems*, trans. by J.B. Leishman, 2nd ed., 1949); *Das Marienleben* (1913; *The Life of the Virgin Mary*, trans. by Stephen Spender, 1951); *Duineser Elegien* (1923; *Duino Elegies*, trans. by J.B. Leishman and Stephen Spender, 4th ed., 1963); *Die Sonette an Orpheus* (1923; *Sonnets to Orpheus*, trans. by J.B. Leishman, 2nd ed., 1946); *Späte Gedlichte* (1934; *Later Poems*, trans. by *J.B.* Leishman, 1938); *Poèmes français* (1935); *Aus dem Nachlass des Grafen C.W.* (1950; *From the Remains of Count C.W.*, trans. by J.B. Leishman, 1952). Other translations by J.B. Leishman include: *Selected Works*, vol. 2, *Poetry* (1954–60); and *Poems* 1906–26, 2nd ed. (1959).

PROSE: *Vom lieben Gott und Anderes* (1900; *Stories of God*, 1932); *Worpswede* (1903); *Auguste Rodin* (1903; Eng. trans., 1919); *Die Weise von Liebe und Tod des Cornets Christoph Rilke* (1906; *The Tale of the Love and Death of Cornet Christopher Rilke*, trans. by *M.D.* Herter Norton, 1932); *Die Aufzeichnungen des Malte Laurids Brigge* (1910; *The Notebook of Malte Laurids Brigge*, trans. by John Linton, 1930); *Aus der Friihzeit Rainer Maria Rilkes. Vers. Prosa. Drame. 1894–1899* (1921); *Ewald Tragy* (1927–28); *Erzählungen und Skizzen aus der Friihzeit* (1928). Other translations of Rilke's prose are in *Selected Works*, vol. 1, *Prose*, by G. Craig Houston (1954–60).

BIBLIOGRAPHY

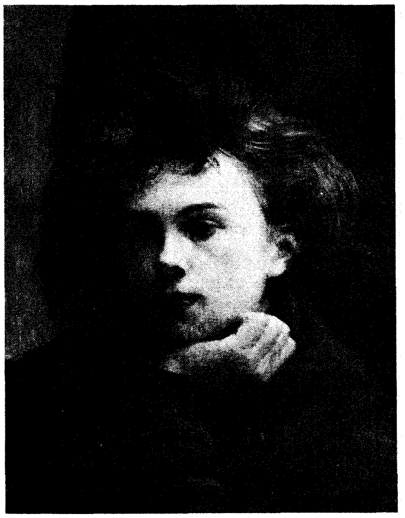
Editions and correspondence: Gesammelte Werke, 6 vol. (1927); *Sämtliche Werke*, 6 vol. (1955–66); *Werke*, 3 vol. (1966); *Gedichte in franzosischer Sprache* (1949); *Gesammelte Briefe*, 6 vol. (1936–39); *Briefe*, 2 vol. (1950); *Tagebücher aus der Friihzeit* (1942); *Briefe an einen jungen Dichter* (1929; Eng. trans., *Letters to a Young Poet*, rev. ed., 1954); *Briefe an eine junge Frau* (1931; Eng. trans., *Letters to a Young Woman*, 1945).

Biography and criticism: Because there is not yet a comprehensive, scholarly Rilke biography, the following may be recommended as sources of information: HANS E. HOLTHUSEN, *Rainer Maria Rilke in Selbstzeugnissen und Bilddokurnenten* (1967; Eng. trans., *Portrait of Rilke*, 1971), an illustrated biographical-critical essay; INGEBORG SCHNACK (ed.), *Rilkes Leben und Werk im Bild* (1956), a biographical survey in photographic documents; and JEAN R. DE SALIS, *Rainer Maria Rilkes Schweizer Jahre* (1952; Eng. trans., *Rainer Maria Rilke: The Years in Switzerland*, 1964), which treats only the last part of Rilke's life in detail. Memoirs that can be designated as classic are MARIE VON THURN UND TAXIS, *Erinnerungen an Rainer Maria Rilke* (1933); and the brilliant essays of RUDOLF KASSNER in *Buch der Erinnerung* (1938), *Umgang der Jahre* (1949), and *Geistige Welten* (1958). Two earlier monographs by friends of the poet are LOU ANDREAS-SALOME, *Rainer Maria Rilke* (1928); and NORA WYDENBRUCK, *Rilke, Man and Poet* (1949). Among the critical works on Rilke in English, the following are prominent: ELIZA M. BUTLER, *Rainer Maria Rilke*, 2nd ed. (1945); E.C. MASON, *Rilke, Europe, and the English-Speaking World* (1961) and *Rilke* (1963); also, H.W. BELMORE, *Rilke's Craftsmanship* (1954), a special investigation interested in formal questions. A comprehensive presentation from the psychoanalytic point of view is ERICH SIMENAUER, *Rainer Maria Rilke: Legende* und *Mythos* (1953). So far as the extensive German Rilke literature is concerned, after the enthusiasm of the 1930s and 1940s, it passed through a phase of disillusionment, indeed even of devaluation. At the end of the 1960s, however, in books such as JACOB STEINER, *Rilkes Duineser Elegien*, 2nd ed. (1969); and KAETEHAMBURGER (ed.), *Rilke in neuer Sicht* (1971), a younger generation of scholars came forth with new arguments on the classical merit of Rilke.

(H.E.H.)

Rimbaud, Arthur

Arthur Rimbaud was a French poet and adventurer, who, although his creative life lasted only from his 15th to his 20th year, became, unbeknown to himself, a revered model for the Symbolist movement, whose devotees sought in their poems to combine a technical perfection of language with subtly mysterious imagery.



Rimbaud, detail of an oil painting by Henri Fantin-Latour, 1872. In the Louvre, Paris.
Giraudon

Few poets have been the object of more passionate study or have exercised greater influence on modern poetry. It is in the prose poems *Illuminations* that he reached the highest peak of his originality, and this is the form best suited to his elliptical and esoteric style. He stripped the prose poem of all the anecdotal, narrative, and even descriptive content of his predecessors and, by divesting words of their dictionary meaning or logical content, endowed the poem with a quasi-magical power intended to evoke a state of mind—"&tat*d'âme*," the Symbolists were to say. He also showed how much rich material for poetry exists in the subconscious mind and in the half-remembered sensations of childhood. His writings are still capable of expressing intensely modern revolt and recoil from the very essence of life in a so-called civilized society.

Childhood. Rimbaud was born at Charlesville in the Ardennes, in northeastern France, on October 20, 1854, the second son of an army captain and a local farmer's daughter: his brother was a year older, and there were two younger sisters. In 1860 Captain Rimbaud and his wife separated, and the children were brought up by their mother, who hoped to keep her sons from following the example of her two wastrel brothers. Arthur, who early displayed unusual intellectual ability, showed a gift for writing at the age of eight. Later, he was the most brilliant pupil at the Collège de Charleville. He had a particular talent for Latin verse and in August 1870 won the first prize for a Latin poem at the Concours Acadtmique. His first published poem had appeared in January 1870 in La *Revue pour Tous*, and in May he had submitted to the distinguished poet Théodore de Banville, one of the selection committee for *Le Parnasse Contemporain,* a number of poems including one entitled "Credo in Unam," which, although derivative, showed signs of an original genius.

Early escapades

The outbreak of the Franco-Prussian War (July 1870) ended his formal education. In August he ran away to Paris but was arrested for travelling without a ticket and spent some days in prison. His friend and schoolmaster Georges Izambard paid his fine and had him sent to Douai, where he spent three weeks at the house of Izambard's aunts and where he joined the national guard. Izambard then took him home, but in October he disappeared again, wandering through northern France and Belgium in the wake of the invading armies. Again, he arrived at Douai, where he made fair copies of the poems composed during his two weeks of freedom, hunger, and rough living. These express an innocent joy in life and liberty and are his first wholly original works. His mother had him brought back by the police, but in February 1871 he sold his watch and again went to Paris, where for a fortnight he lived in great poverty.

Revolt and poetic vision. In early March he returned home on foot, a completely changed character, perhaps as a result of some shattering, yet enlightening, experience. He repudiated his early verses as false and wrote some of his most violent and blasphemous poems, expressing a disgust with life, a desire to escape into a world of innocence, and a sense of struggle between good and evil. His behaviour matched his poetic mood—he refused to work and spent his days drinking in cafés, in conscious revolt against religion, morality, and every kind of discipline. At the same time, he read books on occult philosophy, Kabbalism, magic, and alchemy and formulated his new aesthetic doctrine, expressed in two letters (May 13 and 15, 1871), later called "Lettres du voyant." This title is based on the belief that the poet must become a seer, a *"voyant,"* who can penetrate infinity and who, by breaking down the restraints and controls that make up the conventional conception of individual personality, must become the instrument for the voice of the eternal. But, because the new visions cannot be rendered in contemporary forms, a new language must be invented, a language accessible to all the senses—this is the ideal of "total art" first formulated by Charles Baudelaire and adopted by the Symbolists.

At the end of August 1871, on the advice of a literary friend in Charleville, Rimbaud sent to the poet Paul Verlaine specimens of his new poetry, among them the sonnet "Voyelles" in which he attributes to each vowel a different colour—"A noir, E blanc, I rouge, U vert, O bleu." The poem has been variously interpreted, but its true beauty lies in the originality of its images. Verlaine, impressed by the brilliance of the poems, summoned Rimbaud to Paris and sent the money for his fare. In a burst of self-confidence, Rimbaud composed "Le Bateau ivre." Although traditional in versification, it is a poem of astonishing verbal virtuosity and of daring choice of images and metaphors, inspired by a deep emotional and spiritual experience. In this masterpiece, Rimbaud reached one of the highest peaks of his art.

Meeting with Verlaine

Arriving in Paris in September 1871, he stayed for three months with the Verlaines, who were living with Mathilde Verlaine's parents, and met most of the well-known poets of the day but antagonized them all—except Verlaine himself—by his arrogance, boorishness, and obscenity. Requested to leave, he lived for some weeks in destitution and then in an attic in the house where Banville was living. Again his behaviour caused complaints, and in January 1872 he moved to a room where he lived, on money subscribed by Verlaine and his friends, a life of drink and debauch and was involved in a homosexual relationship with the older poet that gave rise to scandal. In March he returned to Charleville so that Verlaine could attempt a reconciliation with his wife but was recalled in May by his friend, who vowed that he could not live without him.

During this period (September 1871–July 1872), Rimbaud composed the last of his poems in verse, which show an advance in technical freedom and originality on those written earlier. This freedom is daring for that period then at its height, during which prevailed the strictly impersonal themes and descriptive imagery of the Parnassian poets, so-called because of the anthology, Le *Parnasse Contemporain*, to which they contributed. At this time he also composed the work that Verlaine called his masterpiece, "La Chasse spirituelle," the manuscript of which disappeared when Verlaine and Rimbaud went to England. Some critics also assign to this creative period the transcendental prose poems *Illuminations,* not one of which was dated by Rimbaud himself.

In July 1872 Verlaine abandoned his wife and fled with Rimbaud via Belgium to London, where they lived in Soho. Rimbaud may there have composed some of his *Illuminations.* He returned home for Christmas but was recalled (January 1873) by Verlaine, who dramatized an illness to play on his sympathy. In April Rimbaud went to the farm at Roche, near Charleville, where his mother and sisters were staying, and began to write what he called his "Livre païen, ou Livre nègre," which eventually became *Une Saison en enfer.* Verlaine was meanwhile

Quarrel with Verlaine

staying at Jehonville, near Roche, and in May he persuaded Rimbaud to accompany him to London. Rimbaud's sense of guilt at yielding to an influence from which he had wished to escape caused him to treat Verlaine with sadistic cruelty, varied by remorseful kindness. Finally, at the beginning of July, after one of their frequent quarrels, Verlaine abandoned Rimbaud and went to Belgium. Failing, however, to effect a reconciliation with his wife, he sent for Rimbaud and begged him to return with him to London. When Rimbaud tried to leave, Verlaine shot at him, wounding him in the wrist, and threatened to do so again. Rimbaud, terrified, asked a policeman for protection; Verlaine was arrested, jailed, and later sentenced to two years' imprisonment. After a week in the hospital, Rimbaud returned to Roche. There, he finished *Une Saison en enfer*, an account of his spiritual descent into hell and of his failure in art and love, which acquired some of its tragic character from the events in Brussels. It was printed in Belgium, and, in October 1873, he went to Brussels to collect the copies and then to Paris to arrange for its publicity. Discouraged by its reception there and unable to pay the printers, he abandoned the whole edition and, returning on foot to Charleville, he is said to have burned his manuscripts and papers. The bales of the book remained in the attics of the printers, where they were discovered in 1901 by the Belgian bibliophile Léon Losseau. Out of consideration for his fellow bibliophiles, who possessed the only copies known to exist, Losseau did not make his discovery public until 1915.

In February 1874 Rimbaud returned to London with Germain Nouveau, a wild bohemian poet. There they earned a precarious living in menial employment: Rimbaud may also have been composing some of the *Illuminations*. Nouveau returned to Paris in June; and Rimbaud seems to have fallen ill or to have suffered acutely from poverty, for his mother and elder sister went over to help him find a job. At the end of July he took up a post at a coaching establishment at Reading, Berkshire, but went home for Christmas, never to return.

In January 1875, he went to Stuttgart to learn German, for his interest then was in languages. There, he was visited by Verlaine, but this, their last meeting, ended in a violent quarrel. It was probably then that Rimbaud gave Verlaine the manuscript of *Illuminations*.

Traveller and trader. In May Rimbaud left for Italy, crossing the Alps on foot but, taken ill with sunstroke, was repatriated by the French consul at Leghorn. He spent the winter at home, learning Arabic, Hindustani, and Russian, and, in April 1876, set out for Russia. After being robbed of money and luggage in Vienna, he was arrested as a beggar and again sent home. In May he went to Holland, where he enlisted in the Dutch colonial army. He arrived at Batavia (now Djakarta, Indonesia) on July 23 but deserted on August 15 and eventually reached home at the end of December. In the spring of 1877, he went to Hamburg, and, failing to find work on a ship sailing east, he is said to have joined a circus going to Scandinavia, though he returned to France in the summer. In the autumn he set out for Alexandria but, taken ill on board ship, was landed in Italy and again returned home. There, he remained, enfeebled by illness, until October 1878, when he again set off for Egypt. He took a boat from Genoa to Alexandria and then went to Cyprus, where he worked as a labourer; but in June 1879 he fell ill with typhoid fever and returned home to recover.

Life in Ethiopia

It was during this winter of illness that he apparently decided to abandon his life of wandering and to plan for the future. Returning to Cyprus in the spring, he found work as a builder's foreman; in August, however, he quarrelled with his employers and again set out in search of work. A coffee exporter at Aden, Pierre Bardey, sent him to open a trading post at Harer (Harar), in the interior, at that time occupied by the Egyptians, but, after the Mahdist uprising in the Sudan and the consequent Egyptian evacuation of Harer, the trading post was closed, and Rimbaud went back to Aden. From Harer, however, he had been sent by Bardey on a journey of exploration into Ogaden, the region south of Harer, where no white man had ever penetrated. His report of this journey, in the proceedings of the Société de Géographie (February 1884), aroused some interest.

After the Egyptian withdrawal from Ethiopia the contest for power between the emperor, Yohannes IV, and Menelik II, king of Shewa (Shoa), gave rise to an arms race. Rimbaud, deciding to risk his savings on an expedition to sell arms to Menelik, resigned his position with Bardey (October 1885) and, in October 1886, after a series of frustrating delays, set out for the interior. He reached the capital of Shewa in February 1887, but unforeseen circumstances and the rapacity and double-dealing of Menelik robbed him of his expected profits; and he returned to Aden no better off than when he had left it. Exhausted and embittered, he had no prospect of employment. He spent the summer in Cairo, where an account of his journey was published in *Le Bosplzove Égyptien* (August 1887), and then returned to Aden, where he looked in vain for work. In May 1888 he returned to Harer as manager of a trading station for a firm called Tian, which, as well as exporting coffee, gum, ivory, and hides, engaged in the profitable traffic in arms and ammunition. At first, Rimbaud's chief business was supplying Menelik with arms, but he may also have been in some way associated with the slave trade. After Yohannes IV was killed (1889), however, and Menelik became emperor, the profits from gunrunning declined. In addition, although Harer was the chief trading centre of the empire and the more settled conditions encouraged commerce, Rimbaud lacked the hardheaded business sense needed for success. He lived as simply as the poorest native, spending as little as possible in order to save money so that he could one day retire and live at leisure. His meanness to himself was matched by unobtrusive generosity to others, and the little house where he lived with a native woman became the meeting place of the Europeans in Ethiopia.

His gift for languages and his humane treatment of the Ethiopians made him popular with them, and by his honesty and sincerity he even managed to win the confidence of the chiefs and of Menelik's nephew, the Governor of Harer, who became his close friend. He began to wish to do "something good, something useful," and his letters to his mother reveal a desire for affection and intellectual companionship. The marriage of his servant and only close companion, Djami, made him long for a family of his own, and he planned to go home for a holiday in order to look for a wife, in the spring of 1891.

Growing reputation in France

While Rimbaud was in Ethiopia, he had become known as a poet in France. Verlaine had written about him in *Les Poètes maudits* (1884) and had published a selection of his poems. These had been enthusiastically received, and in 1886, unable to discover where Rimbaud was or to get an answer from him, Verlaine published the prose poems, under the title *Illuminations*, and further verse poems, in the Symbolist periodical *La Vogue*, as the work of "the late Arthur Rimbaud." It is not known whether Rimbaud ever saw these publications. But he certainly knew of his rising fame after the appearance of *Les Poètes maudits*, for, in August 1885, he had received a letter from an old schoolmate, Paul Bourde, who told him of the vogue, of his poems—especially the sonnet "Voyelles"—among avant-garde poets. Also, he did preserve, for it was found among his papers, a letter received in July 1890 from a review inviting him to return to France and put himself at the head of the new literary movement, but he does not seem to have answered it. In February 1891 he developed a tumour on his right knee, and, when he left Harer on April 7, he had to be carried by stretcher for the week's journey to the coast. At Aden, the treatment tried had no result, and he was sent back to France. He landed at Marseilles in May and was taken to the Hospital of the Immaculate Conception, where his right leg was amputated. His mother had gone to Marseilles to be with him, but, despite the growing warmth of their letters toward the end of his time at Harer, she remained unable to show her affection for him, to his great disappointment. In letters to his sister Isabelle, he poured out his sense of frustration and despair, and, when, in July, he returned to Roche, a helpless cripple, it was she who looked after him.

He still hoped to marry and return to Ethiopia, but his health grew steadily worse. At last, made wretched by his helplessness and by the gloom of a bad northern summer, in August 1891 he set out on a nightmare journey to Marseilles. His disease was diagnosed as cancer, and Isabelle, who had accompanied him, was told that his case was hopeless Rimbaud himself, however, still believed that he might be cured and endured agonizing treatment. Shortly before he died, Isabelle persuaded him to make his confession to a priest. This conversion seemed to bring him new peace and to reawaken the poetic imagination of his early youth so that he became once more a *voyant,* seeing visions that, according to his sister, surpassed in depth and beauty those that had inspired the ***Illuminations.*** For this account, however, one has only the word of Isabelle, who in other respects (notably in her "emendations" of some of Rimbaud's letters from Ethiopia) has been proved a false witness. The conversion in *extremis,* accompanied by a reawakening of visionary power, belongs to the realm of uncertainty and legend. Rimbaud died in Marseilles on November 10, 1891, at the age of 37.

MAJOR **WORKS**

"Le Bateau ivre" (written 1871, first published in complete collection, 1898; ***The Drunken Boat; 36 Poems,*** trans. by B. Hill, 1952); "La Chasse spirituelle" considered by Verlaine his masterpiece—the manuscript written 1871–72 disappeared; ***Une Saison en enfer*** (1873; ***A Season in Hell,*** trans. and introd. by G.F. Lees, 1932; trans. by L. Varèse, 1945; and by N. Cameron, 1949); ***lllurninations*** (1886; ***Illuminations, and Other Prose Poems,*** trans. by H. Rootham, 1932; and by L. Varèse, 1957; also ***Rimbaud's Illuminations: A Study in Angelism,*** trans. by Wallace Fowlie, 1953, reprinted 1971); see also the ***Oeuvres complètes*** (1948); ***Selected Verse Poems of Arthur Rimbaud,*** trans. by N. Cameron (1942); ***Rimbaud,*** introd. and ed. by O. Bernard, with plain prose translation of each poem (1962); and ***Rimbaud: Complete Works, Selected Letters,*** trans. and introd. by Wallace Fowlie (1966).

BIBLIOGRAPHY

Editions and correspondence: The first collected edition of Rimbaud's writings was that of PATERNE BERRICHON and ERNEST DELAHAYE (1898). The principal editions of the ***Oeuvres complètes*** are'those of the "Pléiade," the first edited by ROLLAND DE RENEVILLE and JULES MOUQUET (1946; rev. eds., 1954, 1960, and 1965), and the new edition by ANTOINE ADAM (1972). The best annotated edition is that of SUZANNE BERNARD (1960; 2nd ed., 1962). HENRY DE BOUILLANE DE LACOSTE'S critical editions of ***Poésies*** (1939); ***Une Saison en enfer*** (1941); ***lllumінations*** (1949), have been followed by ALBERT PY'S annotated edition, with commentaries, of ***Illuminations*** (1967). Rimbaud's correspondence from the East was edited by PATERNE BERRICHON in ***Letfres de Jean-Arthur Rimbaud. Égypte, Arabie, Bthiopie*** (1899). Earlier letters were edited by J.-M. CARRE as ***Lettres de la vie litte'raire d'Arthur Rimbaud, 1870–1875,*** 5th ed. (1931). Important, recently discovered letters have been published in ***Arthur Rimbaud, Correspondance, 1888–1891,*** with preface and notes by JEAN VOELLMY (1965).

Biography: The indispensable biography for English readers is ENID STARKIE, ***Arthur Rimbaud,*** 3rd ed. (1961). It may be supplemented by HENRI MATARASSO and PIERRE PETITFILS, ***Vie d'Arthur Rimbaud*** (1962). Of the earlier notices one cannot ignore the authentic insights of PAUL VERLAINE in ***Les Poètes maudits*** (1884); but the first complete biography, albeit of dubious frankness, is that of PATERNE BERRICHON, ***La Vie de Jean-Arthur Rimbaud*** (1897).

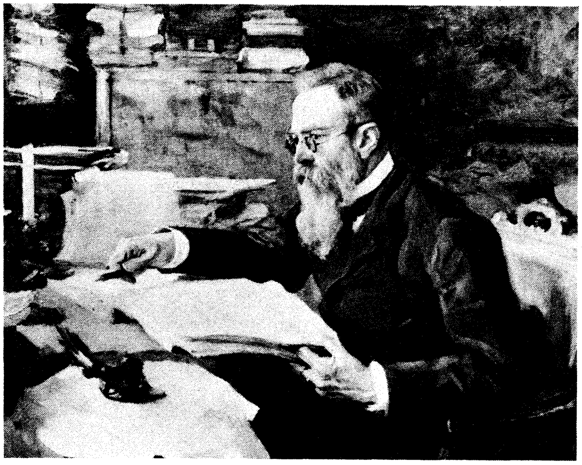
Criticism: In the vast corpus of literary criticism the reader may first consult EDGELL RICKWORD, ***Rimbaud, the Boy and the Poet,*** new ed. (1963); and the authoritative work by ROLLAND DE RENEVILLE, ***Rimbaud le voyant,*** rev. ed. (1947). The legend of Rimbaud is magisterially surveyed by the eminent critic RENE ETIEMBLE in ***Le Mythe de Rimbaud,*** 2 vol. (1952–54; new ed., 1961–68), to which his ***Rimbaud,*** new ed. (1950), in collaboration with YASSU GAUCLERE, may be considered as a preliminary. Subsequent studies include HENRY DE BOUILLANE DE LACOSTE, ***Rimbaud et le problème des "lllurninations"*** (1949); JACQUES GENGOUX, ***La Pensée poe'tique de Rimbaud*** (1950); CHARLES CHADWICK, ***Btudes sur Rimbaud*** (1960); W.M. FROHOCK, ***Rimbaud's Poetic Practice: Image and Theme in the Major Poems*** (1963); J.P. HOUSTON, ***The Design of Rimbaud's Poetry*** (1963); C.A. HACKETT, ***Autour de Rimbaud*** (1967); JACQUES PLESSEN, ***Promenade et Poésie, l'expe'rience de la marche et du mouvement dans I'oeuvre de Rimbaud*** (1967); and MARCEL RUFF, ***Rimbaud*** (1968), in French.

(En.S.)

Rimsky-Korsakov, Nikolay

A celebrated Russian composer and educator, Nikolay Andreyevich Rimsky-Korsakov is historically significant as a founder of what has become known as the Russian school of composition. He contributed to Russian musical culture by his teaching, conducting, and editing of works by Russian composers, including those of his own students. He composed in all genres, excelling particularly in scenes of fantasy and magic, which he depicted in his operas and symphonic works in luxuriant colours often imbued with Oriental inflections. In this respect he followed the traditions of Mikhail Glinka, who is often called the father of Russian music. In his last works, Rimsky-Korsakov approached the frontiers of Modernism in ultrachromatic and strikingly dissonant harmonies.

H. Roger-Viollet



Rimsky-Korsakov, portrait by V.A. Serov (18651911). In the Tretyakov Museum, Moscow.

Early life and naval career

Rimsky-Korsakov was born at Tikhvin, near Novgorod, on March 18 (March 6, old style), 1844. His father was a government official of liberal views; his mother was well educated and could play the piano. His uncle was an admiral in the Russian Navy and his elder brother a marine officer; from them Rimsky-Korsakov acquired his abiding love for the sea. When he was 12 the family moved to St. Petersburg (now Leningrad), where he entered the naval academy. At the age of 15 he began taking piano lessons with Theodore Canillé, a professional pianist, who also taught him the rudiments of composition. In 1861 he met the composer, Mily Balakirev, a man of great musical culture, and under the older man's guidance he began to compose a symphony. In 1862 he graduated from the naval academy. Soon afterward he sailed on the clipper ship "Almaz" on a long voyage, the vessel anchoring in New York, Baltimore, and Washington, D.C., at the height of the U.S. Civil War. The Russian sailors were cordially welcomed there, since Russia was politically sympathetic toward the North. Subsequent ports of call were Brazil (where Rimsky-Korsakov was promoted to the rank of midshipman), Spain, Italy, France, England, and Norway. The ship returned to its Russian home port in Kronstadt in May 1865. For young Rimsky-Korsakov the voyage confirmed a fascination with the sea. Aquatic scenes abound in his operas and symphonic works: the ocean in ***Scheherazade, Sadko,*** and ***The Tale of the Tsar Saltan,*** the lake in ***The Legend of the Invisible City of Kitezh and the Maiden Fevronia.***

On his return to St. Petersburg, he completed the symphony begun before his voyage, and it was performed with gratifying success in St. Petersburg on December 31, 1865, when the composer was only 21 years old. It was an auspicious beginning to his career; it was also the first performance of a full-fledged symphony by a Russian. His next important work was ***Fantasy on Serbian Themes*** for orchestra, first performed at a concert of Slavonic music conducted by Balakirev in St. Petersburg, on May *24, 1867.* The occasion was of historic significance, for, in reviewing the concert, the critic Vladimir Stasov proudly proclaimed that henceforth Russia, too, had its own "mighty little heap" (*moguchaya kuchka)* of native composers. The title caught on quickly and found its way into music history books, with specific reference to Rimsky-Korsakov, Balakirev, Aleksandr Borodin, César Cui, and Modest Mussorgsky, who became known collectively as "The Five," and whose purpose was to assert the musical independence of Russia from the West. Of these Rimsky-Korsakov was the most learned and the most productive; his works embrace all genres, but he excelled mostly in the field of opera.

Teacher and conductor

So high was Rimsky-Korsakov's reputation that in *1871,* still as a very young man, he was engaged to teach composition at the St. Petersburg Conservatory. In his autobiographical *My Musical Life,* he frankly admitted his lack of qualifications for this important position; he himself had never taken a systematic academic course in musical theory, even though he had profited from Balakirev's desultory instruction and by Tchaikovsky's professional advice. Eager to complete his own musical education, he undertook in *1873* an ambitious program of study, concentrating mainly on counterpoint and the fugue. He ended his studies in *1875* by sending *10* fugues to Tchaikovsky, who found them impeccable.

In *1873* Rimsky-Korsakov left the Russian naval service and assumed charge of military bands as inspector and conductor. Although he lacked brilliance as an orchestral leader, he attained excellent results in the training of inexperienced instrumentalists. His first professional appearance on the podium took place in St. Petersburg on March *2, 1874,* when he conducted the first performance of his *Third Symphony.* In the same year he was appointed to be director of the Free Music School in St. Petersburg, a post that he held until *1881.* He served as conductor of concerts at the court chapel from *1883* to *1894.* Between *1886* and *1900* he was chief conductor of the Russian symphony concerts. In *1889* he led concerts of Russian music at the Paris World Exposition; in the spring of *1907* he conducted in Paris two Russian historic concerts in connection with Sergey Diaghilev's Ballets Russes.

Music editor

Rimsky-Korsakov rendered an inestimable service to Russian music as the de facto editor and head of a unique publishing enterprise financed by the Russian industrialist Belayev and dedicated exclusively to the publication of music by Russian composers. After Mussorgsky's death, Rimsky-Korsakov edited his colleague's scores for publication, making radical changes in what he considered to be awkward melodic and harmonic progressions, and he practically rewrote Mussorgsky's opera *Khovanshchina.* His edited and altered version of *Boris Godunov* evoked sharp criticism as a pedantically professorial arrangement of a great, innovative masterpiece; but his masterly handling of the materials cannot be denied. Mussorgsky's score was later published in *1928* and had several performances in Russia and abroad, but ultimately the more effective Rimsky-Korsakov version prevailed in opera houses. Rimsky-Korsakov also edited (with the composer Glazunov) the posthumous works of Borodin.

A strict disciplinarian in artistic matters, he was also a severe critic of his own music. He made constant revisions of his early compositions, in which he found technical imperfections. As a result, double dates, indicating early and revised versions, frequently appear in his catalog of works. He was at his best and most typical in descriptive orchestration, in suggesting a place or an ambience. With but two exceptions *(Servilia* and *Mozart and Salieri),* the subjects of Rimsky-Korsakov's operas are taken from Russian or other Slavonic fairy tales, literature, and history. The most important among them are *Snow Maiden, Sadko, The Tsar's Bride, Tale of the Tsar Saltan, The Legend of the Invisible City of Kitezh and the Maiden Fevronia,* and *Le Coq d'or (The Golden Cockerel).* Although these operas are part of the regular repertory in Russian opera houses, they are rarely heard abroad; only *Le Coq d'or* enjoys occasional production in western Europe and America.

Of the composer's orchestral works, the best known are *Capriccio espagnol (1887;* Russian *Zspanskoye Kaprichchio),* the symphonic suite *Scheherazade* (1888), and *Russian Easter Festival* overture *(1888).* "The Flight of the Bumble Bee" from *The Tale of the Tsar Saltan* and the "Song of India" from *Sadko* are perennial favourites in a variety of arrangements. Rimsky-Korsakov's songs are distinguished by simple elegance and fine Russian prosody; his chamber music is of less importance. He also wrote a piano concerto.

As professor of composition and orchestration at the St. Petersburg Conservatory from *1871* until the end of his life (with the exception of a brief period during *1905* when he was dismissed by the reactionary directorate because of his defense of students on strike), Rimsky-Korsakov taught two generations of Russian composers, and his influence, therefore, was pervasive. Igor Stravinsky studied with him privately for several years. His *Practical Manual of Harmony (1884)* and *Fundamentals of Orchestration* (posthumous, *1913)* are still used as basic musical textbooks in the Soviet Union. Rimsky-Korsakov died at his estate in Lyubensk, on June *21* (June *8,* old style), *1908.*

MAJOR WORKS

Operas

Sixteen, including *Pskovityanka* (first performed *1873; The Maid of Pskov*); *Mayskaya noch (1880; May Night,* performed in French as *La Nuit de Mai); Snegurochka (1882; The Snow Maiden); Sadko (1898); Motsart i Salyeri (1898; Mozart and Salieri); Tsarskaya nevesta (1899; The Tsar's Bride); Skazka o tsare Saltane (1900; The Tale of the Tsar Saltan*), *Servilia (1902); Kashchey Bessmertny (1902; Kashchey the Immortal); Skazaniye o nevidimom grade Kitezhe i deve Fevroniy (1907; The Legend of the Invisible City of Kitezh and the Maiden Fevronia); Zolotoy petushok (1909;* performed in French as *Le Coq d'or* and in English as *The Golden Cockerel).*

Choral Works

Song of Oleg the Wise, after Pushkin, for tenor, bass, male chorus, and orchestra (composed *1899);* various settings of folk songs.

Orchestral

SYMPHONIES: *No.* I *in E Flat Minor (1861–65,* rev. *1884); No.* 2, *Antar (1868,* rev. *1876); No.* **3** *in C Major (1874,* rev. *1886).*

OTHER ORCHESTRAL MUSIC: *Overture on Russian Themes (1866); Piano Concerto in C Sharp Minor (1882–83); Fantasy on Russian Themes* for violin and orchestra *(1886); Fantasy on Serbian Themes (1867); Capriccio espagnol (1887);* symphonic suite *Sheherazade (1888);* overture *Russian Easter Festival (1888).*

Chamber Music

String Quartet in F Major (1875); String Sextet in A Major (1876): Quintet for Piano and Winds in B Flat Major (1876); String Quartet in G Major (1897).

Songs

More than *80* songs, mostly written in sets of four.

BIBLIOGRAPHY. The primary source is Rimsky-Korsakov's posthumously published *Летопись моей музыкальной Жизни (1909;* Eng. trans. *My Musical Life, 1923).* The most important documentary study was compiled by Rimsky-Korsakov's son Andrey, *HA. Римский-Корсаков, жизнь и творчество,* 5 fasc. *(1933–46).* A useful compendium is A. SOLOVTZOV, *Жизнь и творчество Н.А. Римского-Корсакова (1964).* See also GERALD ABRAHAM, *Rimsky-Korsakov: A Short Biography (1948).*

(N.Sl.)

Rio de Janeiro (State)

Rio de Janeiro is one of the states (*estados*) of the republic of Brazil. Located in the southeastern region of the country, it is bounded by the states of Espirito Santo and Minas Gerais (to the north) and São Paulo (to the west); to the south lies the Atlantic Ocean. The state's name is derived from the city of Rio de Janeiro, which exerted a strong influence on its formation. The state has an area of *17,092* square miles *(44,268* square kilometres), of which *300* square miles consist of coastal lagoons and other internal waters. At the time of the *1980* census it had a population of almost *11,500,000* and a population density of more than *675* persons per square mile *(260* persons per square kilometre). The capital is the city of Rio de Janeiro, which had a population in 1980 of about 5,100,000.

History. The history of the state is enmeshed with that of the city of Rio de Janeiro, its chief economic and political centre from the early 16th century until 1834, when the city first became a separate entity. In 1835 Niteroi became the capital of the province of Rio de Janeiro. In 1889, when the Brazilian republic was proclaimed, the province became a state, and in 1890 Teresopolis became the capital; in 1902, however, the seat of government returned to Niteroi. When the capital of Brazil was moved to the newly established city of Brasilia in 1960, the territory that formerly had been the Federal District became the new Guanabara state, which existed as an enclave within Rio de Janeiro state. In 1975 the two states were merged into the reorganized State of Rio De Janeiro. The city of Rio de Janeiro was then made the capital once again.

From the time of its temtorial formation, the province depended on sugar production as the basis of its economy. By the end of the 18th century it had about 168,000 inhabitants, almost half of whom were slaves who were working. either in agriculture or in one of 600 sugar mills in the province. During the 19th century coffee replaced sugar as the most commercially significant crop, enriching the landowners of the Paraiba Valley, who constituted the ruling group of the Brazilian Empire until the abolition of slavery in 1888 and the proclamation of the republic in 1889.

The three regions

Relief and environment. The state's relief has three distinct features: the plain, or coastal lowland; the mountainous highland; and the plateau of the interior. The coastal lowland—broken by occasional massifs or rocks that sometimes extend far into the sea, as happens at Cabo Frio and Saquarema and along the stretch of coast opposite the offshore island of Ilha Grande—is narrower to the west, where the mountains of the Serra do Mar compress it against the sea. The prevailing climate is hot and humid and is characterized by summer showers. In winter the climate is modified by cold air masses from the south. The average temperature is generally above 72" F (22" C).

The mountainous highland comprises part of the Serra do Mar and, farther inland, part of the Serra da Mantiqueira, both of which run parallel to the coast in a roughly southwest-to-northeast direction. Some important tourist and holiday resorts—Petropolis, Teresopolis, and Nova Friburgo—are located in this region, which is characterized by mild temperatures that average below 68° F (20° C) because of the high altitude. The highest summits are found in a range called the Serra dos Orgãos and in the Massif of Itatiaia, where the Pico das Agulhas Negras reaches a height of 9,134 feet (2,784 metres).

The most important area in the landscape of the plateau is the Rio Paraiba do Sul. In its valleys coffee plantations were first developed in the 19th century. The temperature remains mild at the highest altitudes but grows progressively hotter toward the bed of the Rio Paraiba do Sul, which flows northeastward before turning eastward to drain into the Atlantic in the northeastern part of the state. At the level of the river itself the climate is tropical, and the temperatures are high.

Destruction of the humid tropical forest that originally covered the territory of the present state of Rio de Janeiro began in the 16th century with the introduction of the *queimadas* (slash-and-burn) technique, used by Indians and European settlers alike to clear tracts of land for temporary cultivation. The clearing of the forest cover continued with the cultivation of sugarcane on plantations, was intensified with the expansion of coffee growing, and was completed with the progress of urbanization. In the mid-20th century the Brazilian government undertook reforestation of large areas on the hillsides and massifs, granting the areas protection as national parks and thus helping to preserve remains of the original forest. Part of the area, commonly known as the Forest of Tijuca (8,200 acres [3,300 hectares]), was established by the federal government as a nature reserve in 1961 to help preserve both vegetation and animal life. About 74,000 acres of the Serra dos Órgãos have been made a national park, and Itatiaia National Park was created on the Itatiaia massif in the Mantiqueira chain. Trees represented in the parks include jequitibas, perobas, and imbuias, as well as ferns and palm trees.

Apart from these public parks some patches of forest vegetation still survive on a few hillsides near the city of Rio de Janeiro, but these are disappearing as the urbanized area is gradually enlarged. On the Santa Cruz, Campo Grande, and Jacarepagua plains, grassland prevails, whereas on the muddy coastland red, yellow, and white mangroves grow.

Animals continue to inhabit the forests, but man's predatory activities in earlier periods and the continuing extension of the urban perimeter have forced the animals to flee to the thickest parts of the forest reserves or to a few natural shelters. Still found in these areas are the black-tailed, soft-furred nail monkey, or mico; a type of squirrel called *caxinguelê*; a bird called *macuco*; and the *jaguatirica*, as well as more common animals as parrots, anteaters, armadillos, raccoons, bush dogs, opossums, and snakes.

Government and political institutions. The State of Rio de Janeiro, according to the terms of its 1967 constitution, is an "integrated and inseparable part of the Brazilian Republic" and is entitled, in its territory, to all the power not bestowed on the union (federal government) by the Brazilian constitution. On March 15, 1975, the former State of Guanabara was merged with the surrounding State of Rio de Janeiro, and the new state was named Rio de Janeiro. It has its own symbols, as well as a flag, a hymn, and a coat of arms.

The government of the state consists of legislative, executive, and judiciary powers, the first being held by the state Legislative Assembly, the second by the state governor and his secretaries, and the third by the law courts and judges. The Legislative Assembly is composed of at least 55 deputies of Brazilian nationality and of legal age, who are elected by direct and secret ballot for a four-year mandate. The number of deputies increases as the population grows, at a ratio of one deputy per 100,000 inhabitants or a fraction greater than 50,000. The assembly meets at the city of Rio de Janeiro during the period from March 1 to November 30 each year, with a break in July. Since 1980 the state governor has been elected by direct popular vote. The governor's term also lasts four years. He must be Brazilian by birth and must meet certain other requirements.

The judiciary

The state judiciary power is exercised by a series of institutions—the Tribunal de Justiça (lawcourts); the Tribunal de Alçada (jurisdiction courts); the Conselho da Magistratura (Council of Magistrates); the Corregedoria da Justiça (Magistracy of Justice), as well as judges and appeal courts, and such courts as may be created by law. There is also a Tribunal do Juri (jury court), and tribunals of military justice.

According to the constitution, 22 percent of the total expenditure of the state budget, ratified in the previous session, must be reserved for educational and cultural activities. Teaching of all types and the granting of degrees are the responsibility of the state school system. Private institutions, however, also have freedom to teach. The state university of Rio de Janeiro, which is organized as a foundation, is responsible for education at the undergraduate level and for the promotion of artistic culture. It is autonomous and receives an annual grant of not less than 15 percent of the state's total expenditure on education and culture.

As stipulated by the Brazilian constitution, the state promotes the development of industries, especially of those enterprises in which the greater part of the capital is invested by Brazilian nationals. Factories operating in the urban zone or in other Brazilian states benefit if they transfer to areas zoned for industry in the rural and suburban areas of the state.

Population and administration. The state has one of the highest urbanization rates in Brazil, with more than 80 percent of the population living in the city of Rio de

Janeiro and other urban centres. About 90 percent of the state's inhabitants are Roman Catholic, about 5 percent Protestant, and almost 3 percent Spiritists (believers in spiritualism).

The State of Rio de Janeiro is subdivided into 5 major regions and 14 minor regions with 64 municipalities. In 1980 more than four-fifths of the population of the state lived in the 12 municipalities that had more than 100,000 inhabitants.

Health and Education

Social conditions. Medical facilities in the late 1970s were provided by almost 550 hospitals, which together had about 70,600 beds; there were more than 13,100 doctors. The educational system included some 6,075 primary schools, 1,390 secondary schools, and nine universities; there also were about 100 independent schools that provided higher education.

The economy. About 5 percent of the working population of the state is engaged in agriculture, 25 percent in manufacturing, and 55 percent in service industries. Agriculture, however, accounts for only about 16 percent of the state's income, whereas manufacturing produces about 33 percent, and service industries a little more than 50 percent. The principal industries are metallurgy, printing, shipbuilding, and oil refining, and manufactured products include textiles, foodstuffs, and chemicals. Cement manufacturing, sugar refining, and automobile production are other important economic activities. Agricultural products include sugarcane, oranges, and bananas.

Transportation. More than 12,700 miles of roads, of which about 1,860 miles are paved, cross the state. The Central do Brasil and the Leopoldina railroads link the state with Brazil's national rail network. The Rio–Niteroi Bridge, which is about 9 miles long, connects the city of Rio de Janeiro with Niteroi, located on the east side of the bay. Ferry and hydrofoil service also link the two cities. The state has two airports: Santos Dumont, on Guanabara Bay within the city of Rio; and Galeão, on Governador Island in the bay, which was opened in January 1977 to accommodate international as well as domestic fights.

Cultural life. In the early 1980s the State of Rio de Janeiro had more than 60 museums, almost 300 cinemas and theatres, and some 30 broadcasting stations. There were more than 60 newspapers published in the state, of which 17 were dailies.

BIBLIOGRAPHY. JOSÉ DE SOUSA AZEVEDO PIZARRO E ARAUJO, *Memórias Histdricas do Rio de Janeiro,* 7 vol. (1948); and ALBERTO RIBEIRO LAMEGO, *O Homem e o Brejo* (1945), are two sources for the study of the historical development of the State of Rio de Janeiro. SYLVIO FROES ABREU, *O Distrito Federal e Seus Recursos Naturais* (1957); and ANTONIO TEIXEIRA GUERRA, "Paisagens físicas da Guanabara," *Revista Brasileira de Geografia,* 27:539–568 (1965), describe the state's environment, natural resources, and landscape. The INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATISTICA, *Enciclopédia dos Municípios Brasileiros,* vol. 22–23 (1959–60), gives a comprehensive account of the development of the State of Rio de Janeiro up to the 1960s.

(A.P.G.)

Rio de Janeiro (City)

Rio de Janeiro, a major port city of Brazil and capital of the State of Rio de Janeiro, is located on the Atlantic Ocean, in the southeastern part of the tropical zone in South America. The name was given to the city's original site by Portuguese navigators who arrived at the port on January 1, 1502, and mistook the entrance of the bay for the mouth of a river (***rio*** is the Portuguese word for river and ***janeiro,*** the word for January). When the foundations of the future town were laid in 1565, it was named Cidade de São Sebastião do Rio de Janeiro (City of Saint Sebastian of Rio de Janeiro), for both St. Sebastian and Dom Sebastian, king of Portugal. It is officially and commonly called Rio de Janeiro, but is often referred to, in a shortened form, as Rio.

When the capital of the Brazilian republic was transferred to Brasilia in 1960, the territory belonging to the former Federal District was converted into Guanabara state, which formed an enclave in Rio de Janeiro state. In March 1975 the two states were fused as the State of Rio de Janeiro; Guanabara and the city of Rio de Janeiro (former Guanabara state) became one of the 14 ***municípios*** (municipal districts) of the Metropolitan Region of Rio de Janeiro, or Greater Rio. The city of Rio de Janeiro became the capital of the reorganized State of Rio de Janeiro.

The city of Rio de Janeiro occupies an area of 452 square miles (1,171 square kilometres). It lies on an inlet of the Atlantic, the Baia de Guanabara (Bay of Guanabara), the entrance to which is marked by a point of land called Pão de Açúcar (Sugar Loaf). The core of the city lies on the western shore of the Baia de Guanabara on a plain, composed of marine and continental sediments, that is interrupted by several rocky mountains. At the time of the 1980 census the population of the city of Rio de Janeiro was about 5,100,000.

HISTORY

The city of Rio de Janeiro, surrounded by the sea and the mountains, is an urban agglomeration with exceptional features. The site was won from nature in a long succession of battles, which resulted in the transformation of inhospitable areas into a remarkable landscape, slowly and progressively adapted to the demands of human occupation.

Early settlement

Several years after the Portuguese first explored Brazil, French traders in search of ***pau-brasil*** (a type of brazilwood) explored the rich area extending from the Cabo Frio (Cold Cape) coast to the beaches and islands of the Baia de Guanabara—the economic and, above all, strategic importance of which was already well known. On one of these islands, which now bears his name, a French Huguenot, Nicolas Durand de Villegaignon, founded a colony that was called La France Antarctique (Antarctic France).

The Portuguese wanted to expel the French from Brazil; and the task was given to Estacio de Sa, the nephew of Governor Mem de Sa of Brazil, who in 1565 occupied the plain between the Morro Cara de Cão (Dog Face Hill), and the Sugar Loaf and Urca mounts, thus laying the foundations of the future town of Rio de Janeiro. After two years (1565–67) of bloody battles, in which Estacio de Sá was killed, Mem de Sa chose a new site for the town, further inland on the coast of the bay, at the top of the Morro do Descanso (Hill of Rest), or São Januario (St. Januarius Hill), later called the Morro do Castelo (Castle Hill). In 1568 the settlement was laid out in the form of a medieval citadel, protected by a bulwark and cannons.

As the settlement slowly developed, spreading around the feet of the hills and along the beach, the local marsh did not hinder expansion. The surrounding fertile land, allotted to Portuguese settlers by the Portuguese king in enormous plots called ***sesmarias,*** was planted with sugarcane, which was to provide the colony with its main source of income.

In 1660 the community became the seat of the government of the southern captaincies (Portuguese administrative units, each governed by a captain) of Brazil. In the second half of the 17th century, the captaincy population grew from 7,000 to 8,000 inhabitants, two-thirds of whom were probably Indian or Negro slaves.

At the beginning of the 18th century, Brazil began to engage in mining, which brought about remarkable changes in the colony's economic life. Only three decades after the discovery of gold, diamonds were discovered, resulting in a great migration, both from Brazil itself and from abroad, to the colony's hinterland. As a result the former village became a town and increased its population from 12,000 in 1740 to 24,000 in 1749.

Because of its rapid expansion, quick solutions had to be found for its pressing new problems. The construction of the Carioca aqueduct solved the most important problem—the water supply; this monumental aqueduct brought water from the Rio Carioca to a public fountain in the town. When the colonial capital was transferred from Bahia to Rio de Janeiro in 1763, the resulting demographic growth made the town expand further, far beyond its walls.

Monument to Christ the Redeemer on Morro do Corcovado, overlooking Rio de Janeiro. Baia de Guanabara, and Pão de Açúcar.
Kurt Scholz—Shostal Assoc

At the end of the 18th century, the town's economy, as well as that of the colony as a whole, was in a crisis because of the decline of the mines, as a result of which thousands of gold seekers became vagrants. In 1796 the value of exports from Rio's port was less than half of what it had been in 1760.

Colonial improvements

When the Portuguese royal family resettled in Brazil in 1808, the colony prospered again. By 1815, when Brazil became a kingdom, Rio de Janeiro was large enough to accommodate a foreign population. At about this time the city's initial features were being transformed; from 1808 to 1818, 600 houses and 100 country houses were built, and many older buildings were restored. Many streets were lighted and paved, more land was reclaimed, new roads opened, and new public fountains installed. Among new institutions established were the Royal Press, the Royal Library, the Theatre of Saint John, the Academy of Fine Arts, the Botanical Gardens, and the Bank of Brazil. When King John VI returned to Portugal in 1821, Rio had almost 113,000 inhabitants and 13,500 buildings, and the town had extended both northward and southward.

After Brazil won its independence in 1822, the expansion of coffee plantations in the State of Rio de Janeiro gave a new impulse to the city's development. Nobles and rich bourgeois moved their residences north to the São Cristóvão district. Merchants and English bankers chose to live around the Outeiro da Gloria (Gloria Hill) and Praia do Flamengo (Flamengo Beach) areas in the east, or they established their residences in the nearby Botafogo and Laranjeiras districts. The French, on the other hand, lived in country houses scattered in the Tijuca area farther westward.

In this era the city changed its appearance, and the traces of its colonial past were effaced. In 1829 oxcart traffic was banned from the Rua do Ouvidor, then the city's most elegant highway. In 1838 the first public transportation—horse-drawn buses—began to run to the districts of São Cristóvão, Engenho Velho, and Botafogo. In 1868 the first tramcars, also drawn by animals, were introduced. A steamboat service to Niteroi, on the eastern shore of Baia de Guanabara, began to operate in 1835. The first railroad was built in 1852 to Petropolis to the northeast, and a line reached Queimados to the northwest in 1858. In 1854 gas replaced oil for street-lighting, and wireless telegraphy was inaugurated. Sewerage was installed in 1864, and telephone service began in 1877.

When Rio de Janeiro, which had formerly been the capital of the empire, became capital of the republic of Brazil in 1889, it was already a considerable community. At the time of the 1890 census, it had more than 520,000 inhabitants. In both its population and its urban area (61 square miles), the city of Rio de Janeiro ranked as the largest city in Brazil and as one of the large cities in the world. The 1891 constitution gave it the status and name of the Federal District.

During the federal administration of President Francisco de Paula Rodrigues Alves, from 1902 to 1906, the Federal District was transformed into a modern city. As a result of the work of a team of administrators and technicians, endemic yellow fever and smallpox were subdued, the death rate decreased, health conditions were improved, huge swamps were drained, slums were cleared, and more streets were paved and widened. Important changes were also made in the city's plan: the central avenue (called Avenida Rio Branco from 1912), running northeast to southwest, was opened during this period; Avenida Beira-Mar, running parallel to part of the south shore, was built; and several other important avenues were opened.

Changes in the early 20th century

During the first three or four decades of the 20th century the city underwent no striking changes. The population of the Federal District exceeded 1,000,000 by 1920 and increased to 1,750,000 inhabitants by 1940. The proportion of foreigners decreased from 30 percent in 1890 to 13 percent in 1940. Migration continued to be the main factor in demographic growth, supplying the city with a ceaseless influx of people born in other parts of Brazil. The death rates were lower than the birth rates, and they continued to decline. Industry also developed during this period, industrial establishments in the Federal District increasing from about 1,500 in 1920 to more than 4,000 in 1940.

NATURAL SETTING

Rio de Janeiro lies on a strip of Brazil's coast and faces south. The greater portion of the city—embracing the centre and the North Zone—lies on the western shore of Guanabara Bay and is cut off from the South Zone by coastal mountains. The peaks, ridges, and hills of Rio are offshoots of the Serra do Mar, an ancient gneiss-granite mountain chain.

Although the region's climate is generally tropical, hot, and humid, the climate of Greater Rio is strongly affected by its topography and its proximity to the ocean. Along the coast, the breeze, blowing alternately onshore and offshore, modifies the temperature. Because of its geographic situation, the city is reached by south and the southeast winds, blowing from the Antarctic, which carry large amounts of moisture and cause frequent weather changes. During certain periods of the year these changes are accompanied by strong showers that sometimes provoke catastrophic floods and landslides. The hilly areas register greater rainfall since they constitute a barrier to the humid wind that comes from the Atlantic. The highest rainfall rates are found in the urban districts of Jardim Botânico (more than 63 inches [1,600 millimetres]), Urca, Deodoro, Paqueta, Bangu, and Santa Cruz (more than 47 inches).

The temperature varies according to the altitude, the distance from the coast, the type of vegetation, and the season. Winter (from June 21 to September 23) is particularly pleasant, both because of its mild temperature and because it is, in general, less rainy than the summer (December 21 to March 21), which is both hotter and wetter. The average temperature at Rio is about 73° F (23° C).

THE CONTEMPORARY CITY

Greater Rio. The demographic and economic expansion that took place in Rio during the decades of the mid-20th century transformed the city into the country's second regional metropolis, the first being São Paulo. Many neighbouring settlements have become integrated into its area of influence, since they depend economically and socially on the central nucleus. These local communities, most of which became part of Rio's conurbation after World War II, have changed from commuter areas to suburbs that are firmly integrated in the industrial, commercial, and cultural activities of the metropolis.

Fourteen *municípios* comprise Greater Rio. Five of these municipios—Nova Iguaçu, São Gonçalo, Duque de Caxias, Niteroi, and São João de Meriti—are themselves large urban agglomerations. Petropolis, situated in the Serra do Mar at an altitude of 2,900 feet, is an important tourist resort. The remaining *municípios*—Itaboraí, Itaguai, Mage, Mangaratiba, Maricá, Nilópolis, and Paracambi—are smaller localities scattered in the urban periphery. Together with the municipio of Rio de Janeiro, they form the area that is called Greater Rio. In the 1980 census the total metropolitan population of Greater Rio was almost 9,100,000 people.

Demography. The municipio of Rio de Janeiro, or the city of Rio, had 5,100,000 inhabitants by 1980. It is entirely urbanized and is divided into 24 administrative regions: Portuaria, Centro, Rio Comprido, Botafogo, Copacabana, Lagoa, São Cristóvão, Tijuca, Vila Isabel, Ramos, Penha, Meier, Engenho Novo, Iraja, Madureira, Jacarepagua, Bangu, Campo Grande, Santa Cruz, Ilha do Governador, Ilha de Paquetá, Anchieta, Santa Teresa, and Barra da Tijuca.

Between 1940 and 1960 the city's population expanded at a rate of more than 3 percent a year. After 1960, when the capital of Brazil was transferred to Brasilia, development stabilized in the city of Rio de Janeiro, and the population growth rate fell to 2.3 percent annually. The population density in 1980 reached 11,268 persons per square mile (4,349 persons per square kilometre), with a maximum of 88,126 inhabitants per square mile in the crowded administrative region of Copacabana. At the other extreme, in the scantily populated administrative region of Santa Cruz, a former rural zone, the density was 2,097 per square mile.

The city of Rio de Janeiro

The following were the 12 most populated administrative regions at the time of the 1980 census: Bangu (530,378 inhabitants); Méier (411,343); Anchieta (337,873); Campo Grande (333,941); Jacarepagua (326,855); Penha (315,-837); Madureira (277,537); Iraja (273,281); Botafogo (268,047); Ramos (254,952); Copacabana (228,703); and Lagoa (218,002). The administrative regions of the North Zone are more heavily populated than those of the South Zone.

The population of Rio de Janeiro has always grown primarily as a result of migration, which in some years has accounted for two-thirds of its increase. When the flow of immigrants from abroad began to decline (between the years 1890 and 1960 the proportion of foreigners in the former Federal District decreased from 30 percent to 7 percent), the number of migrants coming from other parts of Brazil increased. Nearly 45 percent of the population of Rio de Janeiro in the late 1970s were migrants from other places in the country. Most migrants were born in the states of Rio de Janeiro, Minas Gerais, or Espirito Santo. Among the foreign-born in Greater Rio in the early 1980s, the largest groups were Portuguese, Italians, and Spaniards.

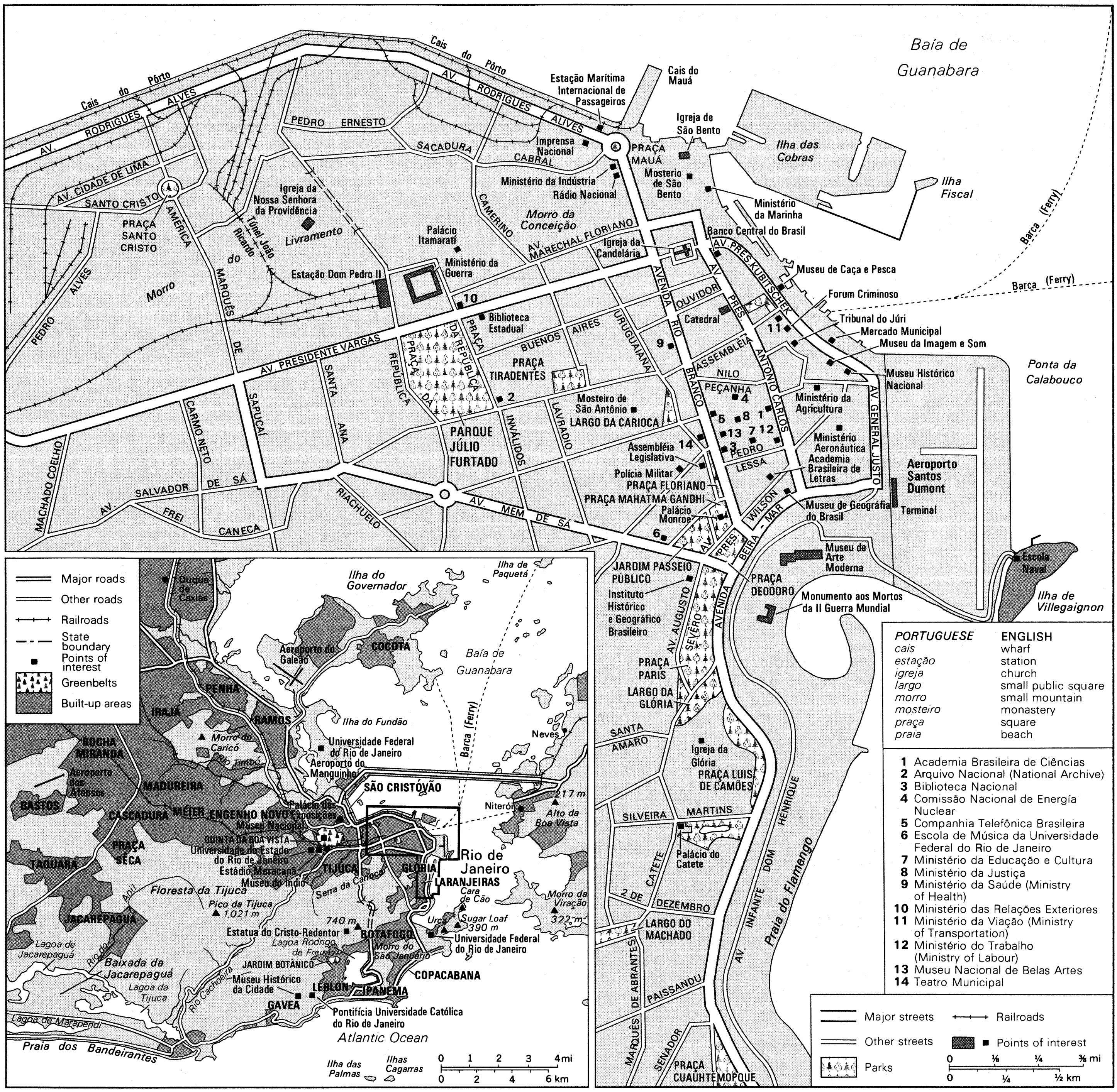
As in other Brazilian urban centres, the inhabitants of the city of Rio de Janeiro are mainly Christian, with about 89 percent Roman Catholic and 4 percent Protestant; Spiritists represent 4 percent and Jews about 1 percent of the population. Most of the people of Rio de Janeiro are white (70 percent), although the proportion of blacks (11 percent) and mestizos (18 percent) is also significant.

Town planning. A great effort of urban renewal in Rio de Janeiro, beginning about 1920, has had encouraging results. Rio is no longer the "soaked sponge," as it was called by a historian referring to the great number of existing lagoons and swamps that characterized it in its earlier periods. After the first decades of the 20th century it became a comfortable and pleasant city and was able to renew, in a short period of time, not only its natural landscape but also the main part of its architectural heritage.

In less than 40 years the urban centre of Rio de Janeiro and some South Zone districts were almost entirely demolished and reconstructed; in some areas older two- or three-storied houses were replaced by skyscrapers that ranged from 10 to 30 floors. Some of these buildings are the product of remarkable architectural concepts and have made the names of such architects as Affonso Eduardo Reidy, Lúcio Costa, and Oscar Niemeyer internationally known.

Urban growth

The accelerating urban development of the city of Rio de Janeiro has attracted many low-income groups from the other parts of the country and especially from rural areas. This is the origin of the rapid growth in the number of barracos ("huts"), or hovels, which are frequently clustered together in agglomerations. Some estimates have put the number of favelas ("shantytowns") at more than 200 and have estimated their total population at about



Central Rio de Janeiro and (inset) its metropolitan area

1,000,000 people. Census data, however, indicate somewhat lower figures—between 1950 and 1960 the census reported that the number of *favelas* increased from 58 to 147, rising to 165 by 1970; their population was reported to have grown from 165,000 in 1950 to 337,000 in 1960 and to 565,000 in 1970.

In the late 1960s, agencies jointly administered by the federal and the state governments began to encourage the purchase of houses and flats that were built for sale in long-term installments to poor families living in the *favela* ~This policy led to the disappearance of some of these agglomerations of substandard dwellings, but it did not prevent the emergence of similar nuclei of huts that, sooner or later, threatened to develop into new shantytowns. At the time of the 1980 census it was reported that there were 192 *favelas* in Greater Rio, with a total population of 628,000.

Economic life. Nearly 75 percent of the internal income of Greater Rio is produced by the service sector of the economy. The contribution of industry amounts to approximately 25 percent.

In 1977 there were almost 10,300 industrial establishments employing some 467,800 people. The most important classes of industry are metallurgy (53,500 employed), machinery (41,000), wearing apparel and footwear (41,700), textiles (35,200), and nonmetallic mineral products (32,100).

The state government and the financial institutions under its control developed, in the early 1970s, a plan to finance and to stimulate industrial growth; the federal government also provided incentives. In accordance with this policy of industrialization, the Metropolitan Region was divided into five industrial sectors (Fazenda Botafogo, Palmares, Paciência, Santa Cruz, and Campo Grande), which together account for about 20,000,000 acres (8,094,000 hectares) for sale to industrial enterprises.

As an important port and trading centre, Rio de Janeiro each year receives and distributes considerable amounts of foreign and national products throughout its zone of economic influence, which includes large regions in southern, northeastern, and southeastern Brazil. Its imports come from other parts of the Americas, from Europe, and from other continents. By value, exports in 1979 totalled U.S. $2,536,000,000, and imports totalled U.S. $1,363,000,000; exports normally greatly exceed imports. From Rio's port almost 5,000,000 tons of commodities are distributed each year to various points in Brazil; this is in addition to a great volume of internal trade by railway and road.

Banks and financial institutions

As one of the most important financial centres of Brazil, the Metropolitan Region of Rio de Janeiro has more than 850 bank branches, 80 percent of which are private. A significant number of banks have their headquarters in the city of Rio de Janeiro. Its stock market is one of the most important in the country. The securities market, which expanded rapidly during the 1970s and early 1980s, is supported by powerful financial corporations that have their headquarters in Rio.

In Greater Rio, which has one of the highest per capita incomes in Brazil, the retail trade is substantial. Although many of the most important establishments are located on the main streets of the city centre, others are scattered throughout the commercial areas of the northern and southern districts, where branches of shops, department stores, warehouses, supermarkets, and other retail businesses have been established.

The building industry is most active and performs an important role in the economy of Rio de Janeiro; it is a major consumer of manufactured products, as well as being a source of employment for large contingents of unskilled labor.

Government and political institutions. The *município* of Rio de Janeiro is governed by a ***prefeito*** (mayor) with the assistance of seven administrative departments (administration, social development, education and culture, financial, public works, planning, and health). The mayor is appointed by the governor of the State of Rio de Janeiro. Municipal legislative power is held by the members of the Municipal Chamber. The assembly meets in the city from March to November each year, with a break in July.

Transportation. Greater Rio is crossed by 700 miles of paved roads, through which it is linked with major cities in Brazil, as well as cities in other countries throughout South America. Nearly 200 road transport enterprises, with equipment totalling more than 6,000 buses and 2,000 trucks, carry some 900,000,000 passengers and 3,000,000 tons of goods a year. The traditional railroads, the Central do Brasil and the Leopoldina railways, both of which are part of the federal rail system, provide daily passenger train service, carrying some 118,000,000 passengers annually between the city centre and the main suburban and rural centres.

Rio has facilities for water traffic along the shores of the Baia de Guanabara, such as at Niteroi, as well as on the islands in the interior of the bay, such as at Paqueta; service is provided by ferry boats, motorboats, and hydrofoils. The construction of a subway was begun in 1972, and the first station was opened in 1979. Traffic on the subway has grown as each new station was opened, and by the early 1980s the subway transported some 250,000 passengers per day. Surface urban traffic is becoming more dense and more complex. Congestion in the city and its environs is increasing as the number of automobiles grows. By the early 1980s there were more than 900,000 automobiles in Greater Rio.

Two modern airports—Galeão for domestic and international services and Santos Dumont for domestic lines only—make Rio the primary centre for air services in Brazil.

The port of Rio de Janeiro is one of the largest in the country, and it is the first in the amount of freight carried (almost 30,000,000 tons in 1980). The port has 24,250 feet of quay.

Municipal services. ***The water and sewer systems.*** As frequently happens in rapidly developing metropolitan areas, the city of Rio de Janeiro throughout its history has confronted serious difficulties in providing urban services and facilities.

Since 1723, when Aires Saldanha, governor of Rio, built the Carioca aqueduct, Rio's water supply has continuously expanded. The historical aqueduct has now become a viaduct, crossed by old and picturesque streetcars. Rio de Janeiro now has a modern water supply furnished by the Companhia Estadual de Águas e Esgotos, a state company. By the late 1970s an average of 500,000,000 gallons of water per day was being supplied, and new reservoirs had been constructed to supply the area's needs. Proposals include the construction of additional reservoirs and the modernization of the entire water supply system.

Rio's first sewer system was constructed in 1864. The present system, like the water supply system, is operated by the Companhia Estadual de Águas e Esgotos. The sewer network is 4,287 miles long, and the number of the buildings served is estimated to be more than 200,000. A sewage disposal facility has been constructed in deep water several miles from the coast in order to avoid pollution of the beaches in the area.

Electricity. The capacity of Rio de Janeiro state's thermoelectric power installations is 662 megawatts, but the supply system is interlinked with a system of power stations among which Light—Serviços de Eletricidade S.A., a state enterprise, is predominant. The total installed power in the Rio area is 1,750 megawatts, of which 1,088 megawatts come from hydroelectric power stations. The consumption of electric power in the area is growing at a rate of more than 10 percent per year. By 1979 it had reached 16,500,000 megawatt-hours, of which 33 percent went to industry, 25 percent to residential uses, and 17 percent to commerce. The remaining 25 percent was allocated to public lighting, electrified transportation, and other purposes.

Health and security. Health services are provided by almost 200 hospitals. About two-thirds are private, and the remainder are run by the state. They provide about 40,000 beds. The staffs of the hospitals are composed of about 15,000 medical doctors and other licensed personnel.

Military agencies

A great variety of civilian and military agencies are responsible for internal security. The city of Rio de Janeiro is the seat of the 1st Military Region, which comprises the army elite forces; of the 1st Naval District, as well as of naval technical, industrial, and training organizations; and of the Third Zone Air Command and other Brazilian air force organizations. The fire brigade of the city has modern equipment and has a tradition of great efficiency. The civilian and the military police forces consist of more than 20,000 men. A total of more than 90,000 people in Rio are engaged in work concerned with the public security and national defense.

Educational and cultural life. The city of Rio de Janeiro has one of the highest literacy and education rates in Brazil. The literate population comprises approximately 90 percent of those age 10 or older. The national average was around 80 percent in 1978. In the late 1970s education at the primary level was provided by almost 1,800 school units, with a staff of almost 36,100 teachers and with 887,000 students. Education at the secondary level in 1978 consisted of 340 school units, almost 13,100 teachers, and 201,300 students.

The three most important universities in Rio de Janeiro are the Universidade Federal do Rio de Janeiro, the Universidade do Estado do Rio de Janeiro, and the Pontifícia Universidade Catolica do Rio de Janeiro. In addition, there are some 50 other university establishments. Some of these are grouped in a federation and offer a great variety of courses.

Museums

As the country's cultural capital, Rio de Janeiro has many prestigious artistic, literary, and scientific institutions. These include the Academia Brasileira de Letras (Brazilian Academy of Letters), the Instituto Histórico e Geografico Brasileiro, and the Academia Brasileira de Ciências (Brazilian Academy of Sciences), as well as 28 museums. Of the museums, the National Museum of Fine Arts, founded in 1818, is the oldest. It is located in the former Imperial Palace of the Quinta da Boa Vista. The Museum of the Republic and the Literary Museum are established in the older Catete Palace, where the presidency of the republic was installed during the period when Rio was the Federal District. Other museums of interest include those devoted to special interests such as the historical city, modern art, image and sound, and the Brazilian Indian people.

Greater Rio has some 100 cinemas and some 50 theatres. There are about 45 libraries, with estimated holdings of 5,000,000 volumes. The city of Rio de Janeiro has a municipal library and 14 regional libraries, all of which are maintained by the city government; their services include mobile libraries and a braille section. The most important library in the city, however, is the National Library; it was founded in 1810 with the remains of the Royal Library of Ajuda, which were brought to Brazil from Portugal after being saved from the fire that followed the 1755 earthquake in Lisbon. The National Library has its own building, in which are kept a total of some 3,500,000 volumes, 600,000 manuscripts, and 250,000 maps and prints.

There are more than 20 radio broadcasting stations and six television stations located in Rio de Janeiro. Many periodicals are published there, including some 14 daily newspapers.

Tourism

The majority of tourists (60 percent) who travel to Brazil visit the city of Rio de Janeiro. Among the picturesque places frequently toured are the Morro do Corcovado (Mount Corcovado), 2,310 feet (704 metres) high, on top of which is found the monument to Christ the Redeemer (Cristo-Redentor); Pão de Açúcar, 1,296 feet (395 metres) high, which affords a spectacular view of the South Zone of the city and which is reached by a funicular railway; the Quinta da Boa Vista, a park in which the National Museum and the Zoological Garden are located; and the Botanical Gardens (Jardim Botânico), which dates from 1808 and displays more than 60,000 species. The Tijuca National Park, located in the Forest of Tijuca, is crossed by a road that reaches several places of touristic interest; the road also links the South and the North zones of the city.

The city of Rio de Janeiro has a number of excellent first-class hotels. The chief hotels are located in the area along the beaches. Rio's well-known carnival, which lasts for four days each year, is a traditional festival in which the people of the city actively participate. The most popular sport in Rio de Janeiro, as in Brazil as a whole, is association football (soccer); the most important matches take place in the Municipal Stadium (Estadio Municipal do Maracanã). A monumental athletic ground that was inaugurated in 1950, the stadium can seat 200,000 spectators.

BIBLIOGRAPHY. JOSÉ VIEIRA FAZENDA, *Antiqualhas e Memórias do Rio de Janeiro* (1928); GASTAO CRULS, *Aparência do Rio de Janeiro: Notícia Histórica e Descritiva da Cidade, 2* vol. (1949); and VIVALDO COARACY, *O Rio de Janeiro no Século 17.* (1944), are the best accounts of the historical development of Rio de Janeiro. ALBERTO RIBEIRO LAMEGO, *O Homem e a Guanabara,* 2nd ed. (1964); SYLVIO FROES ABREU, *O Distrito Federal e Seus Recursos Naturais* (1957); and ANTONIO TEIXEIRA GUERRA, "Paisagens Fisicas da Guanabara," *Revista Brasileira de Geografia,* 27:539–568 (1965), study the environment, natural resources, and landscape of the city and its surroundings. The problem of slums and the living conditions of their inhabitants is treated by ALBERTO PASSOS GUIMARÃES, *As Favelas do Distrito Federal e o Recenseamento de 1950* (1953); and REMULO COELHO, *As Favelas do Estado da Guanabara, Segundo o Censo de 1960* (1966). ALBERTO PASSOS GUIMARÃES *et al., Enciclopédia dos Municipios Brasileiros,* vol. 23 (1960), gives a comprehensive account of the city's history and development up to the 1960s. Statistical data on population and other aspects of Rio de Janeiro are given in FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, *Sinopse Preliminar do Censo Demogrdfico: VIII Recenseamento Geral—1970* (1971); *Censo Demogrdfico de 1970* (1971); *Pesquisa Industrial: Brasil* (1977); *Anudrio Estatistico do Brasil* (annual); *Região Sudeste, Censo Demogrdfico de 1980 (Resultados Preliminares, Sinopse Preliminar do Censo Demográfico de 1980) Rio de Janeiro,* vol. 1; and *Anudrio Estatistico do Estado do Rio de Janeiro* (annual).

(A.P.G.)

Rio Grande

The fifth longest river of North America, the Rio Grande forms the entire border between the U.S. state of Texas and Mexico, in which country it is known as the Rio Bravo del Norte. The river's basin drains an area of about 172,000 square miles (445,000 square kilometres) and has a population of about 5,000,000 persons. The Rio Grande rises as a clear, snow-fed stream more than 12,000 feet (3,700 metres) high in the Rocky Mountains of Colorado and descends across steppes and deserts, watering rich agricultural regions on its way to the Gulf of Mexico.

Exploration and early use. Probably the first Europeans to see any part of the Rio Grande were those of an expedition sent out in 1519 to survey the coast of the Gulf of Mexico. The maps that illustrated this voyage, however, show only nameless indentations on a smooth coast for the mouths of rivers. The name Rio Bravo shows up for the first time on a map of 1536 compiled by a royal Spanish cartographer. In about 1536, the Spanish explorer Álvar Núñez Cabeza de Vaca and three companions crossed the Rio Grande in their travels, and Núñez later described their river crossings in his narrative *Naufragios* (1542). The expedition led by Francisco Vázquez de Coronado in 1540 to locate rumoured rich cities to the north of Mexico resulted in encounters with various Pueblo Indian communities and explorations in the Middle Rio Grande and upper Pecos areas.

The first thorough exploration of the basin of the Rio Grande, however, was made preliminary to the mining and agricultural settlements that were founded sporadically from the latter part of the 16th century into the 18th century. The earliest settlements were mining communities in the upper Conchos drainage in 1563; intermediate was the colonization of the Upper Rio Grande area in New Mexico in 1598; and the last colonization was begun in 1749 along the Lower Rio Grande. With the Mexican explorations of Gov. Juan Bautista de Anza in the San Luis Valley of modern Colorado in 1779, the exploration of the entire Rio Grande Basin was completed. Because most of the entire narratives of exploration and the corresponding maps remained unpublished in the various archives of the Spanish government, however, historians in the United States and Europe have tended to stress later but unpublished non-Spanish explorations such as those led by Zebulon M. Pike in 1807 and by John C. Frémont in 1848–49 in the Upper Rio Grande Area.

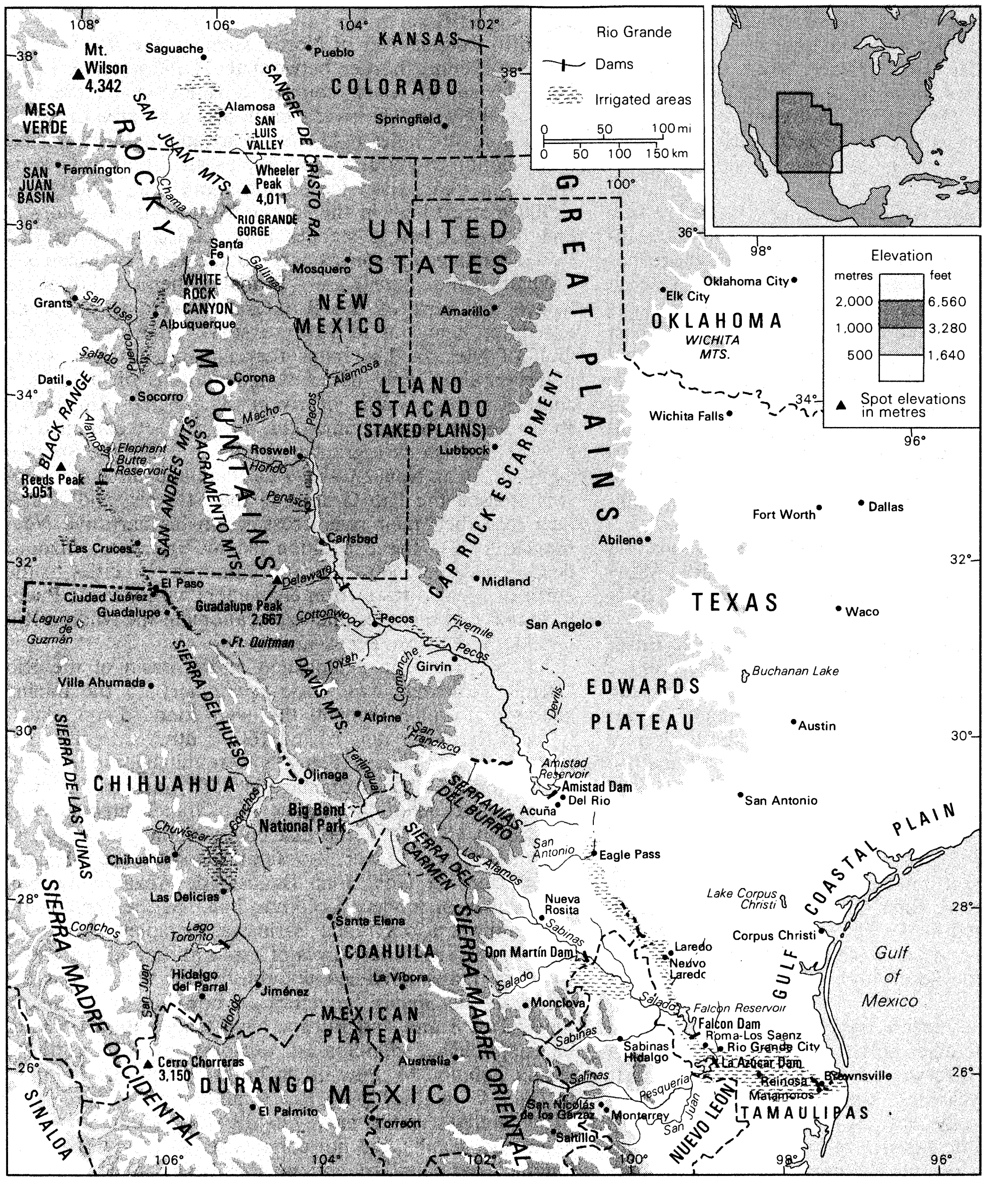
The careful scientific survey of the river, accompanied by good cartography, did not commence until the first of the international boundary commissions began its fieldwork in 1853, directed by a Mexican commissioner and surveyor and their counterparts from the United States. Small steamboats were used on the Lower Rio Grande up to Rio Grande City, and even to Roma when the river was high, from the 1850s until the great hurricane of 1874 swept the river clean of all man-made structures. Since then, accelerated erosion, silting, and sandbar formation have precluded navigation on the Rio Grande and have forced the United States and Mexico to spend much money and time in adjusting the boundary to the numerous changes in the river channel. On October 28, 1967, the United States formally returned to Mexico the Chamizal area, between El Paso and Ciudad Juarez, which a shift of the river in 1864 transferred to the left bank.

Erosion, silting, and changes in course

The river course. From its sources in the San Juan Mountains of southwestern Colorado, the Rio Grande flows southeast and south 175 miles (282 kilometres) in Colorado, southerly about 470 miles across New Mexico, and southeasterly between Texas and the Mexican states of Chihuahua, Coahuila, Nuevo Leon, and Tamaulipas for about 1,240 miles to the Gulf of Mexico. The total length of the river is approximately 1,885 miles.

Diversity of landscapes and environments

Its early course follows a canyon through forests of spruce, fir, and aspen into the broad San Luis Valley in Colorado, after which it cuts the Rio Grande Gorge and White Rock Canyon of northern New Mexico and enters the open terrain of the basin and range and the Mexican Highland physiographic provinces. There, declining elevation, decreasing latitude, and increasing aridity and temperature produce a transition from a cold steppe cli-



The Rio Grande Basin and its drainage network.

mate with a vegetation of piiion, juniper, and sagebrush to a hot steppe and desert climate characterized by mesquite, creosote bush, cactus, yucca, and other desert plants. Shortly before entering the Gulf Coastal Plain, the Rio Grande cuts three canyons between 1,500 and 1,700 feet in depth across the faulted area occupied by the "big bend," where the Texas side of the river comprises the Big Bend National Park. Along the remainder of its course the river wanders sluggishly across the Coastal Plain to end in a true delta in the Gulf of Mexico.

The principal tributaries of the Rio Grande are the Pecos, Devils, Chama, and Puerco rivers in the United States and the Conchos, Salado, and San Juan in Mexico. The peak of flow may occur in any month from April to October. In the upper reaches it usually is in May or June because of melting snow and occasional thunderstorms, whereas the lower portion commonly has its highest water in June or September because of summer rainstorms. It has been estimated that the Rio Grande has an average annual yield of more than 9,000,000 acre-feet (1,000,000 hectare metres), of which about one-third reached the Gulf before the building of the Falcon Dam, upstream from Rio Grande City, in 1953.

Human exploitation. Irrigation has been practiced in the Rio Grande Basin since prehistoric times, notably, among the ancestors of the Pueblo Indians of New Mexico. Increases in population and in the use of water made necessary the 1905–07 and 1944–45 water treaties between the United States and Mexico, as well as the Rio Grande Compact (1939) among Colorado, New Mexico, and Texas, concerning shared use of the waters of the Upper Rio Grande sub-basin (above the site of former Ft. Quitman, Texas), and the Pecos River Compact (1948) between New Mexico and Texas concerning the Pecos above Girvin, Texas. Essentially all of the average annual production of more than 3,000,000 acre-feet in the Upper Rio Grande (including the 60,000 acre-feet allotted to Mexico by treaty) is consumed within this sub-basin. Not only below Ft. Quitman but also in many stretches of the river from the New Mexico–Colorado border to below Brownsville, there has been no surface flow at various times. In some places the depth has varied from nearly 60 feet to a bare trickle or nothing. Below Ft. Quitman the Rio Grande is renewed by the Conchos and other Mexican rivers, which produce about two-thirds of the available water. A number of large springs in the area between Hot Springs in the Big Bend National Park and the town of Del Rio, Texas, including many in the bed of the river, are important and dependable producers of water.

Reservoirs and dams

The total storage capacity of the reservoirs in the basin is nearly 19,000,000 acre-feet, with a normal storage of more than 5,000,000 acre-feet, chiefly in the Falcon Reservoir on the Lower Rio Grande, Lago Toronto (La Boquilla Dam) on the Conchos, Elephant Butte on the Rio Grande in New Mexico, Marte Gómez (La Azúcar Dam) reservoir on the San Juan, and Venustiano Carranza (Don Martin Dam) on the Salado. The international Amistad Dam, below the confluence of Devils River, designed for a capacity of some 5,325,000 acre-feet, was completed in 1969 under terms of a U.S.–Mexico treaty. Considerable amounts of hydroelectricity are produced within the basin. More than 3,000,000 acres (1.200.000 hectares) are irrigated within the basin (about two-thirds of these in the United States). The leading crops raised by irrigation vary from potatoes and alfalfa in Colorado to cotton, citrus fruits, and vegetables in the valley of the deltaic Lower Rio Grande in Texas and Tamaulipas.

After agriculture and animal husbandry, the leading industries of the Rio Grande area are mining (petroleum, natural gas, coal, uranium ore. silver, lead, gold, potash, and gypsum), and recreation (national and state parks and monuments, dude ranches, fishing and hunting, summer and winter resorts). Urban communities include Monterrey, Ciudad Juárez, Chihuahua, Saltillo, Matamoros, Guadalupe, Nuevo Laredo, Reinosa, and San Nicolás de los Garzas in Mexico; Albuquerque, New Mexico; and El Paso, Laredo, and Brownsville, Texas.

BIBLIOGRAPHY. PAUL HORGAN, *Great River: The Rio Grande in North American History*, rev. ed., 2 vol. (1960), the best historical treatment available of the main river and the U.S. basin—Mexican portions are treated less adequately; LAURA GILPIN, *The Rio Grande, River of Destiny: An Interpretation of the River, the Land, and the People* (1949), a satisfying book of photographs and sketch maps, with a useful text; *Report of the President's Water Resources Policy Commission: Ten Rivers in America's Future*, vol. 2, pp. 285–351 (1950), a good physical description of the basin and a good map; NORRIS HUNDLEY, JR., *Dividing the Waters: A Century of Controversy Between the United States and Mexico* (1966), a detailed discussion of the Rio Grande to 1963, featuring political aspects.

(D.D.Br.)

Rio Grande do Norte

Rio Grande do Norte, a primarily agricultural and salt-producing state in northeastern Brazil and one of the smallest of all Brazilian states, is bounded by the Atlantic Ocean on the north and east, by the State of Ceará on the west, and by the State of Paraiba on the south. Its area is 20,469 square miles (53,015 square kilometres), and its population in 1970 was 1,611,606. The state capital is Natal, so named for the date of its official founding, Christmas Day (Natal), 1599.

In this northeastern region Brazil projects farthest eastward into the Atlantic, and largely for this reason an air base at Natal became an important stop on the Allied air route to and from Africa during World War II. The airport at Parnamirim, near Natal, is still important in serving airlines in the transatlantic trade.

Physical features

The coastal strip around Natal and southward is forested, and the salt marshes in this area are economically important (the saltworks produce about 95 percent of Brazil's salt). The northern seaboard, north of Natal, is low and sandy, with dunes and coconut palms; here the deep-sea fishermen common to the Northeast—the *jangadeiros*—ride the waves in rafts formed of lightweight tree trunks and powered by small triangular sails. From this narrow coastland in the north the land rises gradually to some low mesas (*taboleiros*). Inland from Natal and the south, the land rises abruptly to the northern edge of a hilly upland known as the Planalto da Borborema, which stretches southward into the States of Paraiba and Pernambuco. In the western interior of the state are several mountain ranges.

Except for the coastland from Natal southward, where the prevailing winds from the Atlantic bring abundant rainfall to support a forest, the state is semi-arid. The rainfall is usually so slight that the few rivers—mainly the Açu (Piranhas), the Apodi, and the Potengi—flow only intermittently. Average temperatures in Natal vary from 77" to 86° F (25" to 30° C), but higher elevations are cooler.

Much of the forests that formerly existed were sacrificed to sugarcane production, but there are vast areas of carnauba trees in the lowlands and coconut palms on the coast, as well as a scattering of various other tropical trees and flora. Animal life is scarce, because hunters have reduced or wiped out most of the native species. There are, however, some armadillos, deer, rhea birds, cariama birds, raccoon, skunks, foxes, and wildcats; in the ponds and lagoons are teal, ducks, jacanas, finfoots, bitterns, and herons, as well as alligators and numerous kinds of fish.

The territory was first settled by the Portuguese in the late 16th century. As early as 1534 the Portuguese crown had considered establishing fiefdoms, or captaincies, in the region, but not until 1598, after successfully repelling local French pirates, did the Portuguese succeed in establishing the Fort of the Three Wise Kings on the future site of Natal and laying the foundation of a government. From then until 1822, when Brazil proclaimed independence, the captaincy was ruled by a succession of Portuguese commanders and governors and, for a while (1633–54), by Dutch invaders. From 1822 to 1889, as part of the Brazilian Empire, the province of Rio Grande do Norte was governed by presidents; ever since, as a state in the republic, there have been a ruling governor and legislature.

Socio-economic conditions

In the colonial period the economy centred chiefly on sugar plantations, ruled by a few wealthy families and manned largely by slaves. Cotton, introduced in the 18th century, became, and remains today, the leading crop in an economy that is basically agricultural; sugarcane, though, is still grown. Other crops include corn, rice, manioc, millet, red beans, potatoes, coconut, and cassava. Cattle and horses are also raised. In addition to the saltworks, there is some mining in the Serra de Borborema, the tungsten mines being the most important in Brazil and constituting an important export. Other mineral products include gypsum, limestone, marble, monazite, gold, and beryl. Outside of Natal and such towns as Mossoró and Caicó, socio-economic conditions are very poor, the rural working families being mostly illiterate and living in hovels on very low incomes. In the towns a measure of industrialization has improved conditions somewhat; there are now factories producing textiles, clothes, oils, leather, furniture, food, tools, plastics, ceramics, paper, and cement. Schools, hospitals, housing, and other welfare items are being expanded. There are two universities, one at Natal and one at Mossoró.

There are three railroads—from the harbour in Areia Branca to the city of Sousa, Paraiba, from the capital to São Rafael, and from Natal to Recife, the capital of Pernambuco. There are several highways—from Natal to the south of the country, cutting through various states; from the north of the country to Ceará; and from the coast to the western interior.

BIBLIOGRAPHY. Few works are devoted strictly to Rio Grande do Norte, but a considerable number deal with the northeastern region of Brazil of which it forms part. Titles particularly recommended include LUIS DA CAMARA CASCUDO, *Histdria do Rio Grande do Norte* (1955), a detailed and well-documented study by an historian native to the area; GILBERTO FREYRE, *Casa-grande y Senzala*, 5th ed., 2 vol. (1943; Eng. trans., *The Masters and the Slaves*, 2nd ed., 1956), a classic work by an internationally esteemed scholar in the field of anthropology that describes a Brazilian family of the northeast as it existed in colonial times; and *Sobrados e Mucambos*, 2nd ed., 3 vol. (1951), a continuation of the preceding study, describing the transformation of rural patriarchal society as a result of growing urbanization; SERGIO BUARQUE DE HOLANDA, *História Geral da Civilização Brasileira*, vol. 1, *A Epoca Colonial* (1960), traces in detail the colonization and subsequent development of the northeast (as well as other areas) during the colonial period, written in a popular style; and ANTONIO DA SILVA MELLO, *Nordeste Brasileiro*, 2nd ed. (1964), a panoramic view of the complexities of the area, written by an outstanding scholar.

(J.E.F.P.)

Rio Grande do Sul

The southernmost state *(estado)* of Brazil, Rio Grande do Sul is bordered by the State of Santa Catarina on the north, Argentina on the west, Uruguay on the south, and the Atlantic Ocean on the east. It covers about 108,951 square miles (282,184 square kilometres), or 3 percent of the nation's total area. The population early in the 1970s was over 6,750,000. The capital of PBrto Alegre is the state's main industrial area and port. The state is a major agricultural and livestock region. Poor in mineral resources it also has little industrial development. (For a related article see BRAZIL.)

History. The state was originally thinly inhabited by Tupi-Guarani, Ge, and Guaicurú Indians. It was first explored and colonized by the Portuguese during the late 17th century. Long disputed between Spain and Portugal, the region was the site of intermittent warfare between 1754 and 1870. The state was also wracked by wars of secession during the 19th century. (For further discussion of the state's history see BRAZIL, HISTORY OF.)

Relief features

The environment. The north occupies part of the Paraná Plateau, which is composed of outpourings of basaltic lava solidified into sheets of rock known as diabase. The plateau stands between 2,000 and 3,000 feet (600 and 900 metres) above sea level. It has been dissected into rolling hills by streams, but its margins are marked by steep cliffs. Cliffs also cap the Serra Geral, the southern extremity of the Serra do Mar—the eastern edge of the Brazilian Highlands—along the Atlantic coast. North of PBrto Alegre, the cliffs turn westward and descend southward along the Rio Jacui Valley.

In the south, the Rio Jacui and its tributary, the Taquari, drain a lowland along the base of the plateau. South of the river, gently rolling hills stand between 1,000 and 1,500 feet in elevation. West of Livramento are tabular landforms of diabase; the Uruguay River cuts through the diabase and is broken by rapids. The coast is lined with sandbars and lagoons including the Lagoa dos Patos and Lagoa Mirim.

The climate is generaliy mild. In winter, cold air masses from the south bring heavy rains and occasional snow to the higher elevations. In summer the prevailing northeast winds bring less precipitation and hot weather, especially inland. Temperatures range from a minimum of 18° F (−8° C) to a maximum of 109" F (43" C), with an annual average of 68° F (20" C). Precipitation measures about 52 inches (1,300 millimetres) annually.

Most of the state is tall-grass prairie, with pine and tropical forests in the higher altitudes and deeper river valleys. Animal life includes deer, rodents, otters, armadillos, monkeys, and porcupines, but there are no large wild mammals. Of the abundant birdlife, the Brazilian lapwing is common. There are about 40 species of snakes, five of which are poisonous. Coastal and inland waters abound with fish (anchovies, dolphins, king fish, flat fish, mullet) and shrimp.

The population. The state's 6,755,000 inhabitants include descendants of the Portuguese in the Jacui Valley, Germans on the lower slopes of the Paraná Plateau and above the Jacui, and Italians on the plateau. There are also descendants of Polish and other European immigrants. About 6 percent of the population is composed of blacks and mulattos (persons of mixed white and black ancestry), and there is a small number of Indians in the northwest and Asians.

The main language is Portuguese. About 80 percent of the people are adherents of Roman Catholicism; other Christian denominations are also represented.

Administration and social conditions. As a federated state of Brazil, Rio Grande do Sul is governed by its own constitution, based on that of the federal government. Governmental powers are divided among the executive, legislative, and judicial branches. Local government is carried out through more than 230 municipalities.

Educational services include primary, secondary, and technical schools and several universities such as the Universidade Federal do Rio Grande do Sul (founded in 1934) and the Universidade Católica de Pelotas (1960). About 27 percent of the population is illiterate.

Role as the "Granary of Brazil"

The economy. The region was long known as the "Granary of Brazil." About 25 percent of Brazil's rice is raised on the floodplain of the Jacui and Taquari rivers. Wheat and maize (corn) are grown on the Paraná Plateau and the terraces above the Jacui. Other crops include grapes and tobacco. The southern plains serve as vast pastures for the state's livestock industry. Herds of cattle and sheep are tended by the gauchos, the herdsmen of the Llanos, whose animals feed across the vast unimproved pasturelands of the plains and produce mainly a lean, tough meat. Pigs are also raised.

Industrialization did not reach the south until the 1930s, and the state contributes only a small portion of the national industrial output. Industries are concentrated at PBrto Alegre, Rio Grande, and Pelotas. Coal is mined at São Jerônimo on the Rio Jacui and shipped downstream to PBrto Alegre; it is used to generate electricity and manufacture gas.

Transportation. The state is served by about 106,000 miles (171,000 kilometres) of roads. There are 2,500 miles (4,000 kilometres) of railways. The main line runs west from PBrto Alegre through Santa Maria and Alegrete to the Argentine border at Uruguaiana. Branch lines run north to São Paulo state, south to the Uruguay border, and southeast to the port of Rio Grande.

There are about 800 miles of inland waterways on the Jacui and Taquari rivers and the Lagoa dos Patos. There is also shipping along the state's 390-mile ocean coastline. The busy network of airports serves more than 30 cities.

Cultural life. Folkloric tradition centres upon the near-legendary courage, generosity, and romantic chivalry of the gaucho and his life under the wide skies of the plains. The Museu "Julio de Castilhos" in PBrto Alegre contains exhibits of national history and Indian artefacts, and the Museu Oceanogrbfico is located in Rio Grande. There are public libraries in Pelotas, PBrto Alegre, and Rio Grande.

BIBLIOGRAPHY. Enciclopédia Rio-Grandense, 5 vol. (1956–58); *Rio Grande do Sul, Terra e Povo,* 2nd ed. (1969); AZIZ NACIB AB'SABER and JEAN ROCHE, *Três Estudos Rio-Grandenses* (1966), cover geographic–economic aspects of the state. STUART CLARK ROTHWELL, *Ports and Hinterlands of Rio Grande do Sul State, Brazil* (1960); ARTHUR FERREIRA FILHO, *História Geral do Rio Grande do Sul, 1503–1957* (1958); and BALDUIN RAMBO, *A Fisionomía do Rio Grande do Sul,* 2nd ed. rev. (1956), deal with the geography of Rio Grande state. See also FRANKLIN DE OLIVEIRA, *Rio Grande do Sul: um Novo Nordeste,* 2nd ed. rev. (1961); and RIO GRANDE DO SUL, *Politica de Desenvolvimento Urbano* (1970).

(F.M.C.)

Ritschl, Albrecht

A German Lutheran theologian, Albrecht Ritschl initiated and, with his followers, led one of the most significant Christian movements in the years between 1875 and 1918, first in Germany and subsequently in the Anglo-Saxon world. Ritschl's interpretation of Christianity was influential because he was able to synthesize the teaching of the Scriptures and of the Protestant Reformation with some aspects of modern knowledge so as to show both the religious and ethical relevance of the Christian faith.

Early life and career

Ritschl was born in Berlin on March 25, 1822. His father and grandfather were Lutheran clergymen, and his lifelong interest in the Protestant Reformation was due, in part, to his family background. He completed both his secondary schooling and university theological studies with distinction. Ritschl was apparently well suited for his vocation intellectually, temperamentally, and by virtue of his education. In addition, his father had inculcated in him a strong sense of duty, the fear of 'God, and humility.

Ritschl was trained in theology and philosophy at the universities of Bonn (1839–41) and Halle (1841–43). He moved gradually from a more conservative, traditional theological position to a more liberal one. This shift was precipitated, in part, by the influence of the philosophy of G.W.F. Hegel upon most of Ritschl's generation. While a student Ritschl studied the classical theological disciplines and also developed a special interest in the doctrine of the redemptive work of Jesus Christ and, more specifically, the doctrine of the Atonement—that Christ died for man's sins. Other German philosophers, in addition to Hegel, influenced Ritschl: Immanuel Kant (1724–1804) and Friedrich Schleiermacher (1768–1834), the father of liberal Protestantism. After receiving his doctorate in 1843, Ritschl joined the ranks of the Tübin-

Ritschl.

By courtesy of the Niedersachsische Staats und Universitatsbibliothek, Gottingen, West Germany

gen school, a dominant theological movement involved in reconstructing the origins of Christianity and the early history of the church and its theology.

Works and significance

Ritschl taught at the University of Bonn (1846–64) and at Gottingen from 1864 until his death on March 20, 1889. His first significant publication, ***Die Entstehung der altkatholischen Kirche*** ("The Origin of the Old Catholic Church"; 1850, 2nd ed. 1857), revealed both his initial indebtedness to and gradual breach with the Tübingen school, which, in their analysis of the early history of Christianity, he found too indebted to Hegelian presuppositions.

In his interpretation of Christianity, Ritschl interacted with the Scriptures, the whole history of the church and its theology, and, especially, with the theology of Martin Luther and the confessions of the Reformation, as well as with the philosophical methodology and ethics of Kant and the theology of Schleiermacher. Virtually all of his research came to fulfillment in his major work, ***Die christliche Lehre von der Rechtfertigung und Versohnung (The Christian Doctrine of Justification and Reconciliation),*** published in three volumes (1870–74); it deals, respectively, with the historical and biblical materials (vols. 1–2) and with Ritschl's own reconstruction (vol. 3). Ritschl depicted Christianity as an ellipse with two focuses; namely, (1) the reconciliation established between God and man through the divine love manifested in the life and death of Jesus and (2) the Kingdom of God as the goal toward which all of God's activities are directed. Ritschl stressed the significance of the Kingdom of God in the teaching of Jesus and was critical of Protestantism's neglect of the Kingdom.

Over against Protestant Pietism, which emphasized the spiritual piety of the individual from a rather antiworldly perspective, Ritschl argued persuasively for the ethical development of man in his individual and corporate life. If man's forgiveness (justification) represents the religious pole of Christianity, reconciliation—which has to do with the practical consequences of justification—represents the ethical pole. Forgiveness of sin should give rise to a new life-style marked by doing God's will in both the personal and social spheres, thereby advancing the growth of the Kingdom of God (the brotherhood of all men united by the love of God) in society.

Ritschl was both esteemed and vilified by critics both during and after his lifetime. The dominant 20th-century Protestant movement known as Neo-orthodoxy, which restated Reformation principles such as justification and grace, attacked Ritschl as an advocate of cultural Christianity—the identification of Christianity with prevailing cultural norms. Contemporary reassessments of 19th-century theology are, however, more appreciative of Ritschl's interpretation of Christianity and the Kingdom of God and of their relevance to man's personal and corporate existence.

BIBLIOGRAPHY. OTTO RITSCHL, ***Albrecht Ritschls Leben,*** 2 vol. (1892–96), is the definitive biography by Ritschl's son; contains a complete bibliography of Albrecht Ritschl's writings from 1842–89. English translations of these writings include: ***A Critical History of the Christian Doctrine of Justification and Reconciliation,*** trans. *by* JOHN S. BLACK (1872); ***The Christian Doctrine of Justification and Reconciliation: The Positive Development of the Doctrine,*** 2nd ed., trans. by H.R. MAC [illegible] A.B. MACAULAY 19[illegible] reprinted 1966) and 'Instruction in th[illegible] Christi[illegible] [illegible]eli[illegible] trans. by ALICE MEAD SWING, in ALBERT T. SWING, ***The Theology of Albrecht Ritschl*** (1901). Works discussing his theology are: ALFRED E. GARVIE, ***The Ritschlian Theology,*** 2nd ed. (1902), a standard treatment; PHILIP HEFNER, ***Faith and the Vitalities of History*** (1966), an analysis of aspects of Ritschl's theology and Hefner's suggestions concerning a viable theological method (bibliography included); GOSTA HOK, ***Die elliptische Theologie Albrecht Ritschls*** (1942), the best treatment of Ritschl's theological method; includes an extensive bibliography of secondary sources; DAVID L. MUELLER, ***An Introduction to the Theology of Albrecht Ritschl*** (1969), an assessment of the major lines of Ritschl's theology including a statement concerning his relevance today; and ROLF SCHAFER, ***Ritschl*** (1968), a new analysis including current bibliography.

(D.L.M.)

Ritual

Exhibited by all known societies, ritual is a specific, observable kind of behaviour based upon established or traditional rules. It is thus possible to view ritual as a way of defining or describing man.

Parallel between man's ritual and verbal behaviour

Nature and significance. Man is sometimes described or defined as a basically rational, economic, political, or playing species. Man may, however, also be viewed as a ritual being, who exhibits a striking parallel between his ritual and verbal behaviour. Ritual, a pervasive kind of behaviour, can be seen as basic to the understanding of man. Just as language is a symbolic system based upon arbitrary rules, ritual may be viewed as a symbolic system of acts based upon arbitrary rules.

The intricate, yet complex, relation between ritual and language can be seen in the history of various attempts to explain ritual behaviour. In most explanations, language becomes a necessary factor in the theory concerning the nature of ritual, and the specific form of language that is tied to explanations of ritual is the language of myth. Both myth and ritual remain fundamental to any analysis of religions.

Three general approaches to a theory about the nature and origin of ritual prevail.

The origin approach. The earliest approach was an attempt to explain ritual, as well as religion, by means of a theory concerned with historical origin. In most cases, this theory also assumed an evolutionary hypothesis that would explain the development of ritual behaviour through history. The basic premise, or law, for this approach is that ontogeny (development of an individual organism) recapitulates phylogeny (evolution of a related group of organisms), just as the human embryo recapitulates one stage of human evolutionary history in the womb; *e.g.*, the gill stage. The solution to explaining the apparently universal scope of ritual depended upon the success in locating the oldest cultures and cults. Scholars believed that if they could discover this origin, they would be able to explain the contemporary rituals of man.

Early emphasis on sacrifice

There are almost as many solutions as authors in this approach. In the search for an origin of ritual, research turned from the well-known literate cultures to those that appeared to be less complex and preliterate. The use of the terms primitive religion and primitive cultures comes from this approach in seeking an answer to the meaning of ritual, myth, and religion. Various cultures and rituals were singled out, sacrifice of either men or animals becoming one of the main topics for speculation, though the exact motivation or cause of sacrificial ritual was disputed among the leading authors of the theory. For W. Robertson Smith, a British biblical scholar who first published his theory in the ninth edition of *Encyclopædia*

Britannica (1875–89), sacrifice was motivated by the desire for communion between members of a primitive group and their god. The origin of ritual, therefore, was believed to be found in totemic (animal symbolic clan) cults; and totemism, for many authors, was thus believed to be the earliest stage of religion and ritual. The various stages of ritual development and evolution, however, were never agreed upon. Given this origin hypothesis, rituals of purification, gift giving, piacular (expiatory) rites, and worship were viewed as developments, or secondary stages, of the original sacrificial ritual. The Christian Eucharist (Holy Communion), along with contemporary banquets and table etiquette, were explained as late developments or traits that had their origin and meaning in the totemic sacrifice.

The influence of Robertson Smith's theory on the origin of ritual can be seen in the works of the English anthropologist Sir James Frazer, the French sociologist Émile Durkheim, and Sigmund Freud, the father of psychoanalysis. Though they were not in complete agreement with Smith, sacrifice and totemism remained a primary concern in their search for the origin of religion. For Frazer, the search led to magic, a stage preceding religion. Both Smith and Frazer led Durkheim to seek the origin of ritual and religion in totemism as exemplified in Australia. Durkheim believed that in totemism scholars would find the original form of ritual and the division of experience into the sacred and the profane. Ritual behaviour, they held, entails an attitude concerned with the sacred; and sacred acts and things, therefore, are nothing more than symbolic representations of society. In his last published work, *Moses and Monotheism,* Freud also remained convinced that the origin of religion and ritual is to be found in sacrifice.

The functional approach. The second approach to explaining ritual behaviour is certainly indebted to the work of such men as Smith, Freud, and Durkheim. Yet very few, if any, of the leading contemporary scholars working on the problems of religion, ritual, and myth begin with a quest for origins. The origin-evolutionary hypothesis of ritual behaviour has been rejected as quite inadequate for explaining human behaviour because no one can verify any of these bold ideas; they remain creative speculations that cannot be confirmed or denied.

Turning from origin hypotheses, scholars next emphasized empirical data gathered by actual observation. Contemporary literature is rich in descriptions of rituals observed throughout the world. If the term origin can be used as central to the first approach, the term function can be used as indicative of the primary focus of the second approach. The nature of ritual, in other words, is to be defined in terms of its function in a society.

Individual needs and social equilibrium

The aim of functionalism is to explain ritual behaviour in terms of individual needs and social equilibrium. Ritual is thus viewed as an adaptive and adjustive response to the social and physical environment. Many leading authorities on religion and ritual have taken this approach as the most adequate way to explain rituals. Bronisław Malinowski, A.R. Radcliffe-Brown, E.E. Evans-Pritchard, Clyde Kluckhohn, Talcott Parsons, and Edmund Leach, all English or American anthropologists, have adopted a functional approach to explain ritual, religion, and myth.

Most functional explanations of ritual attempt to explain this behaviour in relation to the needs and maintenance of a society. The strengths of this approach are dependent upon a claim that it is both logical and empirical. It is a claim, however, that is open to serious criticism. If the aim of functionalism is to explain why rituals are present in a society, it will be necessary to clarify such terms as need, maintenance, and a society functioning adequately, and this becomes crucial if they are to be taken as empirical terms. From a logical point of view, functionalism remains a heuristic device, or indicator, for describing the role of ritual in society. If it is asserted that a society functions adequately only if necessary needs are satisfied; and if it is further asserted that ritual does satisfy that need, scholars cannot conclude that, therefore, ritual is present in that society without committing the logical fallacy of affirming the consequent. To assert that the need is satisfied "if and only if" ritual is present is a tautology and a reversal of the claim to be empirical.

The history of religions approach. A third approach to the study of ritual is centred on the studies of historians of religion. The distinction between this approach and the first two is that though many historians of religions agree with functionalists that the origin-evolutionary theories are useless as hypotheses, they also reject functionalism as an adequate explanation of ritual. Most historians of religions, such as Gerardus van der Leeuw in The Netherlands, Rudolf Otto in Germany, Joachim Wach and Mircea Eliade in the United States, and E.O. James in England, have held the view that ritual behaviour signifies or expresses the sacred (the realm of transcendent or ultimate reality). This approach, however, has never been represented as an explanation of ritual. The basic problem with it remains that it cannot be confirmed unless scholars agree beforehand that such a transcendent reality exists (see also RELIGION, STUDY OF).

Functions of ritual. Ritual behaviour, established or fixed by traditional rules, has been observed the world over and throughout history. In the study of this behaviour, the terms sacred (the transcendent realm) and profane (the realm of time, space, and cause and effect) have remained useful in distinguishing ritual behaviour from other types of action.

The sacred and the profane

Although there is no consensus on a definition of the sacred and the profane, there is common agreement on the characteristics of these two realms by those who use the terms to describe religions, myth, and ritual. For Durkheim and others who use these terms, ritual is a determined mode of action. According to Durkheim, the reference, or object, of ritual is the belief system of a society, which is constituted by a classification of everything into the two realms of the sacred and the profane. This classification is taken as a universal feature of religion. Belief systems, myths, and the like, are viewed as expressions of the nature of the sacred realm in which ritual becomes the determined conduct of the individual in a society expressing a relation to the sacred and the profane. The sacred is that aspect of a community's beliefs, myths, and sacred objects that is set apart and forbidden. The function of ritual in the community is that of providing the proper rules for action in the realm of the sacred as well as supplying a bridge for passing into the realm of the profane.

Although the distinction between the sacred and profane is taken as absolute and universal, there is an almost infinite variation on how this dichotomy is represented—not only between cultures but also within a culture. What is profane for one culture may be sacred to another. This may also be true, however, within a culture. The relative nature of things sacred and the proper ritual conducted in relation to the sacred as well as the profane varies according to the status of the participants. What is set apart, or holy, for a sacred king, priest, or shaman (a religious personage having healing and psychic transformation powers), for example, will differ from the proper ritual of others in the community who are related to them, even though they share the same belief systems. The crucial feature that both sustains these relations and sets their limits is the ritual of initiation.

Three further characteristics are generally used to specify ritual action beyond that of the dichotomy of sacred and profane thought and action. The first characteristic is a feeling or emotion of respect, awe, fascination, or dread in relation to the sacred. The second characteristic of ritual involves its dependence upon a belief system that is usually expressed in the language of myth. The third characteristic of ritual action is that it is symbolic in relation to its reference. Agreement on these characteristics can be found in most descriptions of the functions of ritual.

The scholarly disputes that have arisen over the functions of ritual centre around the exact relation between ritual and belief or the reference of ritual action. There is little agreement, for example, on the priority of ritual or myth. In some cases, the distinction between ritual, myth, and belief systems is so blurred that ritual is taken to include myth or belief (see also SACRED OR HOLY; MYTH AND MYTHOLOGY).

The variety of ritual references

The function of ritual depends upon its reference. Once again, although there is common agreement about the symbolic nature of ritual, there is little agreement with respect to the reference of ritual as symbolic. Ritual is often described as a symbolic expression of actual social relations, status, or the role of individuals in a society. Ritual is also described as referring to a transcendent, numinous (spiritual) reality and to the ultimate values of a community.

Whatever the referent, ritual as symbolic behaviour presupposes that the action is nonrational. That is to say, the means–end relation of ritual to its referent is not intrinsic or necessary. Such terms as latent, unintended, or symbolic are often used to specify the nonrational function of ritual. The fundamental problem in all of this is that ritual is described from an observer's point of view. Whether ritual man is basically nonrational or rational, as far as his behaviour and his belief system are concerned, is largely dependent upon whether he also understands both his behaviour and belief to be symbolic of social, psychological, or numinous realities. It is difficult to imagine a Buddhist, a Christian, or an Australian aborigine agreeing that his ritual action and beliefs are nothing but symbols for social, psychological, or ultimate realities. The notion of the sacred as a transcendent reality may, however, come closest to the participant's own experience. The universal nature of the sacred-profane dichotomy, however, remains a disputed issue.

What is needed is a new theory that will overcome the basic weaknesses of functional descriptions of ritual and belief. Until such a time, ritual will remain a mystery. The progress made in the study of language may be of help in devising a more adequate explanation of nonverbal behaviour in general and of ritual in particular.

Types of ritual. Because of the complexities inherent in any discussion of ritual, it is often useful to make distinctions by means of typology. Although typologies do not explain anything, they do help to identify rituals that resemble each other within and across cultures.

Imitative. All rituals are dependent upon some belief system for their complete meaning. A great many rituals are patterned after myths. Such rituals can be typed as imitative rituals in that the ritual repeats the myth or an aspect of the myth. Some of the best examples of this type of ritual include rituals of the New Year, which very often repeat the story of creation. In a passage from an Indian *Brāhmaṇa* (a Hindu scripture) the answer to the question of why the ritual is performed is that the gods did it this way "in the beginning." Rituals of this imitative type can be seen as a repetition of the creative act of the gods, a return to the beginning.

The "myth-ritual" school

This type of myth has led to a theory that all rituals repeat myths or basic motifs in myths. A version of this line of thought, often called "the myth-ritual" school, is that myth is the thing said over ritual. In other words, myths are the librettos for ritual. The works of such scholars as Jane Harrison and S.H. Hooke are examples of this theory. Although it cannot be denied that some rituals explicitly imitate or repeat a myth (*e.g.*, a myth of creation), it cannot be maintained that all rituals do so. The ritual pattern of the ancient Near East, which Hooke considers basic to the festival celebrating the creation, is itself a typological construction. In any case, al-, though there is a combat and killing narrated in the festival myth, no known evidence exists of ritual killing or of king-sacrifice in the ancient Near East. Nevertheless, some rituals do repeat the story of a myth and represent an important type of ritual behaviour, even though the type cannot be universalized as a description of all ritual action.

Positive and negative. Rituals may also be classified as positive or negative. Most positive rituals are concerned with consecrating or renewing an object or an individual, and negative rituals are always in relation to positive ritual behaviour. Avoidance is a term that better describes the negative ritual; the Polynesian word *tabu* (English, taboo) also has become popular as a descriptive term for this kind of ritual. The word taboo has been applied to those rituals that concern something to be avoided or forbidden. Thus, negative rituals focus on rules of prohibition, which cover an almost infinite variety of rites and behaviour. The one characteristic they all share, however, is that breaking the ritual rule results in a dramatic change in ritual man, usually bringing him some misfortune.

Variation in this type of ritual can be seen from within a culture as well as cross-culturally. What is prohibited for a subject, for example, may not be prohibited for a king, chief, or shaman. Rituals of avoidance also depend upon the belief system of a community and the ritual status of the individuals in their relation to each other. Contact with the forbidden or transgression of the ritual rules is often offset by rituals of purification.

Negative ritual, as noted above, is always in polarity with positive ritual. The birth of a child, the consecration of a king, a marriage, or a death are ritualized both positively and negatively. The ritual of birth or death involves the child or corpse in a ritual that, in turn, places the child or the corpse in a prohibitive status and thus to be avoided by others. The ritual itself, therefore, determines the positive or negative characteristic of ritual behaviour.

Sacrificial. Another type of ritual is classified as sacrificial. Its importance can be seen in the assessment of sacrificial ritual as the earliest or elementary form of religion (see also SACRIFICE).

Differentiation between sacrifice and other ritual acts

The significance of sacrifice in the history of religions is well documented. One of the best descriptions of the nature and structure of sacrifice is to be found in *Essai sur la nature et le fonction du sacrifice,* by the French sociologists Henri Hubert and Marcel Mauss, who differentiated between sacrifice and rituals of oblation, offering, and consecration. This does not mean that sacrificial rituals do not at times have elements of consecration, offering, or oblation but these are not the distinctive characteristics of sacrificial ritual. Its distinctive feature is to be found in the destruction, either partly or totally, of the victim. The victim need not be human or animal; vegetables, cakes, milk, and the like are also "victims" in this type of ritual. The total or partial destruction of the victim may take place through burning, dismembering or cutting into pieces, eating, or burying.

Hubert and Mauss have provided a very useful structure for dividing this type of ritual into subtypes. Though sacrificial rituals are very complex and diverse throughout the world, nevertheless, they can be divided into two classes: those in which the participant or participants receive the benefit of the sacrificial act and those in which an object is the direct recipient of the action. This division highlights the fact that it is not just individuals who are affected by sacrificial ritual but in many instances objects such as a house, a particular place, a thing, an action (such as a hunt or war), a family or community, or spirits or gods that become the intended recipients of the sacrifice. The variety of such rituals is very extensive, but the unity in this type of ritual is maintained in the "victim" that is sacrificed.

Life crisis. Any typology of rituals would not be complete without including a number of very important rites that can be found in practically all religious traditions and mark the passage from one domain, stage of life, or vocation into another. Such rituals have often been classified as rites of passage, and the French anthropologist Arnold van Gennep's study of these rituals remains the classic book on the subject.

The basic characteristic of the life-crisis ritual is the transition from one mode of life to another. Rites of passage have often been described as rituals that mark a crisis in individual or communal life. These rituals often define the life of an individual. They include rituals of birth, puberty (entrance into the full social life of a community), marriage, conception, and death. Many of these rituals mark a separation from an old situation or mode of life, a transition rite celebrating the new situation, and a ritual of incorporation. Rituals of passage do not always manifest these three divisions; many such rites stress only one or two of these characteristics.

Rituals of initiation

Rituals of initiation into a secret society or a religious vocation (viz., priesthood, ascetic life, medicine man) are often included among rites of passage as characteristic rituals of transition. The great New Year's rituals known throughout the world also represent the characteristic passage from old to new on a larger scale, that includes the whole society or community.

One of the dominant motifs of the life-crisis ritual is the emphasis on separation, as either a death or a return to infancy or the womb. In India, a striking example is the Hindu rite of being "twice born." The young boy who receives the sacred thread in the *upanayana* ritual, a ceremony of initiation, goes through an elaborate ritual that is viewed as a second birth. Rituals such as Baptism in early Christianity, Yoga in India, and the complex puberty rituals among North American Indian cultures exemplify this motif of death and rebirth in rites of passage.

Rituals of crisis and passage are often classified as types of initiation. An excellent description of such rites is found in *Birth and Rebirth* by Mircea Eliade. From Eliade's point of view, rituals, especially initiation rituals, are to be interpreted both historically and existentially. They are related to the history and structure of a particular society and to an experience of the sacred that is both transhistorical and transcendent of a particular social or cultural context. Culture, from this perspective, can be viewed as a series of cults, or rituals, that transform natural experiences into cultural modes of life. This transformation involves both the transmission of social structures and the disclosure of the sacred and spiritual life of man.

Initiation rituals can be classified in many ways. The patterns emphasized by Eliade all include a separation or symbolic death, followed by a rebirth. They include rites all the way from separation from the mother to the more complex and dramatic rituals of circumcision, ordeals of suffering, or a descent into hell, all of which are symbolic of a death followed by a rebirth. Rites of withdrawal and quest, as well as rituals characteristic of shamans and religious specialists, are typically initiatory in theme and structure. Some of the most dramatic rituals of this type express a death and return to a new period of gestation and birth and often in terms that are specifically embryological or gynecological. Finally, there are the actual rituals of physical death itself, a rite of passage and transition into a spiritual or immortal existence.

The various typologies of ritual that can be found in texts on religion and culture often overlap or reveal a common agreement in the way in which ritual behaviour can be classified. There is a striking contrast in the use of these typologies to interpret the meaning of ritual. In general, this contrast can be described in terms of two positions: the first emphasizes the sociopsychological function of ritual; the second, although not denying the first, asserts the religious value of ritual as a specific expression of a transcendental reality.

Nonverbal communication and meaning

Conclusion. Ritual behaviour is obviously a means of nonverbal communication and meaning. This aspect of ritual is often overlooked in the stress on the relation of ritual to myth. Thus, the meaning of ritual is often looked for in the verbal, spoken, or belief system that is taken as its semantic correlate. The spoken elements in a ritual setting do often reveal the meaning of a ritual by reference to a belief system or mythology, but not always. Such a connection has led to an overemphasis on the importance of the belief system or myth over ritual. To assert that myths disclose more than ritual ever can is an oversimplification of the complex correlation of these two important aspects of religion. A partial explanation of this emphasis is undoubtedly the fact that a vast amount of data, both primary and secondary, is literary in form. Theories about ritual are either deduced from the primary literature of a religious tradition or are translated into written language as a result of observation.

Ritual can be studied as nonverbal communication disclosing its own structure and semantics. Scholars have only recently turned to a systematic analysis of this important aspect of human behaviour; and progress in kinesics, the study of nonverbal communication, may provide new approaches to the analysis of ritual. This development may well parallel the progress in linguistics and the analysis of myth as an aspect of language.

A complete analysis of ritual would also include its relation to art, architecture, and the specific objects used in ritual such as specific forms of ritual dress. All of these components are found in ritual contexts, and all of them are nonverbal in structure and meaning.

Significance of "sacred time" and "sacred space"

Most rituals mark off a particular time of the day, month, year, stage in life, or commencement of a new event or vocation. This temporal characteristic of ritual is often called "sacred time." What must not be forgotten in the study of ritual is a special aspect of ritual that is often described as "sacred space." Time and place are essential features of ritual action, and both mark a specific orientation or setting for ritual. Time and space, whether a plot of ground or a magnificent temple, are ritually created and become, in turn, the context for other rituals. Examples of ritual time and ritual space orientation can be found in the rituals for building the sacrifice in Brahmanic Indian ritual texts; for the building of a Hindu temple or a Christian cathedral; and for consecrating those structures that symbolize a definite space-time orientation in which rituals are enacted. The shape, spatial orientation, and location of the ritual setting are essential features of the semantics of ritual action.

When particular ritual objects, dances, gestures, music, and dress are included in the study of ritual, the total structure and meaning of ritual behaviour far exceed any one description or explanation of ritual man. Most descriptions are selective and are dependent upon the theory and intent with which rituals are to be studied.

In recent years there has been little consensus among scholars on an adequate theory, or framework, for explaining or describing ritual. Though the term has often been used to describe the determined, or fixed, behaviour of both animals and men, the future study of ritual may disclose that this behaviour, found throughout history and cultures, is as unique to man as his capacity for speaking a language and that change in ritual behaviour is parallel to, or correlated with, change in language. Although great progress has been made in the analysis of man as the species who speaks, the syntax and semantics of ritual man are yet to be discovered.

BIBLIOGRAPHY. WILLIAM LESSA and EVON Z. VOGT, *Reader in Comparative Religion,* 3rd ed. (1971), is the best general anthology on classical and modern positions on religion, ritual, and myth (mainly concerned with nonliterate cultures). The *Reader* also includes an excellent bibliography. *Gods and Rituals,* ed. by JOHN MIDDLETON (1967), contains a good collection of essays on ritual practices in nonliterate cultures, with a fine bibliography for further study. Among the classic texts dealing with the origin of ritual and religion, there are three authors who remain important because of their enduring influence: W. ROBERTSON SMITH, *Lectures on the Religion of the Semites* (1889); EMILE DURKHEIM, *Les Formes élémentaires de la vie religieuse* (1912; Eng. trans., *The Elementary Forms of the Religious Life,* 1915); and SIGMUND FREUD, *Totem und Tabu* (1913; Eng. trans., *Totem and Taboo,* 1918). Among the classic positions on a functional approach to ritual are those of BRONISLAW MALINOWSKI, in his *Coral Gardens and Their Magic, 2* vol. (1935); and A.R. RADCLIFFE-BROWN, *The Andaman Islanders* (1922). Among more recent anthropological texts, E.E. EVANS-PRITCHARD, *Nuer Religion* (1956); and EDMUND LEACH, *Political Systems of Highland Burma* (1954), are very good examples of the continued development of the functional approach. MELFORD E. SPIRO, *Burmese Supernaturalism* (1967), is one of the best critical texts using data from Burmese Buddhism as support for a revised approach. VICTOR W. TURNER, *The Forest of Symbols* (1967), represents a novel analysis of dominant symbols in belief and ritual. Among valuable approaches by theologians and historians of religion are RUDOLF OTTO, *Das Heilige* (1917; Eng. trans., *The Idea of the Holy,* 1923); and JOACHIM WACH, *The Comparative Study of Religions* (1958). JANE E. HARRISON, *Themis,* 2nd ed. rev. (1927); and S.H. HOOKE (ed.), *Myth, Ritual and Kingship* (1958), are the best examples of the myth-ritual school. An excellent critique of this school may be found in JOSEPH E. FONTENROSE, *The Ritual Theory of Myth* (1966). HENRI HUBERT and MARCEL MAUSS, *Essai sur la nature et le fonction du sacrifice*

(1899; Eng. trans., ***Sacrifice: Its Nature and Function*** (1964), remains the standard analysis of sacrifice as ritual; and ARNOLD VAN GENNEP, ***Les Rites du passage*** (Eng. trans. 1960), although written in 1909, continues to remain an important work on ritual as passage. MIRCEA ELIADE, ***Birth and Rebirth*** (1958), is an excellent study of ritual as initiation from a history of religions viewpoint; has a good bibliography.

(Ha.P.)

River Deltas

Deltas are low-lying plains composed of stream-borne sediments deposited by a river at its mouth as it enters the sea. Some 2,500 years ago, Herodotus recognized that the land bound by the seaward-diverging distributary branches of the Nile and the sea was deltoid in shape; and so used the Greek letter Δ (delta) to describe it. Although many of the world's great deltas are deltoid or triangular in shape, notable exceptions occur; often the delta shape is controlled by the outline of the water body being filled. For this reason, the term delta is now normally applied, without reference to shape, to the exposed and submerged plain formed by a river at its mouth.

Deltas are widely distributed; they form along the coasts of virtually every landmass on the globe and occur in all climatic regions, from harsh arctic and desert climates to humid tropical and temperate zones (Table). Of the larger deltas in the world, 11 are located in the U.S.S.R., 7 are in Southeast Asia, 6 are in South America, and 4 each are in Africa and North America.

Deltas display enormous variety in size, shape, structure, composition, and genesis. These differences result from the same event taking place in a wide range of settings. Numerous factors influence the character of a delta; the more important are: (1) geologic setting and sediment sources in the drainage basin, (2) climatic conditions, (3) tectonic stability (magnitude and frequency of uplift), (4) river slope and flooding characteristics, (5) intensities of depasitional and erosional processes, and (6) tidal range and offshore energy conditions. Combinations of these factors and time result in the wide variety of modern deltas and their changing configurations.

Deltas have been important to mankind since prehistoric times. The abundant wildlife and edible plants in deltaic areas attracted early man. Mazes of interconnecting waterways provided natural avenues for communication and trade. Sands, silts, and clays deposited by floodwaters were extremely productive; and, as man's agricultural technology increased, huge civilizations flourished in the deltaic plains of the Nile and Tigris-Euphrates. Excavations by archaeologists have revealed the grandeur of those civilizations.

River mouths give seagoing ships access to interior ports, and many of the world's great harbours are located in delta plains. Alexandria on the Nile and New Orleans on the Mississippi, both flourishing seaports, owe their success to their location along rivers.

In recent years geologists have discovered that much of the world's petroleum resource is found in ancient deltaic rocks. Thus understanding of deltas is important to petroleum exploration.

Man has failed to exploit the potential of the majority of the deltas. Even though the Irrawaddy, Ganges-Brahmaputra, and Mekong deltas are the rice bowls of Asia, present agricultural practices do not allow maximum utilization of the fertile lands found there. Improved farming techniques, especially those developed by the United Nations-sponsored Economic Commission for Asia and the Far East (ECAFE), will undoubtedly increase future yields in these areas. Other large tropical deltas, such as the ·Niger, Amazon, Orinoco, and Magdalena, have been virtually untouched, and only a thorough knowledge of deltaic processes will lead to development of these areas.

This article treats the processes that are involved in river delta formation, the morphology, stratigraphy, and structure of deltas, and deltaic changes that occur through time by reason of changes in the supply and removal of sediment. A concluding section on experimental studies treats computer-simulation models of the growth of deltas. For additional detail on those aspects of flowing water and current systems that are relevant to delta formation, growth, and maintenance, see RIVERS AND RIVER SYSTEMS; FLUVIAL PROCESSES; DENSITY CURRENTS; and WATER WAVES. The ultimate source of deltaic sediments is covered in SEDIMENT YIELD OF DRAINAGE SYSTEMS, and the general distribution of sediment along coastal areas is treated in the articles BEACHES; COASTAL FEATURES; and CONTINENTAL SHELF AND SLOPE. Lacustrine sedimentation is dealt with in LAKES AND LAKE SYSTEMS and the terrestrial analogues of deltas are separately discussed in ALLUVIAL FANS. For information on sea level changes that have affected the world's deltas in the geological past, see PLEISTOCENE EPOCH and HOLOCENE EPOCH. Finally, the interested reader should consult the articles COMPUTER and MATHEMATICS AS A CALCULATOR-SCIENCE for further insight to computer modelling generally.

PROCESSES RESPONSIBLE FOR DELTA FORMATION AND OCCURRENCE

The presence of a delta represents the continuing ability of rivers and river systems (*q.v.*) to supply and deposit stream-borne sediments more rapidly than they can be removed by water waves and ocean currents (*qq.v.*). Because delta building is a contest between the river and the sea, it presents one of the most dynamic situations in nature. Many factors, some resulting from fluvial processes (*q.v.*) and others from marine processes, affect the riverine-marine balance. Conditions existing in the drainage basin or source area of a river exert considerable influence on the formation of the delta. The size, shape, relief, soil development, and geologic type of rocks in the basin control, to a large degree, the type and amount of sediment available for transport to the sea.

Runoff and sediment transport

One of the most important factors is the type of climate that exists in the drainage area. In tropical climates, rainfall is distributed abundantly over the basin throughout the year; maximum downpours occur during a few months of the year. Water and sediment discharge are therefore continuous all year, but maximums coincide with periods of heaviest rainfall. In arid climates rainfall is not continuous throughout the year; rather, it occurs sporadically, causing rapid and catastrophic runoff. Rivers draining this type of basin have erratic floodwater and sediment discharges. Arctic climates result in similar runoff characteristics, except that throughout most of the year the water available for runoff is frozen and cannot serve as a transport agent until the period of thaw. In the three cases cited the total yearly rainfall could be identical in the basins, but because the frequency of occurrence and the availability of the rivers to serve as transport agents differ, the sediment supply and water reaching the delta cause quite different morphological features to develop. In the case of a rather constant supply of sediment and water all year, channel systems in the delta can adjust quite easily to handling the volume of water and sediment, and as a consequence, channels are quite stable and do not tend to migrate rapidly. Erratic flood conditions, such as exist under arid and Arctic climates, supply large quantities of sediment and water to the delta during a short period of time. Distributary channels therefore never completely adjust to the large, rapid influx of sediments and water. New channels form rapidly during flood, function for a short period of time, and then are abandoned and filled as floodwaters recede. Channels are unstable and therefore tend to migrate considerably. Other climatic conditions, not quite so harsh as those described, impose similar control on delta morphology. Thus climatic factors in the basin determine the quantity of both water and sediment supplied to the delta and also control the frequency of the input to the delta.

Sediment transport within the delta proper is in the form of bed load and suspended load. Bed load consists of coarser particles that travel close to the bed of the channel; in contrast, the suspended load consists of finer particles that travel primarily above the bed. Channels carrying large volumes of bed load will normally be wide and shallow and will tend to rapid lateral migration. A high percentage of suspended load results in deep, narrow

Major River Systems of the World

river system	country	drainage area (000 sq km)	delta area (sq km)	mean discharge (00 cu m/sec)	yearly sediment yield (000,000 metric tons/year)	tidal range at mouth (m)	climate in delta area
Amazon	Brazil	7,050	estuary	1,800	400	5.70	tropical rainforest
Rio de la Plata–Paraná	Argentina	4,144	14,245	220	150	1.00	humid subtropical
Congo	Zaire	3,457	2,072	413	71	1.70	tropical savanna
Nile	Egypt	3,349	20,228	31	122	0.50	humid subtropical
Mississippi-Missouri	U.S.	3,221	26,159	184	495	0.50	subtropical desert
Ob-Irtysh	U.S.S.R.	2,975	2,849	158	20	0.70	subarctic
Yenisey	U.S.S.R.	2,580	2,460	190	11	0.40	tundra
Lena	U.S.S.R.	2,490	25,900	163	12	0.30	tundra
Yangtze	China	1,959	estuary	340	500	4.20	humid subtropical
Niger	Nigeria	1,890	36,260	61	5	2.20	tropical rainforest
Amur	U.S.S.R.	1,855	estuary	124	20	2.30	subarctic
Mackenzie	Canada	1,841	12,200	113	...	0.40	tundra
Ganges-Brahmaputra	Bangladesh	1,621	59,570	385	2,400	5.60	tropical rainforest
St. Lawrence-Great Lakes	Canada	1,463	estuary	102	...	...	humid continental
Volga	U.S.S.R.	1,360	9,970	80	283	...	middle latitude desert
Zambezi	Mozambique	1,330	7,148	71	...	4.00	tropical savanna
Indus	Pakistan	1,166	7,770	55	480	4.20	subtropical desert
Shatt al-Arab (Tigris-Euphrates)	Iraq	1,114	—	14	62	2.80	subtropical steppe
Nelson	Canada	1,072	—	23	...	5.20	subarctic
Murray-Darling	Australia	1,057	—	4	35	2.80	Mediterranean
Orinoco	Venezuela	948	24,553	198	95	2.20	tropical rainforest
Tocantins	Brazil	906	—	102	...	4.30	tropical rainforest
Yukon	U.S.	828	5,802	59	...	1.20	subarctic
Danube	Romania	816	4,299	72	80	0.09	humid continental
Mekoug	South Vietnam	795	50,000	110	187	3.50	tropical savanna
Huang Ho	China	745	1,940	33	2,080	3.40	middle latitude steppe
São Francisco	Brazil	673	—	28	...	2.50	tropical savanna
Kalyma	U.S.S.R.	647	3,704	38	...	0.06	tundra
Dnepr (Dnieper)	U.S.S.R.	504	640	17	9	0.09	middle latitude steppe
Amu Darya	U.S.S.R.	465	3,522	—	130	—	middle latitude desert
Don	U.S.S.R.	422	647	—	5	—	middle latitude steppe
Irrawaddy	Burma	411	19,943	130	330	5.50	tropical rainforest
Indigirka	U.S.S.R.	360	9,169	18	...	0.12	tundra
Dvina, Northern	U.S.S.R.	357	1,100	34	...	3.00	subarctic
Godāvari	India	298	—	36	—	1.80	middle latitude desert
Magdalena	Colombia	284	2,460	75	—	0.60	tropical steppe
Fraser	Canada	238	—	27	—	3.10	marine west coast
Rhine	The Netherlands	160	—	22	—	5.50	marine west coast
Rhône	France	96	1,683	17	—	0.20	Mediterranean, subtropical

Source: Modified and updated from J.M. Coleman "Deltaic Evolution," *Encyclopedia of Geomorphology* (1967'. Table compiled from various sources.

channels that are relatively stable. Large sediment concentrations transported in a delta also increase the density of river water. The greater the contrast in density between river water and seawater, however, the more compressed the sediment plume will be as it moves offshore from a river mouth. Thus, to some degree the amount of sediment load controls lateral spreading of stream-borne material along the coast at a river mouth (see also DENSITY CURRENTS).

Marine processes and tectonic effects

Marine processes also play a key role in molding the landforms in the delta. Wave-energy levels, wind intensity and direction, coastal currents, tidal action, and offshore slope are the major factors involved. High-energy levels along the front of a delta rapidly rework the stream-borne sediment introduced to the sea and tend to spread sediments laterally, clogging channels. Thus high-energy conditions favour the formation of fewer channel outlets and of sandy beach deposits between channel outlets. Sediment spreading at river mouths is also enhanced by offshore currents and tidal action. Low-energy conditions, weak currents, and small tidal range result in less lateral spreading of sediment and in a highly indented delta coastline with many river mouths. See further WATER WAVES; COASTAL FEATURES; BEACHES.

Many of the world's large and important deltas occur in subsiding basins and are controlled by regional tectonics or local fault systems. There is still controversy as to whether it is the subsiding basin that attracts the river system or whether it is the load of the sediment deposited by the river that causes the earth's crust to buckle and subside. In addition to regional subsidence, local compaction caused by differential sediment loading is an important process in delta building. Although compaction is operative in nearly all deltas, it is much more prevalent in deltas that transport large fine-grained sediment loads. In such cases, heavier sandy deposits form at the river mouths, sink into the underlying soft muds, forcing large masses of clay to rise toward the surface. In some deltas these clay deposits will reach the surface and form mud islands around the river mouths. Such a condition exists in the Mississippi River Delta, where the islands, referred to as mud lumps, offer serious obstructions to ships navigating the river mouths.

MORPHOLOGY OF DELTAS

In plan or map view, deltas appear to be hoeplessly complex mazes of channels, lakes, and marshes. There is, however, an orderly arrangement of component parts, three of which are found in most deltas. The most landward section is called the upper delta plain, the middle one the lower delta plain, and the third the subaqueous delta, which lies seaward of the shoreline and forms below sea level. Variations in the proportions of each of these components give rise to the differing sizes and shapes of the world's deltas (Figure 1).

Delta plains

That part of the river confined by valley walls, so that the river commonly is restricted from spreading out, is called the alluvial plain. It may be thought of as a conduit through which sediment and water derived from the drainage basin are brought to the sea. At some point downstream the plain broadens out, and most river channels break up into more than one course. This is the apex of the delta and the beginning of the upper delta plain. This part of the delta possesses many morphologic surfaces and processes that are similar to those of the alluvial plain. All of this land lies at an elevation above the effective intrusion of tidal water and is formed entirely by riverine processes. Areas between channels usually support broad freshwater marshes, swamps, or shallow lakes. The lower delta plain is periodically inundated by tidal waters, and landforms result from the interaction of both riverine and marine processes. Areas between the channels show a variety of landforms, ranging from brackish water bays, marshes, and mangrove swamps to

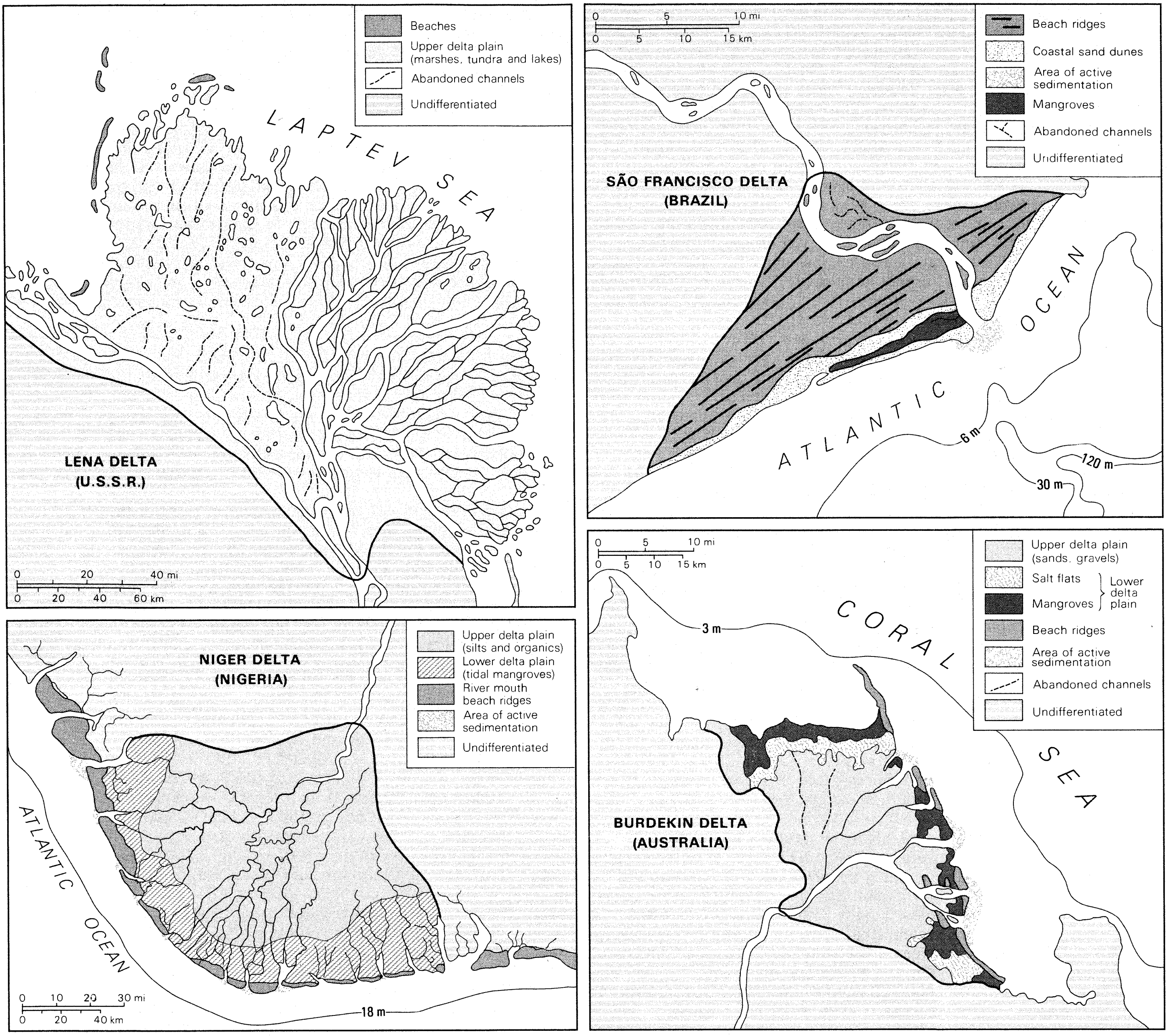


Figure 1: The component parts of four representative river deltas.

hypersaline tidal flats and beach ridges. The subaqueous delta forms entirely below the level of the sea and commonly constitutes the obvious bulge on the continental shelf seen seaward of many deltas. Marine processes are dominant, riverine factors playing a secondary role. The importance of this component, however, should not be underrated, because it is the foundation for construction of the exposed portion of the delta.

One of the largest delta plains, the Ganges-Brahmaputra in Bangladesh, encompasses some 60,000 square kilometres (23,166 square miles), over half of which is inundated by tidal waters (lower delta plain), forming vast mangrove swamps locally referred to as the sunderbands. Deltas such as the Ganges-Brahmaputra and Niger (Figure 1), in which high tidal ranges occur, normally display broad lower delta plains at the expense of upper delta plains. Other deltas, such as the Lena (U.S.S.R.; Figure 1), Nile (Egypt), and Volga (U.S.S.R.), in which tidal range is small, have poorly developed lower delta plains but large, well-developed upper delta plains. Formation of a large subaqueous delta plain appears to be controlled by offshore slope and sediment load. Low, sloping continental shelfs such as exist off the deltas of the Amazon (Brazil), Orinoco (Venezuela), and Huang Ho (China) favour development of broad, widespread, subaqueous delta plains.

Channels and levees

Channel patterns within most delta plains can be grouped into three general patterns: single or simple, long, straight channels; complex, multiple-braided channels; and continuous seaward-bifurcating or seaward-branching channels. The first type most commonly develops in deltas that carry a fine-grained sediment load and that have nonerratic flood discharges and high wave energy offshore (Mekong, Congo, San Francisco del Norte). The second type is characterized by complex anastomosing patterns within the lower delta plain; it results from high wave energy offshore, large tidal ranges, somewhat erratic flooding, and large sediment (bedload) volumes. The Niger (Figure 1) and Zambezi provide excellent examples of this type. Weak wave energy, fine-grained sediments, and a low tidal range tend to form seaward-bifurcating channel patterns; for example, such as those that exist in the Mississippi River and Volga River deltas.

Bordering the channels in most deltas are slightly higher areas called natural levees. These form as a result of the deposition of sediment during river floods. As floodwaters top the channel banks, velocity is reduced, causing deposition of the coarser suspended sediment near the channel margins. Natural levees are best developed in rivers that flood and reach bankfull stage each year and that carry a heavy suspended sediment load. The Mississippi, Mekong, Volga, and Nile have well-developed natural levees, whereas the Lena, Mackenzie, and San Francisco have poorly developed levees. Between the channels within a delta are usually found low-lying interlevee basins or interfluves. These areas vary considerably from one delta region to another. Climatic factors, offshore energy conditions, and tidal range exert the major controlling factors. Thus in regions of low energy, small tidal range, and tropical and temperate climate, interlevee basins will commonly show broad, flat, marsh plains, open-water bays, or broad freshwater swamps. In regions having high tidal ranges and tropical climates, vast mangrove swamps and numerous tidal channels are most common. In high-tide arid deltas, broad salt flats and occasional areas of wind-blown dune fields are encountered. Extreme wave action, when combined with a heavy sediment load, will result in a considerable number of beach ridges in the interdistributary area.

Delta front and river mouth

The subaqueous delta plain, or delta front, forms the submarine platform across which the exposed delta will eventually build. Broad subaqueous delta plains are normally associated with shallow, sloping continental shelves, whereas little or no development is associated with steeply sloping continental shelves or with areas where large submarine canyons originate near river mouths. For example, the Congo River carries an extremely heavy load of sediment, yet has a poorly developed subaqueous delta. A large submarine canyon exists immediately off the river mouth, and all sediments delivered to the sea by the river are funnelled down this canyon for deposition beyond the edge of the continental shelf. In deltas where the continental shelf has a low slope, the sediment leaves the seaward termination of the distributary channels and is deposited on the continental shelf, thereby forming a bulge in the offshore contours (see also CANYONS, SUBMARINE; CONTINENTAL SHELF AND SLOPE).

River mouths are highly variable. Some display broad, bell-shaped patterns, and others protrude as narrow bands into the sea, resembling the artificial jetties constructed by man at many river mouths. The reason for these variations in river-mouth patterns is poorly understood, but most scientists agree that offshore energy levels, density differences between river water and seawater, and tidal range are important processes in producing the differences that exist in many world deltas. Seaward of the river mouths are shoal areas referred to as river-mouth bars. As the river water leaves the confines of its channel banks, current velocities are reduced, and the coarser sediment is deposited as a shoal in front of the river mouth. Here it is reworked and redistributed by waves and tidal currents. The finer grained suspended sediments do not drop out immediately, but rather spread out and are deposited seaward of the bar, forming the resulting delta front.

DELTAIC STRATIGRAPHY AND STRUCTURE

Sediment distribution and aggradation

The presence of a delta along a coast represents the ability of the river to bring sediments to a coast and deposit them faster than marine currents and waves can remove them. Essentially a dispersal mechanism, delta building consists of distributing and sorting the sediments derived from the drainage basin and brought to the coast by the river. The sediment is dispersed in two distinct ways during a single year or during a flood cycle. During low river stage or times of normal discharge, transported sediment essentially remains within the channels throughout its travel through the delta and does not begin to be dispersed laterally until it reaches the river mouth. During this time river-current velocities are generally low near the river mouths, and marine currents and waves can easily rework and distribute the sediment laterally along the front of the delta. The coarser grained sediments are concentrated by waves near the river mouth, whereas the finer grained sediments are moved laterally along the delta coastline or offshore. When flood period ensues, however, transported sediment and water are no longer confined to the channels within the delta but spread over the river banks into the adjacent interdistributary areas. Silts and clays are therefore introduced rapidly into areas that normally have very low rates of sediment accumulation. This process is the major factor in building up or aggrading the subaerial delta plain. At the river mouth, river-current velocities are appreciably higher than those which prevail during low river stage and tend to overpower marine currents and wave action. Sediments are dispersed much farther seaward during flood season than during low stage when they are distributed laterally along the coast. Silts, and occasionally sands, are transported across the river-mouth shoal and are deposited along the delta front or steeper sloping portion of the subaqueous delta. This seaward spreading of sediments is the mechanism by which the delta grows seaward. The seaward movement of the delta is known as prograding.

The depositional sequence

The dispersal patterns and sorting of sediment result in a well-drained, well-developed depositional sequence both horizontally and vertically. In the subaerial delta plain, fine-grained silts and clays are deposited by the river along with the organic material that accumulates *in situ* in the interdistributary areas. Immediately in front of the delta mouths, sands and silts are accumulating and are constantly being reworked and sorted by marine processes. These deposits are building seaward and thus have a fairly steep seaward slope. Farther offshore, the fine-grained sediments, predominantly clays, are accumulating beyond the delta front. Thus, the horizontal sequence of deltaic sediments is a gradation from silts, clays, and organic material in a landward position, through coarser sands and silts near the delta mouths, to finer grained marine clays offshore. Vertically, this same sequence also can be found: the clays and organic sediments are found at the highest levels, sands and silts at the intermediate level (shallow nearshore water depths), and marine clays at the lowest level 30.5- to 71-metre (100- to 200-foot) water depths. This horizontal and vertical sequence has long been recognized, and in 1890 G.K. Gilbert, the American geologist, used the terms topset, foreset, and bottomset deposits to denote the three units. The topset deposits correspond to the subaerial delta sediments, the foreset deposits to the coarse sands and silts laid down along the delta front, and the bottomset deposits to the marine clays accumulating beyond the delta front. These relationships are shown clearly in the cross section of a typical delta (Figure 2). Gilbert used these terms to describe the sediments composing small lake deltas in the Pleistocene glacial Lake Bonneville. In large, complex deltas where many processes are interacting, with differing intensities, and where compaction of sediments is an active process, it is more difficult to recognize these simple depositional components. As more details of different types of deltas become better known, the concept still appears valid, but sediments composing each of these components become increasingly more complex and highly variable.

The subaerial delta, or topset component, consists primarily of channel sands, natural levee silts and clays, and interdistributary peats and clays (Figure 2). Channel deposits commonly consist of poorly sorted sands that contain a high percentage of included clays, wood and clay fragments, heavy minerals, and a normally high mica content. In some cases, channel deposits are composed of interbedded silts, clays, and transported organic sediments instead of sands. In these cases, the process of channel abandonment was quite slow, and only fine-grained material was available for filling the channel. Regardless of composition, channel deposits are normally lenticular (lens shaped) in cross-sectional shape and elongate in nature. Thickness is dependent on channel scour depth as well as intensity of compaction. Natural-levee sediments overlie channel deposits and are much finer grained. They are composed of suspended sediments, interbedded silts and clays being predominant. Because they are exposed to air

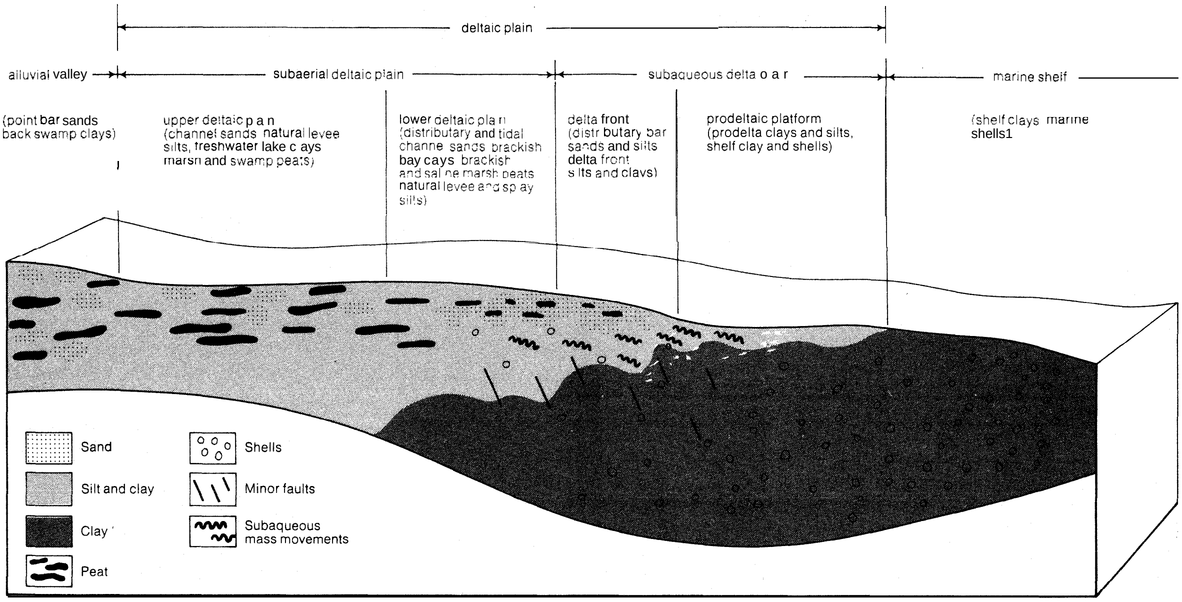

deltaic plain
alluvial valley
subaerial deltaic plain
subaqueous delta o a r
marine sheif
(point bar sands back swamp clays)
upper deltaic p a n (channel sands natural levee silts, freshwater lake c ays marsn and swamp peats)
lower deltaic pla n (distributary and tidal channe sands brackish bay cays brackish and sal ne marsh peats natural levee and sp ay silts)
delta front (distr butary bar sands and silts delta front s lts and clays)
prodeltaic platform (prodelta clays and silts, shelf clay and shells)
(shelf clays marine shells1
Sand
Silt and clay
Clay
Peat
Shells
Minor faults
Subaqueous mass movements

Figure 2: Characteristic distribution of the kinds of deposits occurring in the submerged and exposed parts of a delta.

From E.H. Rainwater "The Geological importance of Deltas," *Deltas in Their Geological Framework.* Houston Geological Society (1966)

a large part of the time, oxidation and a constantly fluctuating water table are common. These processes impart a red coloration to the sediment and result in inclusions of calcium carbonate. Encasing the channel and natural levee deposits are interdistributary sediments. These deposits are highly variable from one delta to another, but most commonly they include finely laminated clays containing zones of lenticular silt and peat layers. In some cases, burrowing organisms rework the clays as they accumulate, and mottled deposits result. In many deltas these deposits form the highest percentage of the topset stratum.

The delta front or foreset deposits are laid down in subaqueous environments immediately seaward of the delta coastline. Nearest the river mouths and in shallow water, coarse-grained sands accumulate as river-mouth bars (Figure 2). Normally, they are the coarsest material found in delta deposits, are relatively clean, and contain occasional thin stringers of clay and transported (rafted) organics. In a seaward direction and radiating away from the river mouth, the sands grade laterally into silts that are deposited in somewhat deeper water. Such deposits commonly consist of interbedded silt, silty clay, and clay layers; the coarser material is deposited farther offshore during flood stage. Quite often subaqueous mass movement will result in distortion of these sediments. In deeper water and still farther offshore, only the finest of the stream-borne material slowly settles to the bottom. These deposits are referred to as prodelta clays, they commonly consist of fine-grained clays containing an abundant marine faunal assemblage.

Thus, delta deposits grade from coarse to fine in an offshore direction and from fine to coarse in an upward direction in a vertical section.

DELTAIC CHANGES THROUGH TIME

The discussion of the processes controlling delta formation thus far has largely ignored the element of time, yet geologic time and the numerous combinations of processes result in dynamically changing conditions within deltas. Because the mechanisms of supply and removal of sediment cannot remain constant throughout time, one of the major characteristics of deltas is that in time they tend to shift position. As a delta progrades or builds farther and farther into the sea, the gradient and sediment-carrying capacity of the river gradually decreases, and shorter, more efficient routes to the sea can be found in adjacent areas. Eventually the river will seek the shorter route, and active deltaic sedimentation will be initiated in a new site. The new delta will then begin to prograde seaward rapidly. This phase is referred to as the constructional phase of delta building. The abandoned or inactive delta, however, will be rapidly attacked by the sea because it will no longer be fed sediment by the river; the balance between sea and river forces will then favour erosion, and the delta shoreline will retreat. This phase of delta evolution has been termed the destructional phase. The distinction between these two phases is normally clear in most deltas, but in others it is less apparent. The major reason for this difference is that not all deltas switch their courses in the same manner. Three major types of deltaic change can be recognized: (1) lobe switching, (2) channel switching, and (3) alternate channel extension. These three types are illustrated diagrammatically in Figure 3.

Lobe switching

In type 1, the delta builds a complex of distributaries that prograde simultaneously seaward, producing a lobe of sediment that protrudes seaward from the coastline. Adjacent areas are deprived of sediment; and, because of rapid subsidence or heavy wave attack, these areas open up and form open-water areas adjacent to the active delta. At some point, generally upstream from the delta apex, the river will eventually break out of its levees and seek a shorter route to the sea, which will normally correspond to the adjacent indented coastline. The new channels will rapidly prograde, forming another site of active deposition. The abandoned delta will then be attached by the sea, and sandy islands commonly will form offshore. Thus, in this type of switching, the delta plain will be composed of alternate and overlapping lobes of deltaic distributaries that are active at different periods through geologic time. The condition is shown by the upper three diagrams in Figure 3. The Mississippi and Rurdekin deltas are examples of this type.

Channel switching

Type 2, channel switching (Figure 3), is characterized by the Niger, Lena, and Ganges-Brahmaputra. At any one time, a series of distributary channels is active at different sites within the delta plain. Development of an excessive number of bars within a channel, closure of the river mouth by heavy wave action or migrating coastal sand spits, decrease of channel gradient by excessive seaward buildout, as well as several other factors, will result in the abandonment of the channel and formation of a new one, generally in an area of more favourable gradient. Thus, the delta is constantly being prograded seaward in different parts of the delta plain simultaneously. In this case all parts of the plain are likewise active, and switching is dependent on changes in individual channels rather than in the location of the delta lobe, as in type 1.

Alternate channel extension (type 3) is commonly found

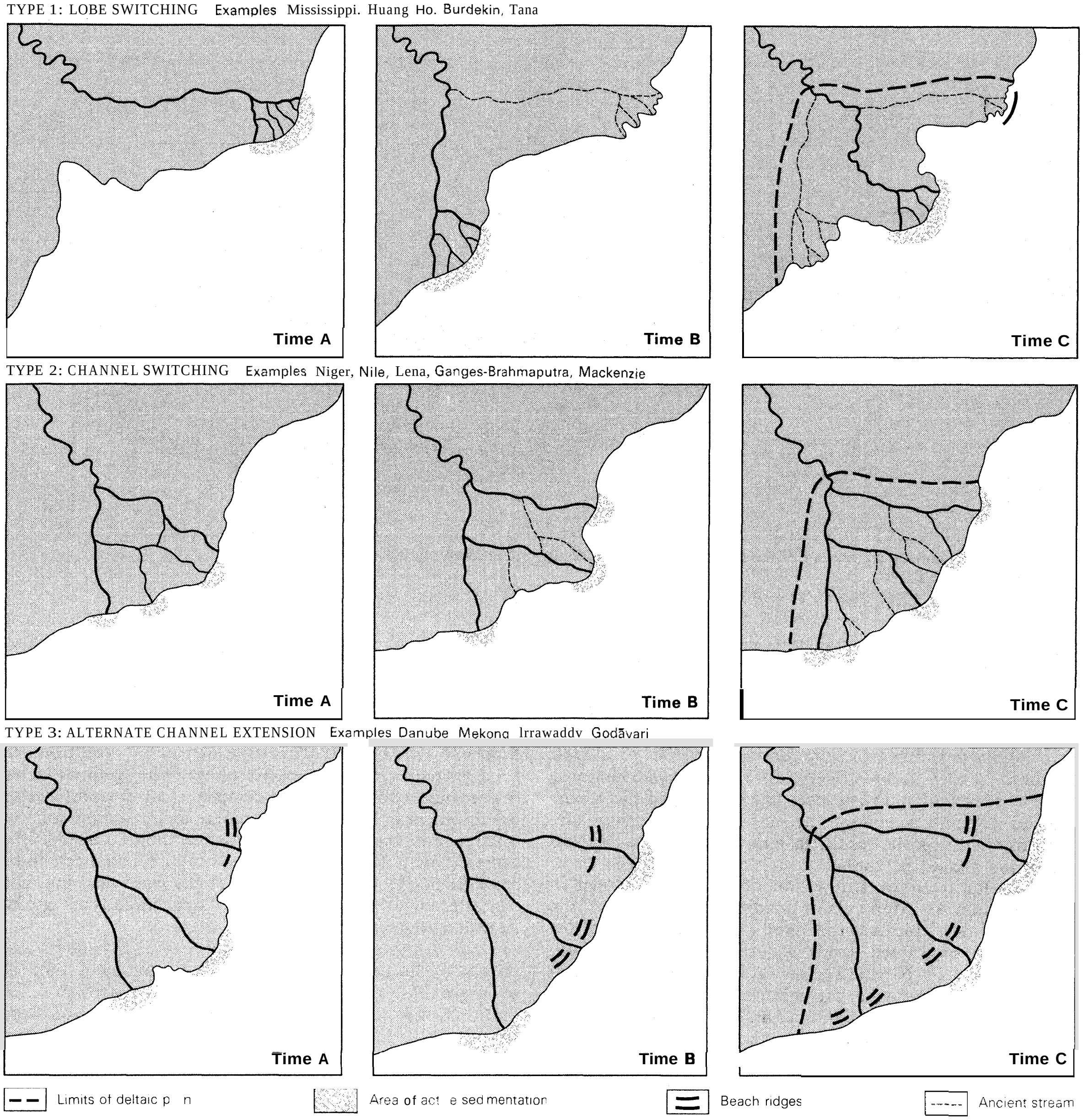

TYPE 1: LOBE SWITCHING Examples Mississippi. Huang Ho. Burdekin, Tana
Time A
Time B
Time C
TYPE 2: CHANNEL SWITCHING Examples Niger, Nile, Lena, Ganges-Brahmaputra, Mackenzie
Time A
Time B
Time C
TYPE 3: ALTERNATE CHANNEL EXTENSION Examples Danube Mekong Irrawaddy Godāvari
Time A
Time B
Time C
Limits of deltaic plain
Area of active sedimentation
Beach ridges
Ancient stream

Figure 3: Modes of growth of river deltas through time (see text).

Alternate channel extension

in deltas where stable basement conditions are encountered, such as in the Danube River Delta. In this type (Figure 3), distributary channels will remain in constant positions through relatively long periods of time and, instead of switching to different sites within the delta plain, tend to extend themselves seaward at alternate times. Those channels receiving the higher percentage of water and sediment discharge will prograde seaward; riverine processes will overpower marine destructive processes. On the other hand, channels receiving smaller volumes of water and sediment will be acted on by marine processes, normally resulting in the formation of river-mouth beach ridges. With time, however, the prograding channels will extend their mouths seaward to a point at which channel gradients become so low that dominant discharge will switch to one of the former partially inactive channels. This channel will then begin to prograde seaward, stranding the beach ridges that were developed at that river mouth. This process is illustrated diagrammatically in the lower three diagrams of Figure 3. Sites of active deposition therefore occur at different time periods within the delta plain, but in more or less specified sites.

These three types of deltaic evolution result in dynamically changing delta configuration and shoreline. The extent of changes in deltas is drastic; for example, within the past 2,000 years the Huang Ho (Yellow) River of China has shifted its delta some six times, the maximum distance between sites being in excess of 200 miles (321 kilometres). Other deltas, especially the lobe switching type, display similar changes through time. Even during periods of relative stability the shoreline is constantly changing; many deltas prograde their shorelines seaward at a rate of several hundred feet a year. Delta switching also controls rates of sedimentation within the delta plain. In various areas of active delta growth, especially in the vicinity of the river mouths, sedimentation rates are regarded as quite high, whereas in the abandoned or inactive parts of a delta accumulation rates are quite low, and in many cases erosion is active.

Deltas in the geological past

Modern deltas of various sizes and shapes are being constructed along the margins of oceans, seas, and lakes, and similar deltas were formed throughout geologic time. A lack of knowledge of the three-dimensional configuration of various types of modern deltas has hampered interpretations of ancient deltaic sequences. In spite of this shortcoming, however, deltaic sequences have been recognized in rocks of essentially all geologic ages and in many parts of the world. Among the ancient rock deltas that have been described in the geologic literature are the Blount Delta (Ordovician) in Tennessee, Queenstone Delta (Ordovician) of New York and Pennsylvania, Catskill Delta (Devonian) of Pennsylvania and New York, Old

Red Sandstone (Devonian) of Great Britain, Red Bedford Delta (Mississippian) of Ohio, Pennsylvania Coal Bearing Cycles of Kentucky, Pennsylvania, and Ohio; Rawlins Delta (Cretaceous) of Wyoming; Mesa Verde Delta (Cretaceous) of Colorado; and the Miocene (Tertiary) subsurface deltas of the Gulf Coast. Many of these deltas, especially the Tertiary deltas of the Gulf Coast, contain vast reserves of petroleum, and therefore a thorough understanding of deltaic processes and deposits is critical for future exploration. (For further information on the structures and sediments of ancient deltaic deposits see SEDIMENTARY ROCKS; and SANDSTONES.) (J.M.C.)

EXPERIMENTAL STUDY OF DELTAS

River deltas are difficult to study experimentally because of their size. In most cases it is totally impractical, and even impossible, to alter parameters such as sediment supply, sea level, or tidal patterns to observe the resulting responses. Furthermore, even if such changes could be made, it would usually take much too long to complete a meaningful experiment, perhaps several months to several years, even for quite small river deltas.

Naturally investigators have turned to model studies, which provide a cheap and fast way of doing experiments. Results obtained from models of any sort have to be treated carefully, of course, because the models may not behave exactly like their prototypes. Physical scale models, such as sedimentation tanks and flumes, have been widely used to build small-scale deltas. In only a few cases, however, has this work led to a systematic quantitative investigation, probably because of difficulties encountered in correctly scaling down the physical properties of the models.

With the development of high-speed computers, some attempts have recently been made to perform delta experiments using simulation models. Although computer-simulation models are at present limited, due to lack of knowledge of the mathematics of the processes involved in deltaic sedimentation, potentially they are enormously useful experimental devices capable of reproducing hundreds of years of deltaic sedimentation in a few seconds of computer time.

Instead of building a delta in a tank, a series of "deposits" representing the delta is formed as output from a computer program. The model can provide for inputs such as river flow, sediment supply, depth of water in the basin, and so forth. Within the program deltaic processes are represented by sequences of logical and mathematical statements that manipulate the values specified for the inputs. Finally the model prints out and automatically plots tables, graphs and other diagrams to show the form and composition of the hypothetical delta that has been produced.

A *computer* model. The first step in developing a computer model is to draw up a flow chart showing the parameters and processes to be included in the model. A computer model that deals with a part of such a flow chart, but in great detail, has been programmed for a digital computer. The model is concerned with the phase of sediment transport and deposition at the mouth of a single delta distributary. The model ignores compaction, tectonic subsidence, and changes in sea level and assumes that the depositional basin is free from all of the energy sources that originate offshore, sources such as tides, currents, and winds.

The aim of the model is to determine the rate at which sediment with different particle sizes accumulates at various positions in front of a river mouth. Specifically, what shape will the sedimentary deposit have in plan view, what will be the "foreset" slope, and how will these factors respond to changes in hydraulic and grain-size parameters? Do mouth bars and subaqueous levees form, and if so where are they to be expected?

The basic form of the model is illustrated in Figure 4. The river channel is subdivided into a number of stream segments, each containing water and sediment particles. The average sediment load in a particular stream tube is represented by a statistical particle travelling at the centre of the tube. It is assumed that in the river, statistical particles travel with the flow parallel to the bottom, but once the river mouth is reached they begin to settle downward. An accounting grid in front of the river mouth "catches" the sediment as it settles to the floor and keeps track of the rate of sediment accumulation.

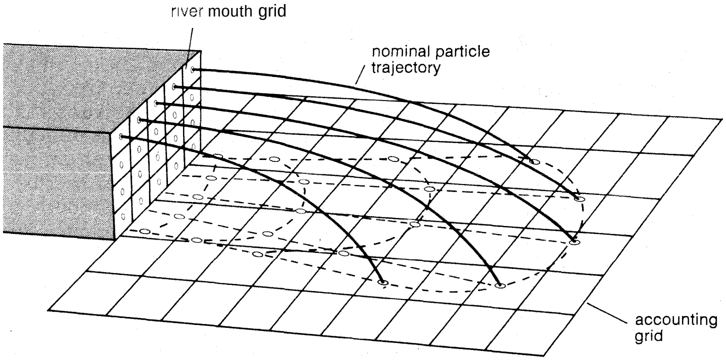


Figure 4: Structure of delta simulation model, showing vertical river grid and horizontal accounting grid. Statistical particles issuing from the centre of each river grid cell (section through a stream tube) are traced along settling trajectories until they land in accounting grid cells (see text).

From G.F. Bonham-Carter and A.J. Sutherland, "Mathematical Model and Fortran IV Program for Computer Simulation of Deltaic Sedimentation," *Computer Contribution* PC (1968); Kansas Geological Survey

The space and time variables are represented in discrete form because the model is programmed for a digital (as opposed to an analogue) computer. Thus the spatial variables are subdivided by grids that consist of a meshwork of rectangular cells. The first part of the model is static, in that only a single time step is considered. The second, or dynamic part of the model, assumes that time is represented by a series of discrete steps.

The key factor in the static model is the calculation of the settling trajectories of statistical particles. The problem is treated ballistically, each trajectory being described by the resultant of settling velocity and forward velocity. This requires knowledge of a three-dimensional velocity field. The model assumes that the flow is like a series of thin two-dimensional horizontal jets placed one on top of the other, with the initial velocity in each being given by the velocity at that elevation in the river channel. It is further assumed that the fresh river water flows out over, and floats on top of, the salty seawater, and that the two do not mix because of different densities.

With these and other assumptions, a series of experiments has been carried out with the static model to determine the effects of changing hydraulic and grain size parameters. The computer output shows that the area of sediment deposition, depicted by the positions of statistical particles in the accounting grid, is rather narrow and cigar shaped. With an arithmetic decrease in grain size, the depositional area increases geometrically, reflecting the relationship between grain size and settling velocity. More interesting is the elevation view, not shown here, in which the "foreset" slope of the deposits is about half of a degree, decreasing slightly with decrease in grain size. Considering the gross assumptions, this figure compares favourably with values for foreset slopes reported for the Grand Rhône distributary (2°), the Mississippi (1°), and the Orinoco (¼°).

Changes in the hydraulic parameters (river width, depth, and slope) produce only small differences in the same basic pattern. The mole typical fanlike shape of delta deposits is not developed. There are a number of possible reasons. First, the jet model is probably not fully appropriate for the velocity field, except possibly during periods of high river discharge when water is literally jetted out of the mouth. During periods of low discharge, the flow in real deltas appears to spread laterally more rapidly than predicted by the jet model. Second, offshore energy factors, which not considered in the model, modify the flow considerably. Third, the presence of river-mouth bars and subaqueous levees tend to disrupt the ideal flow pattern used in the model.

The model in the form described thus far sheds no light on the processes responsible for formation of levees and mouth bars. The theoretical deposits contain no irregularities suggesting the presence of these subaqueous forms. Furthermore the rate of accumulation remains constant with time in each cell of the accounting grid, a situation that rapidly leads to blockage of the river mouth. A dynamic version of the model overcomes this difficulty by cutting off sediment accumulation in any cell already filled to a specified "limiting depth." In other words, a delta platform is allowed to build out from the mouth, on which sediment can no longer settle. The computer program for the dynamic model prevents the particle trajectories from bending downward until the edge of the platform has been crossed.

Several experiments have been carried out using the dynamic model, one of which is illustrated in Figure 5. In this experiment a grain size of 0.3 millimetres (0.01 inch) diameter was employed. Figure 5A shows plots of the particle maps at the end of 2, 4, *6*, 8, and 10 time increments. The delta platform (in black) is shown to build out and gradually force the statistical particles farther from the mouth. However, the velocity field is assumed to be unchanged, and in time the particle trajectories become very steep so that at the distal end of the deposit sediment is being deposited very rapidly. This situation is shown in Figure 5A at time increments 8 and 10, where an increased density of particles can clearly be seen round the terminus of the deposit. In a vertical section parallel to the main axis (Figure 5B), successive sedimentary layers are shown, separated by time lines that simply record the sequence of sediment-water surfaces, one for each time step. The first bar produced is simply the outcome of assuming the form of the limiting depth surface. The second, or distal, bar corresponds to the high density of particles seen in Figure 5A. In further experiments in which the cells were made much smaller in relation to the channel width, thereby giving a higher resolution, deposits were formed marginal to the flow, possibly analogous to subaqueous levees.

From G.F. Bonham-Carter and A.J. Sutheriand. ''Mathematical Mode! and Fortran IV Program for Computei Simulation of Deliaic Sedimentation,'' Computer Contribution 24 (1968); Kansas Geological Survey

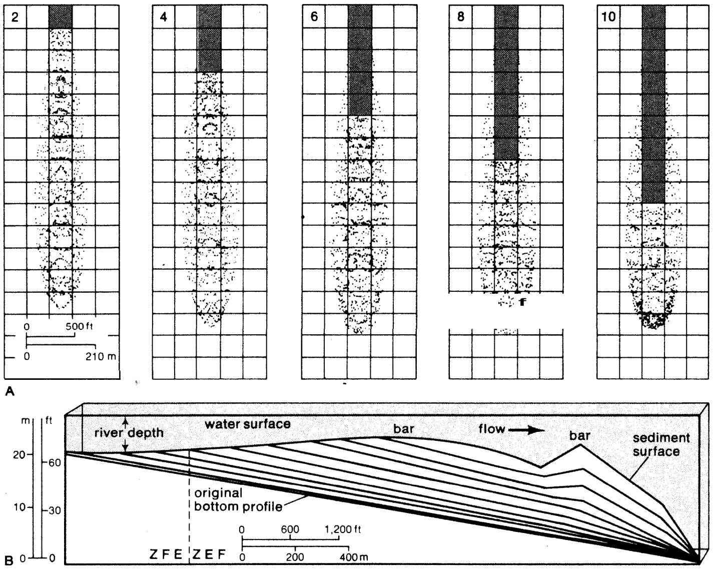


Figure 5: Computer-drawn output of results obtained from dynamic model (see text).
(A) Plan view showing build out of delta platform (solid black) during time increments 2, **4,** 6, **8,** and 10. Note the increased density of points around the distal end in increments 8 and 10.
(B) Vertical section along the centre of the delta showing the buildup of successive sediment layers. The bar on the right corresponds to the increased density of points shown in the plan view for increment 10. Note the grossly exaggerated vertical scale. ZFE and ZEF refer to flow regions in the jet model.

Value of experi-mental studies

The restrictive assumptions of this delta model make it difficult to compare in detail the theoretically produced deposits with actual deltas. Nevertheless, reasonable predictions are made for foreset slopes, and the experiments suggest that the basic shape of river-mouth deposits in plan view is not very sensitive to changes in river flow or grain-size parameters. Even though the existing dynamic model is crude in many respects, it indicates that the sediment-water surface may develop in a complex fashion when a feedback relationship between depth of water near the mouth and sediment deposition is taken into consideration.

The process of developing a simulation model forces the model builder to think very clearly about how a delta works. In order to reach the stage of writing a useful computer program, a great deal of time must first be spent in drawing together a number of theoretical and empirical relationships into a general synthesis. This focusses attention on areas in which knowledge is lacking, suggests areas for new research, and provides a framework to which new data may be related. Although very little work has yet been done on computer models of river deltas, as indicated by the sparse bibliography, it seems likely that simulation will become an increasingly important tool in delta studies. Indeed, it has become important in the geological sciences and hydrological sciences (*qq.v.*) in general. (G.B-C.)

BIBLIOGRAPHY

River deltas in general: I.V. SAMOILOV, *Die Flussmündungen* (1956), German trans. of a classic work published originally in Russian in 1952, containing a detailed discussion of deltaic processes and descriptions of some 65 river deltas, with geologic history, regional setting, geomorphology, hydrology, marine energy, and biological factors thoroughly described for each delta, including an extensive bibliography; M.L. SHIRLEY (ed.), *Deltas in their Geologic Framework* (1966). a collection of 11 papers dealing with both recent and ancient deltaic deposits, descriptions of several types of modern deltas (Mississippi, Colorado, Rhine, and Godavari) and four ancient rock deltas, and outline maps of 23 world deltas; F.P. SHEPARD *et al.* (eds.), *Recent Sediments, Northwest Gulf of Mexico* (1960), a collection of 15 papers dealing with many aspects of deltaic processes in the Mississippi River Delta; of particular interest is the article by Scruton, "Delta Building and the Deltaic Sequence," an excellent discussion of the processes involved in deltaic evolution; R.J. RUSSELL, "Geomorphology of the Rhône Delta," *Ann. Ass. Am. Geogr.*, 32:149–254 (June 1942), a classic work in purely descriptive writing on the geomorphology of a single delta; describes all aspects of geomorphology and should be read by all students interested in geomorphic studies; C.O. DUNBAR and J. RODGERS, *Principles of Stratigraphy* (1957), a general work dealing with most aspects of stratigraphy that provides a good description of several modern deltas and a few ancient rock sequences.

Experimental studies: C.C. BATES, "Rational Theory of Delta Formation," *Bull. Am. Ass. Petrol. Geol.*, 37:2119–2162 (Sept. 1953), discussion of the hypothesis that river outlets are similar to jets, and that a jet velocity field may give a reasonable approximation to velocities at river mouths; G.F. BONHAM-CARTER and A.J. SUTHERLAND, "Diffusion and Settling of Sediments at River Mouths: A Computer Simulation Model," Trans. Gulf-Cst. Ass. Geol. Soc., 17:326–338 (1967), the original publication describing the delta model described in this article, "Mathematical Model and Fortran IV Program for Computer Simulation of Deltaic Sedimentation," *Computer Contribution 24, Kansas Geol. Surv.* (1968), detailed mathematical description and listing of computer program for the model described in this article; J.W. HARBAUGH and G.F. BONHAM-CARTER, *Computer Simulation in Geology* (1970), description of the mathematical and computing tools used in developing geological computer simulation models; ch. 9 devoted to sedimentation models, including a section on the delta model discussed here; G. OERTEL and E.K. WALTON, "Lessons from a Feasibility Study for Computer Models of Coal-Bearing Deltas," *Sedimentology*, vol. 9, no. 2, pp. 157–168 (1967), discussion of the author's conclusion that it is not yet possible to set up a computer model that would be sufficiently detailed for performing useful experiments on coal-bearing deltas; L.M.J.U. VAN STRAATEN, "Some Recent Advances in the Study of Deltaic Sedimentation," *Lpool. Manch. Geol. J.*, vol. 2, pt. 3, pp. 411–442 (1960), a useful review arlicle.

(J.M.C./G.B-C.)

Rivers and River Systems

By original usage, a river is flowing water in a channel with defined banks (ultimately from Latin *ripa*, "bank"). Modern usage includes rivers that are multichannelled, intermittent, or ephemeral in flow and channels that are practically bankless. The concept of channelled surface flow, however, remains central. The word stream (ultimately from the Indo-European root ***srou-)*** emphasizes the fact of flow; as a noun it is synonymous with river and is often preferred in technical writing. Small natural watercourses are sometimes called rivulets, but branch, brook, burn, and creek are more common, occurring regionally to nationaliy in piace-names. Arroyo and (dry) wash connote ephemeral streams or their resultant channels. Tiny streams or channels are known as rills or runnels.

Rivers are nourished by precipitation, by direct overland runoff, through springs and seepages, or from meltwater at the edges of snowfields and glaciers. The contribution of direct precipitation on the water surface is usually minute, except where much of a catchment area is occupied by lakes. River water losses result from seepage and percolation into shallow or deep aquifers (water-bearing layers), and particularly from evaporation. The difference between the water input and loss sustains surface discharge or streamflow. The amount of water in river systems at any time is but a tiny fraction of the earth's total water; 97 percent of all water is contained in the oceans and about three-quarters of fresh water is stored as land ice; nearly all the remainder occurs as groundwater. Lakes hold less than 0.5 percent of all fresh water, soil moisture accounts for about 0.05 percent, and water in river channels for about half as much, 0.025 percent, which is one four-thousandth of the earth's total fresh water.

Water is constantly cycled through the systems of land ice, soil, lakes, groundwater (in part), and river channels, however. The discharge of rivers to the oceans delivers to these systems the equivalent of the water vapour that is blown overland and then consequently precipitated as rain or snow, that is, some 7 percent of mean annual precipitation on the globe, and 30 percent of precipitation on land areas.

Prehistoric importance of rivers

The inner valleys of some great alluvial rivers contain the sites of ancestral permanent settlements, including pioneer cities. Sedentary settlement in Hither Asia began about 10,000 BP (years before present) at the site of Arīhā (ancient Jericho). Similar settlement in the Tigris-Euphrates and Nile valleys dates back to at least 6000 BP. The first settlers are thought to have practiced a hunting economy, supplemented by harvesting of wild grain. Conversion to the management of domesticated animals and the cultivation of food crops provided the surpluses that made possible the rise of towns, with parts of their populations freed from direct dependence on food getting. Civilization in the Indus Valley, prominently represented at Mohenjo-daro, dates from about 4500 BP, civilization in the Ganges Valley from about 3000 BP. Permanent settlement in the valley of the Huang Ho (Yellow River) has a history 4,000 years long, and the first large irrigation system in the Yangtze catchment dates back to roughly the same time. Greek invaders of the Syrdarya, Amu Darya, and other valleys draining to the Aral Sea, east of the Caspian, encountered irrigating communities that had developed from about 2300 BP onward.

The influence of climatic shifts on these prehistoric communities has yet to be worked out. In wide areas, these shifts included episodic desiccation from 12,000 or 10,000 BP onward; in what are now desert environments, increased dependence on the rivers may have proved as much a matter of necessity as of choice. All of the rivers concerned have broad floodplains subject to annual inundation by rivers carrying heavy sediment loads. Prehistoric works of flood defense and irrigation demanded firm community structures and required the development of engineering practice. Highly elaborate irrigation works are known from Mohenjo-daro; the ziggurats (temple mounds) of the Euphrates Valley may well have originated in ancient Egypt in response to the complete annual inundation of the Nile floodplain, where holdings had to be redefined after each flood subsided. It is not surprising that the communities named have been styled hydraulic civilizations. But it would be oversimplistic to claim that riparian sites held the monopoly of the developments described: elaborate urban systems arising in Mexico, Peru, and the eastern Mediterranean from about 4000 BP onward were not immediately dependent on the resources of livers.

Where riverine cities did arise, they commanded ready means of communication; the Two Lands of Upper and Lower Egypt, for instance, were unified by the Nile. At the same time, it can be argued that eariy riverine and river-dependent civilizations bore the seeds of their own destruction, independent of climatic shifts and natural evolutionary changes in the river systems. High-consuming cities downstream inevitably exploited the upstream catchments, especially for timber. Deforestation there may possibly have led to ruinous silting in downstream reaches, although the contribution of this to the eventual decline of civilization on the Euphrates and the Indus remains largely a matter of guesswork. An alternative or conjoint possibility is that continued irrigation promoted progressive salinization of the soils of irrigated lands, eventually preventing effective cropping. Salinization is known to have damaged the irrigated lands of Ur, progressively from about 4400 to 4000 BP, and may have ruined the Sumerian Empire of the time; but the relative importance of environmental and social deterioration in prehistoric hydraulic civilizations remains a matter of debate. Furthermore, defective design and maintenance of irrigation works promote the spread of malarial mosquitoes, which certainly afflicted the prehistoric hydraulic communities of the lower Tigris-Euphrates Valley; these same communities may also have been affected by bilharziasis (blood fluke disease), which requires a species of freshwater snail for propagation and which even today follows many extensions of irrigation into arid lands.

Historical and present importance

In various intervals of history, rivers have provided the easiest, and in many areas the only, means of entry and circulation for explorers, traders, conquerors, and settlers. They assumed high importance in Europe after the fall of the Roman Empire and the dismemberment of its roads; regardless of political structures, control of crossing points was expressed in strongholds· and the rise of bridge-towns. Rivers in medieval Europe dominantly supplied the water that could sustain cities and the sewers that could carry away city waste and were widely used, either directly or with offtakes, as power sources. Western European history records the rise of 13 national capitals on sizable rivers, exclusive of seawater inlets; three (Vienna, Budapest, and Belgrade) lie on the Danube, with two others, Sofia and Bucharest, on feeder streams above stem floodplain level. The location of provincial and corresponding capitals is even more strongly specialized on riparian sites, as is readily illustrated from the United Kingdom, France, and Germany. In modern history, both in North America and northern Asia, natural waterways directed the lines of exploration, conquest, and settlement: in both areas, passage from system to system was facilitated by portage along lines defined by temporary ice-marginal or ice-diverted channels. Many pioneer settlers of the North American interior entered by means of natural waterways, especially in Ohio.

The historical record includes marked shifts in the appreciation of rivers, numerous conflicts in use demand, and an intensification of use that has rapidly accelerated in the present century. External freight trade concentrated in estuarine ports rather than inland ports when ocean craft increased in size; even the port of London, though constrained by high capital investment, has displaced itself toward its estuary. The Amazon remains naturally navigable by ocean ships for 3,700 kilometres (2,300 miles), the Yangtze for 1,000 kilometres (625 miles), and the partly artificial St. Lawrence Seaway for 2,100 kilometres (1,300 miles). Internal freight traffic on the Rhine system and its associated canals amounts to one-quarter or more of total traffic in the basin, and to more than half in some parts. Waterborne freight movement currently is recovering much of the ground lost to railways in the later 19th and the first half of the present century, notably in interior America and the northern plainlands of the U.S.S.R. Extensive commercial navigation, however, usually implies much artificial improvement of natural channels, allied to efforts to bring variations of regime under some control. The nearly 145,000 kilometres of navigable natural waterway in the U.S.S.R., for example, are mainly subject to summer floods.

Use for irrigation and hydroelectric power

Demand on open-channel water incteases as population and per capita water use increase and as underground water supplies fall short. Domestic per capita use, foı example, already has reached about 200 gallons per day in the U.S. and 100 gallons per day in the U.K. Irrigation use undoubtedly constitutes the bulk of total irrigation supply. With a history of at least 5,000 years, controlled irrigation now affects nearly 750,000 square miles of land (1,940,000 square kilometres), three-quarters of it in East and South Asia and two-fifths in mainland China alone. Most of this involves natural floodwater, although reliance on artificially impounded storage is fast increasing; irrigation in the 1,300-kilometre (700-mile) length of the Indus Valley, for instance, depends almost exclusively on barrages (*i.e.*, distributor canals) running down alluvial fans and along floodplains. Rice is the great crop of seasonally irrigated land; perennial irrigation inserts additional crops into the rotation or makes possible highly specialized commercial cropping, as in Australia, which, with about one-fourth of an acre of irrigated land per capita, leads the world in this respect among developed and intensively irrigated countries. Modern demands on rivers as power sources range from the floating of timber, through the use of water for cooling, to hydroelectric generation. Logging in forests relies primarily on flotation during the season of meltwater high flow. Large thermal generating stations are typically sited on rivers, which supply the huge quantities of water needed for cooling; the demand is as great as 1,000 tons of water for every ton of coal burned. Some other types of industrial installation make corresponding demands ranging from 250 to 600 tons of water per ton of steel, wood pulp, or woollen cloth. Hydroelectric power generation has a history about a century long, but the bulk of the existing installations have been built since 1950. All great industrial countries, regardless of their other power supplies, have developed their hydropower—in Europe, to about 30 percent of potential, and in North America, to about 20 percent. Norway and Switzerland are among countries depending almost solely on hydropower for internally generated energy. Capacity of major plants now existing is 6,000,000 kilowatts and the level is still rising. Potential supplies are greatest in Africa and South America (combined, about half the world potential and 1 percent of world developed hydropower). Use demand of more immediate kinds relate to freshwater fisheries (including fish farming) and to dwelling in houseboats. Reliable data for these kinds of dependence on rivers do not exist; published estimates that freshwater and migratory fish provide up to some 15 percent of world catch could be too low. In parts of tropical Asia, the yield of freshwater fish runs up to 1,000 pounds per acre (1,120 kilograms per hectare) and 75 percent of national take, whereas Denmark produces about half the world's trout by pisciculture. Certainly, millions of people are concerned with freshwater fishery and houseboat living, principally in the deltaic areas of eastern Asia, where dwelling, marketing, and travel can be located almost exclusively on the water.

Political observers have recorded the difficulties entailed in using rivers as boundaries or drainage basins as areal entities. Multipurpose works and the reconciliation of conflicting use demands cause difficulties even within a single political unit, as is illustrated by the controversy attending the Tennessee Valley project in the U.S. and the continued rarity of interbasin transfer of water in that country today. Very large scale integrated projects frequently involve more than one country, as with the Columbia River scheme and the St. Lawrence Seaway project. The general tendency seems to be toward international cooperation in water use, even between politically inimical countries. The present extreme of technical possibility is illustrated by the Kariba and Aswan High dams, each of which impounds enough water to cover 200,000 square miles (518,000 square kilometres) a foot deep, permits a carryover of storage, and allows considerable control over discharge regime. Projections of water demand, however, indicate that by the year 2000 the U.S. will expend 75 percent of supply on irrigation, southeast England will expend 50 percent on industrial use, and comparable percentages will be required in Asiatic U.S.S.R. These figures suggest that regional projects of river management that at present seem grandiose will in another generation or so become both necessary and ordinary.

Geological significance

Rivers are 100 times more effective than coastal erosion in delivering rock waste to the sea. Their rate of sediment delivery is equivalent to an average lowering of the lands by one foot in 9,000 years, a rate that is sufficient to remove all the existing continental relief in 25,000,000 years.

Rock waste enters fluvial systems either as fragments eroded from rocky channels or in dissolved form. During transit downstream, the solid particles undergo systematic changes in size and shape, travelling as bed load or suspension load. Except in high latitudes and on steep coasts generally, little or no coarse bed load reaches the sea. Downvalley movement of the solid load is irregular, both because streamflow is irregular and because the material is liable to enter temporary storage, forming distinctive river-built features that range through riffles, midstream bars, point bars, floodplains, levees, alluvial fans, and river terraces. In one sense, such features belong to the same series as deltas, estuary fills, and the terrestrial sediments of many inland basins.

Rates of erosion and transportation, and comparative amounts of solid and dissolved load, vary widely from river to river. Least is known about dissolved load, which at coastal outlets is added to oceanic salt. Its concentration in tropical rivers is not necessarily high, although very high discharges can move large amounts; the dissolved load of the lowermost Amazon averages about 40 parts per million, whereas the Elbe and the Rio Grande average more than 800 parts per million. Suspended load for the world in general perhaps equals two and one-half times dissolved load; well over half of it is deposited at river mouths as deltaic and estuarine sediment. About one-quarter of all suspended load is estimated to come down the Ganges–Brahmaputra and the Huang Ho, which together deliver some 4,500,000,000 tons a year; the Yangtze, Indus, Amazon, and Mississippi deliver quantities ranging from about 500,000,000 to about 350,000,000 tons a year. Suspended sediment transport on the Huang Ho equals a denudation rate of about 3,090 tons per square kilometre (8,000 tons per square mile) per year; the corresponding rate for the Ganges–Brahmaputra is almost half as great. Extraordinarily high rates are recorded for some lesser rivers: for instance, 1,060 tons per square kilometre (2,750 tons per square mile) per year on the Ching and 1,080 tons per square kilometre (2,800 tons per square mile) per year on the Lo, loess-plateau tributaries of the Huang Ho.

This article treats the distribution, patterns, and geometry of steamflow in nature, as well as floods, fluvial landforms, and life in river systems. For further details of relevance on the action of flowing water on the Earth's surface, see FLUVIAL PROCESSES; SEDIMENT YIELD OF DRAINAGE SYSTEMS; and LANDFORM EVOLUTION. See also RIVER DELTAS and ALLUVIAL FANS, sedimentary deposits that occur wherever the competence of fluvial transport diminishes; CANYONS, SUBMARINE, analogues of terrestrial canyons that are cut in the world's continental shelves and slopes by sediment-laden density currents (*q.v.*); and WATERFALLS, which occur at breaks in the longitudinal profile of a given river. Certain aspects of the changes in river systems through time are treated in the articles PLEISTOCENE EPOCH; HOLOCENE EPOCH; and CLIMATIC CHANGE; and the general interrelationship of rivers with other parts of the Earth's hydrosphere is covered in HYDROLOGIC CYCLE.

DISTRIBUTION OF RIVERS IN NATURE

World's largest rivers. Obvious bases by which to compare the world's great rivers include the size of the drainage area, the length of the main stem, and the mean discharge; but reliable comparative data, even for the

The World's Principal Rivers, Ranked According to Drainage Area

river	drainage area			length		mean discharge					
	extent		percent of world's land area	mi	km	(000 cu ft/sec)	(000 cu m/sec)	rank order	percent of world total	cu ft/ sec/sq mi	cu m/ sec/sq km
	(000 sq mi)	(000 sq km)									
Amazon	2,722	7,050	4.8	4,000	6,437	6,350	180	1	19.2	2.33	.0255
Rio de la Plata-Paraná	1,600	4,144	2.8	2,485	4,000	777	22	5	2.3	0.48	.0052
Congo	1,314	3,457	2.3	2,914	4,700	1,458	41	2	4.4	1.11	.0121
Nile	1,293	3,349	2.3	4,132	6,650	110	3	—	0.3	0.09	.0009
Mississippi-Missouri	1,244	3,221	2.2	3,741	6,020	650	18	8	2.0	0.52	.0057
Ob-Irtysh	1,149	2,975	2.0	3,362	5,410	558	15	10	1.7	0.49	.0053
Yenisey	996	2,580	1.7	3,442	5,540	671	19	6	2.0	0.67	.0073
Lena	961	2,490	1.7	2,734	4,400	575	16	9	1.7	0.60	.0065
Yangtze	756	1,959	1.3	3,434	5,494	1,200	34	4	3.6	1.59	.0174
Niger	730	1,890	1.3	2,600	4,180	215	6	—	0.7	0.29	.0032
Amur	716	1,855	1.3	1,755	2,824	438	12	10	1.3	0.61	.0066
Mackenzie	711	1,841	1.2	2,635	4,241	400	11	—	1.2	0.56	.0061
Ganges-Brahmaputra	626	1,621	1.1	1,800	2,897	1,360	38	3	4.1	2.17	.0237
St. Lawrence-Great Lakes	565	1,463	1.0	2,500	4,023	360	10	—	1.1	0.64	.0069
Volga	525	1,360	0.9	2,293	3,690	282	8	—	0.9	0.54	.0058
Zambezi	514	1,330	0.9	2,200	3,540	251	7	—	0.8	0.49	.0053
Indus	450	1,166	0.8	1,790	2,880	194	5	—	0.6	0.43	.0047
Shatt al-Arab (Tigris-Euphrates)	430	1,114	0.8	1,700	2,740	49	1	—	0.1	0.11	.0012
Nelson	414	1,072	0.7	1,600	2,570	81	2	—	0.2	0.20	.0021
Murray-Darling	408	1,057	0.7	2,350	3,780	14	0.4	—	0.04	0.04	.0003
Tocantins	350	906	0.6	1,000	1,610	360	10	—	1.1	1.03	.0112
Danube	315	816	0.6	1,770	2,850	254	7	—	0.8	0.81	.0088
Columbia	258	668	0.5	1,210	1,950	247	7	—	0.7	0.96	.0104
Rio Grande	172	445	0.4	1,885	3,040	3	0.08	—	0.01	0.02	.0001
Rhine	62	160	0.1	820	1,320	78	2	—	0.2	1.26	.0137
Rhône	37	96	—	500	800	60	2	—	0.2	1.62	.0177
Thames	4	10	—	210	340	3	0.08	—	0.01	0.75	.0082

world's greatest rivers, do not exist. Some of the values listed in the Table are approximate. The Nile, the world's longest river, is about 130 miles longer than the Amazon, according to latest revised figures. It is possible that well over 100 of the greatest rivers may exceed a 1,600-kilometre (1,000-mile) length on their main stems.

Drainage area, length, and discharge

Area–length–discharge combinations vary considerably, although length tends to increase with area and area and discharge to increase through their individual ranking series. On all counts except length, the Amazon is the world's principal river; the Congo and the La Plata–Parana are among the first five by area and discharge, but the Mississippi, third in length and fifth in area, is only seventh in discharge. The Ganges–Brahmaputra, third in discharge, is 13th (or lower) in area and well down the list of length for its two main stems taken separately.

Ranking in the Table is by drainage area. In combination, the rivers listed drain some 44,000,000 square kilometres (17,000,000 square miles), 30 percent of the world's land area. If volume of discharge is taken to be the basis of comparison, then certain other rivers not tabulated above must also be mentioned. The most important of these is the Orinoco, mean discharge 19,800 cubic metres (700,000 cubic feet) per second and a basin of 948,000 square kilometres (360,000 square miles). Others are the Irrawaddy, discharge 13,000 cubic metres (460,000 cubic feet) per second, basin 411,000 square kilometres (159,500 square miles); the Mekong, 11,000 cubic metres (390,000 cubic feet) per second, basin 795,000 square kilometres (306,000 square miles); the St. Lawrence, 10,200 cubic metres (360,000 cubic feet) per second, 1,463,000 square kilometres (565,000 square miles); and the Niger, 6,100 cubic metres (215,000 cubic feet) per second, 1,890,000 square kilometres (730,000 square miles). The 20 greatest of these rivers, draining about 30 percent of the world's land area, discharge nearly 40 percent of total runoff, reckoned from a mean equivalent of 29.2 centimetres (11.5 inches) of precipitation. They deliver to the sea about 92 cubic kilometres (22 cubic miles) of water per day, about 33,325 cubic kilometres (8,000 cubic miles) a year. The Amazon, the La Plata–Parana, the Congo, and the Ganges–Brahmaputra, combined, discharge more than 54 cubic kilometres (13 cubic miles) a day and nearly 20,800 cubic kilometres (5,000 cubic miles) a year, one-third of the world's total runoff to the oceans, the Amazon alone accounting for almost one-fifth.

World average external runoff is about 0.01 cubic metre per second per square kilometre (0.6 cubic foot per second per square mile). Great rivers with notably higher discharges are fed either by the convectional rains of Equatorial regions or by monsoon rains that are usually increased by altitudinal effects. The Huang Ho averages 0.046 cubic metre per second per square kilometre (4.25 cubic feet per second per square mile), the Irrawaddy 0.032 cubic metre per second per square kilometre (2.90 cubic feet per second per square mile), the Magdalena and the Amazon 0.026 cubic metre per second per square kilometre (2.40 cubic feet per second per square mile), the Orinoco 0.21 cubic metre per second per square kilometre (1.91 cubic feet per second per square mile), and the Ganges–Brahmaputra, whose mean discharge is above 0.023 cubic metre per second per square kilometre (2.10 cubic feet per second per square mile). Very high mean discharges per unit area are also recorded for lesser basins in mountainous coastlands exposed to the zonal westerlies of midlatitudes. Among great rivers with mean discharges near or not far below world averages per unit area are those of Siberia, the Mackenzie, and the Yukon (828,000 square kilometres, 5,900 cubic metres per second), all affected by low precipitation for which low evaporation rates barely compensate. The basins of the Mississippi, Niger, and Zambezi include some areas of dry climate. The Nelson illustrates the extreme effects of low precipitation in a cool climate, while the Nile, Murray–Darling, and Shatt al-Arab (Tigris–Euphrates) experience low precipitation combined with high evaporation losses.

The lower end of the Table lists comparative data for selected rivers in highly inhabited or otherwise hydrographically interesting valleys. The Rhine, Rhône, and Danube record regimes that vary along the length of their courses, in response to glacier melt in the headwaters and the entry of contrasting tributaries downstream. The Columbia is noteworthy as incorporated in an international project of regulation and power generation; the Rio Grande, like the Orange and the Colorado, suffers progressive downstream losses, both natural and irrigational. The Thames is special, as it experiences a very high tidal range in its estuary; this results in a number of benefits to navigation but also considerable handicaps in flood control.

Principles **governing** distribution and **flow.** Moisture supply sufficient to sustain channelled surface flow is governed primarily by climate, which regulates precipitation, temperature, and evapotranspiration water loss caused by vegetation. In rainy tropical and exposed midlatitude areas, runoff commonly equals 38 centimetres (15 inches) or more of rain a year, rising to more than 102 centimetres (40 inches). Negligible external runoff occurs in subtropical and rain-shadow deserts; perennial, intermittent, and ephemeral lakes, expanding in response to local runoff, prevent the drainage of desert basins from finding escape routes.

Variation of stream regime

Seasonal variation in discharge defines river regime. Three broad classes of regime can be distinguished for perennial streams. In the megathermal class, related to hot equatorial and tropical climates, two main variants occur; discharge is powerfully sustained throughout the year, usually with a double maximum (two peak values), but in some areas with a strong single maximum. In the mesothermal class, some regimes resemble those of tropical and equatorial areas, with single or double summer maxima corresponding to heavy seasonal rainfall, while others include sustained flow with slight warm-season minima. Where midlatitude climates include dry summers, streamflow decreases markedly and may cease altogether in the warm half of the year. In areas affected by release of meltwater, winter minima and spring maxima of discharge are characteristic. Microthermal regimes, influenced by snow cover, include winter minima and summer maxima resulting from snowmelt and convectional rain; alternatively, spring meltwater maxima are accompanied by secondary fall maxima associated with late-season thunder rain, or spring snowmelt maxima can be followed by a summer glacier-melt maximum, as on the Amu Darya. Megathermal regimes, controlled by systematic fluctuations in seasonal rain, and microthermal regimes, controlled by seasonal release of meltwater, may be more reliable than mesothermal regimes.

The regime can vary considerably along the length of a single river in timing and in seasonal characteristics. Spring maxima in the Volga headwaters are not followed by peak flows in the delta until two months later. The October seasonal peak on the upper Niger becomes a December peak on the middle river; the swing from tropical-rainy through steppe climate reduces volume by 25 percent through a 483-kilometre (300-mile) stretch. The seasonal headwater flood wave travels at 0.09 metre (0.3 foot) per second, taking some four months over 2,011 kilometres (1,250 miles), but earlier seasonal peaks are re-established on the lower river by tributaries fed by hot-season rains. The great Siberian rivers, flowing northward into regions of increasingly deferred thaw, habitually cause extensive flooding in their lower leaches, which remain ice-covered when upstream reaches have already thawed and are receiving the meltwaters of late spring and summer.

Extremes of regime characteristics come into question when streams are classified as perennial, intermittent, or ephemeral. These terms are in common use but lack rigid definition. Whereas the middle and lower reaches of streams in humid regions rarely or never cease flowing and can properly be called perennial, almost every year many of their upstream feeders run dry where they are not fed by springs. In basins cut in impermeable bedrock, prolonged droughts can halt flow in most channel reaches. Karst (limestone country) that has some surface drainage often includes streams that are spatially intermittent; frequently it also contains temporally intermittent streams that flow only when heavy rain raises the groundwater table and reactivates outlets above the usual level. Temporally intermittent streams also occur in dry areas where, at low stage, only some channel reaches contain flowing water. There is a continuous progression from perennial streams through intermittent streams to ephemeral streams: the latter flow only in response to occasional storms, their channels remaining dry most of the time. Supplying storms may be more or less randomly distributed or may belong to the outermost limits of seasonal monsoon rains.

Long-term effects expressed in mean seasonal regimes and short-term effects expressed in individual peak flows are alike affected by soil-moisture conditions, groundwater balance, and channel storage. Channelled surface flow begins when overland flow becomes deep enough to be erosive; and depth of overland flow represents a balance between short-term precipitation and soil infiltration. Rate and capacity of infiltration depend partly on antecedent conditions and partly on permeability. Seasonal assessments are possible, however; numbers of commercial crops can take up and transpire the equivalent of 38 centimetres (15 inches) of precipitation during the growing season. In many midlatitude climates, the rising curves of insolation and plant growth during spring and early summer cause soil moisture depletion, leading eventually to a deficit that is often strong enough to reduce runoff and streamflow. Soil-moisture recharge during colder months promotes high values of runoff, frequently in the spring, quite independently of the influence of precipitation regime or snowmelt.

Groundwater and glacier effects

Storage of water in groundwater tables, stream channels, on floodplains, and in lakes damps out variations in flow, whereas snow and ice storage exaggerate peaks. For the world as a whole, groundwater contributes perhaps 30 percent of total runoff, although the proportion varies widely from basin to basin, within basins, and through time. Shallow groundwater tables in contact with river channels absorb and release water, respectively at high and low stage. Percolation to greater depths and eventual discharge through springs delays the entry of water into channels; many groundwater reservoirs carry over some storage from one year to another. Similar carryover occurs with glaciers and to some extent also with permanent snowfields; water abstracted by the icecaps of high latitudes and by large mountain glaciers can be retained for many years, up to about 250,000 years in the central Antarctic cap. Temperate glaciers, however, with temperatures beneath the immediate subsurface constantly near the freezing (or the melting) point, can, like their associated snowfields, release large quantities of water during a given warm season; their losses through evaporation are small.

Meltwater contributions to streamflow, however, can range from well above half the total discharge to well below the level of the snowline; they are vital to irrigation on alluvial fans rimming many dry basins, as in the Great Valley of California and the Tarim Basin of the Takla Makan Desert: meltwater is released during the planting or growing seasons. Within the limiting constraints of precipitation or meltwater input or both, and the outputs of evapo-transpiration and percolation, the actual distribution of rivers in nature is affected by available drainage area, lithology, and vegetation. Vegetation is obviously climate dependent to a large extent but might well be capable of reaching thresholds of detention ability that do not match recognized climatic boundaries: it is, moreover, liable to the influence of climatically independent factors where it has been disturbed by man. Runoff on the plainlands of northern Asia, expressed as a percentage of mean annual precipitation, ranges from about 75 in the tundra, through about 70 in the boreal forest and 50 through boreal forest with perennially frozen ground, down through less than 40 in mixed forest, to five in semidesert. Clear felling of forest increases runoff in the short and medium term because it reduces surface detention and transpiration. In areas of seasonal snow cover, forest considerably influences seasonal regime. But although there may be a jump in short-term runoff characteristics between areas of continuous vegetation (forest and grass sward) on the one hand and discontinuous vegetation (bunch grass and scrub) on the other, comprehensive general studies of precipitation–temperature-runoff characteristics suggest that mean annual runoff decreases, at a decreasing rate through the range that is involved, as temperature increases and as precipitation (weighted in respect of seasonal incidence) decreases.

Constraints of rock type and vegetation

Lithology is significant mainly in connection with permeability. The capacity of karst or limestone country to swallow and to reissue water is well-known, as is the role of permeable strata generally in absorbing water into groundwater tables. An extreme case of a special kind is represented by an artesian aquifer, which in favourable structural conditions can take water for a very long time from the surface and immediately connected circulations, returning it only if the artesian pressure becomes strong enough to promote the opening of flowing springs. Less directly, but with considerable effect or, infiltration and short-term runoff, the mechanical grade of bedrock or of surficial deposits can considerably affect the response to individual storms.

By courtesy of Mount Salus Press Limited, Dublin: photograph, K.G. Govan

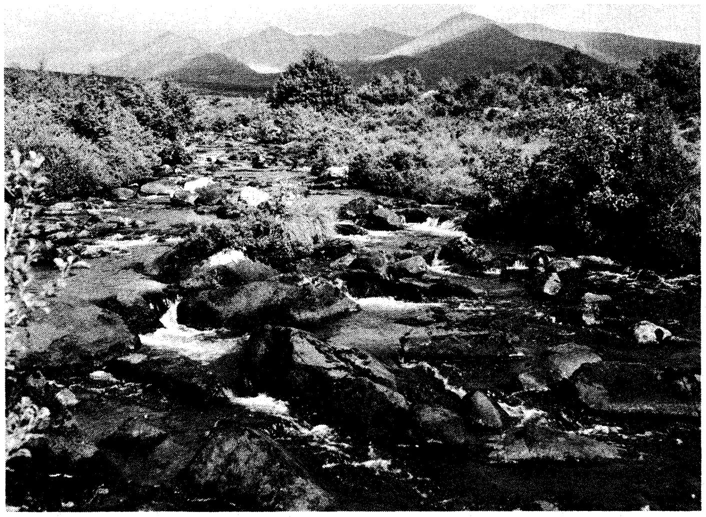


Figure 1: Boulder-bed stream, Ireland.

Both the ultimate possible extent of drainage basins and the opening of individual headwater channels are influenced by available drainage area. A hypothetical limit for very large basins could probably be constructed from considerations of stem length, basin shape, computed area, and continental extent: the Amazon probably approaches the hypothetical maximum. At the other extreme, basin morphometry (geometrical aspects of basins and their measurement) can be made to indicate the limiting average area necessary to sustain a given length of channel; in large areas of midlatitudes, the ratio is close to 2.25 square kilometres (1.4 square miles) of drainage area for a channel 1.6 kilometres (one mile) in length. Estimates for the conterminous United States, an area of about 7,770,000 square kilometres (3,000,000 square miles), give some 5,230,000 kilometres (3,250,000 miles) of channel length. These estimates include 1,500,000 unbranched fingertip tributaries—each having an average length of between 1.6 and 2.4 kilometres (1.0 and 1.5 miles).

DRAINAGEPATTERNS

Distinctive patterns are acquired by stream networks in consequence of adjustment to geologic structure. In the early history of a network, and also when erosion is reactivated by earth movement or a fall in sea level, downcutting by trunk streams and extension of tributaries are most rapid on weak rocks, especially if these are impermeable, and along master joints and faults. Tributaries from those streams that cut and grow the fastest encroach on adjacent basins, eventually capturing parts of the competing networks therein. In this way, the principal valleys with their main drainage lines come to reflect the structural pattern.

Dendritic, trellis, and radial patterns

Flat-lying sedimentary rocks devoid of faults and strong joints and the flat glacial deposits of the Pleistocene Epoch (from 2,500,000 to 10,000 years ago) exert no structural control at all: this is reflected in branching networks. A variant pattern, in which trunk streams run subparallel, can occur on tilted strata. Rectangular patterns form where drainage lines are adjusted to sets of faults and marked joints that intersect at about right angles, as in some parts of ancient crustal blocks; the pattern is varied where the regional angle of structural intersection changes. Radial drainage is typical of volcanic cones, so long as they remain more or less intact. Erosion to the skeletal state often leaves the plug standing in high relief, ringed by concentric valleys developed in thick layers of ash.

Similarly, on structural domes where the rocks of the core vary in strength, valleys and master streams locate on weak outcrops in annular patterns. Centripetal patterns are produced where drainage converges on a single outlet or sink, as in some craters, eroded structural domes with weak cores, parts of some limestone country, and enclosed desert depressions. Trellis (or espalier) drainage patterns result from adjustment to tight regional folding in which the folds plunge. Denudation produces a zigzag pattern of outcrops, and adjustment to this pattern produces a stream net in which the trunks are aligned on weak rocks exposed along fold axes and small feeder streams run down the sides of ridges cut on the stronger formations. Deranged patterns, in which channels are interrupted by lakes and swamps, characterize areas of modest relief from which continental ice has recently disappeared; these patterns may be developed either on the irregular surface of a till sheet (heterogeneous glacial deposit) or on the ice-scoured expanse of a planated crystalline block. Where a till sheet has been molded into drumlins (inverted-spoon-shaped forms that have been molded by moving ice), the postglacial drainage can approach a rectangular pattern. In glaciated highland, postglacial streams can pass anomalously through gaps if the divides have been breached by ice, and sheet glaciation of lowland country necessarily involves major derangement of river networks near the icefront. At the other climatic extreme, organized networks in dry climates can be deranged by desiccation, which breaks down the existing continuity of a net; the largely linear systems of ephemeral lakes in inland Western Australia have been referred to this process.

Spence Air Photos

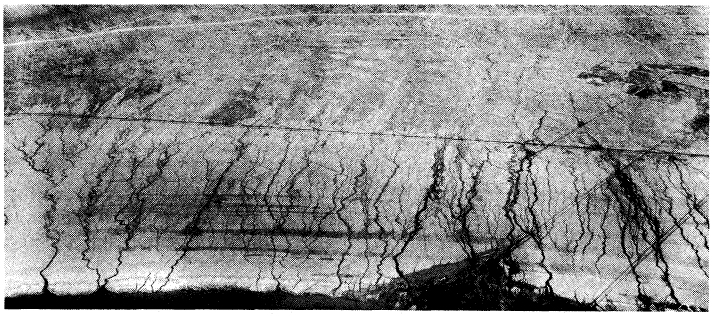


Figure 2: Dendritic drainage patterns in alluvium, Imperial Valley, California.

Adjustment to bedrock structure can be lost if earth movement raises folds or moves faults across drainage lines without actually diverting them; streams that maintain their courses across the new structures are called antecedent. Adjustment is lost on a regional scale when the drainage cuts down through an unconformity into an undermass with structures differing greatly from those of the cover: the drainage then becomes superimposed. Where the cover is simple in structure and provides a regional slope for trunk drainage, remnants of the original pattern may persist long after superimposition and the total destruction of the cover, providing the means to reconstruct the earlier network.

Stream order, stream number, and stream length

Horton's laws of drainage composition. Great advances in the analysis of drainage nets were made by R.E. Horton, an American hydraulic engineer who developed the fundamental concept of stream order: An unbranched headstream is designated as a first-order stream. Two unbranched headstreams unite to form a second-order stream; two second-order streams unite to form a third-order stream, and so on. Regardless of the entry of first- and second-order tributaries, a third-order stream will not pass into the fourth order until it is joined by another third-order confluent. Stream number is the total number of streams of a given order for **a** given drainage basin. The bifurcation ratio is the ratio of the number of streams in a given order to the number in the next higher order. By definition, the value of this ratio cannot fall below 2.0, but it can rise higher, since streams of greater than first order can receive low-order tributaries without being promoted up the hierarchy; some estimates for large continental extents give bifurcation ratios of 4.0 or more (see further SEDIMENT YIELD OF DRAINAGE SYSTEMS).

Although the number system given here, and nowadays in common use, differs from Horton's original in the treatment of trunk streams, Horton's laws of drainage composition still hold, namely:

1. Law of stream numbers: the numbers of streams of different orders in a given drainage basin tend closely to approximate an inverse geometric series in which the first term is unity and the ratio is the bifurcation ratio.

2. Law of stream lengths: the average lengths of streams of each of the different orders in a drainage basin tend closely to approximate a direct geometric series in which the first term is the average length of streams of the first order.

These laws are readily illustrated by plots of number and average length (on logarithmic scales) against order (on an arithmetic scale). The plotted points lie on, or close to, straight lines. The orderly relationships thus indicated are independent of network pattern. They demonstrate exponential relationships. Horton also concluded that stream slopes, expressed as tangents, decrease exponentially with increase in stream order. The systematic relationships identified by Horton are independent of network pattern: they greatly facilitate comparative studies, such as those of the influences of lithology and climate. Horton's successors have extended analysis through a wide range of basin geometry, showing that stream width, mean discharge, and length of main stem can also be expressed as exponential functions of order, and drainage area and channel slope as power functions. Slope and discharge can in turn be expressed as power functions of width and drainage area, respectively. The exponential relationships expressed by network morphometry are particular examples of the working of fundamental growth laws. In this respect, they relate drainage-net analysis to network analysis and topology in general.

Relation of morphometric parameters and river flow. The functional relationships among various network characteristics, including the relationships between discharge on the one hand and drainage area, channel width, and length of main stem on the other, encourage the continued exploration of streamflow in relation to basin geometry. Attention has concentrated especially on peak flows, the forecasting of which is of practical importance; and since many basins are gaged either poorly or not at all, it would be advantageous to devise means of prediction that, while independent of gaging records, are yet accurate enough to be useful.

Problem of predicting peak discharge

A general equation for discharge maxima states that peak discharges are (or tend to be) power functions of drainage area. Such a relationship holds good for maximum discharges of record; but conflicting results have been obtained by empirical studies of stream order, stream length, drainage density, basin size, basin shape, stream and basin slope, aspect, and relative and absolute height in relation to individual peak discharges in the shorter term. One reason is that not all these parameters have always been dealt with. In any event, peak discharge is also affected by channel characteristics, vegetation, land use, and lags induced by interception, detention, evaporation, infiltration, and storage. Although frequency–intensity–duration characteristics (and, in consequence, magnitude characteristics) of single storms have been determined for considerable land areas, the distribution of a given storm is unlikely to fit the location of a given drainage basin; in addition, the peak flow produced by a particular storm is considerably affected by antecedent conditions, seasonal and shorter term wetting and drying of the soil considerably influencing infiltration and overland flow. Nevertheless, one large study attained considerable success by considering rainfall intensity for a given duration and frequency, plus basin area, and main-channel slope expressed as the height–distance relationship of points 85 and 10 percent of stem length above the station for which predictions were made. For practical purposes, the telemetering of rainfall in a catchment, combined with the empirical determination of its response characteristics, appears effective in forecasting individual peak flows.

Evolution of drainage systems. To empirical analysis of the morphometry of drainage networks has been added theoretical enquiry. Network plan geometry is specifically a form of topological mathematics. Horton's two fundamental laws of drainage composition are instances of growth laws. They are witnessed in operation especially when a new drainage network is developing; and, at the same time, probability statistics can be used to describe the array of events and of forms which are produced.

Random-walk analysis

Random-walk plotting, which involves the use of random numbers to lay out paths from a starting point, can produce networks that respond to analysis as do natural stream networks: that is, length and number increase and decrease respectively, in exponential relationship to order, and length can be expressed as a power function of area. The exponential relationship between number and order signifies a constant bifurcation ratio throughout the network; and a greater constancy in this respect would be expected from a randomly predicted network than from a natural network, which contains adventitious streams that join trunks of higher than one additional order. The exponential relationship between length and order in a random network follows from the assumption that the total area considered is drained to, and by, channels; the power relationship of length to area then also follows. The implication of the random-walk prediction of networks that obey the empirically derived laws of drainage composition is that natural networks correspond to, or closely approximate, the most probable states (see further LANDFORM EVOLUTION).

THE GEOMETRY OF RIVER SYSTEMS

Hydraulic geometry. Hydraulic geometry deals with variation in channel characteristics in relation to variations in discharge. Two sets of variations take place: variations at a particular cross section (at-a-station) and variations along the length of the stream (downstream variations). Characteristics responsive to analysis by hydraulic geometry include width (water-surface width), depth (mean water depth), velocity (mean velocity through the cross section), sediment (usually, concentration or transport, or both, of suspended sediment), downstream slope, and channel friction.

Graphs of the values of channel characteristics against values of discharge usually display some scatter or departure from lines of best fit. One main cause is that values on a rising flood often differ from those on a falling flood, partly because of the reduction of flow resistance, and hence the increase in velocity, as sediment-concentration increases on the rising flood. Bed scour and bed fill are also related. Nevertheless, the variations for a given cross section can be expressed as functions of discharge, Q. For instance, width, depth, and velocity are related to discharge by the expressions: $w \propto Q^b$, $d \propto Q^f$, and $v \propto Q^m$, where w, d, v and b, f, m are numerical constants. The sum of the exponents $b + f + m = 1$, because of the basic relation—namely that $Q = wdv$.

Width, depth, and velocity of flow

Similar functions can be derived for downstream variations, but, for downstream comparisons to be possible, the observed values of discharge and of channel characteristics must be referred to selected frequencies of discharge. When data are plotted on graphs with logarithmic scales for each of two discharge frequencies at an upstream and a downstream station, the four points for each channel characteristic define a parallelogram (Figure 3), whereby the hydraulic geometry of the stream is defined in respect of that characteristic. The values of exponents in the power equations differ considerably from one river to another: those shown here are theoretical optimum values. One common cause of difference is that many gaging stations are located where some channel characteristics are controlled, whether naturally as by rock outcrops or artificially as by bridge abutments. Constraints on variation in width, for instance, are mainly offset by increased variation in depth.

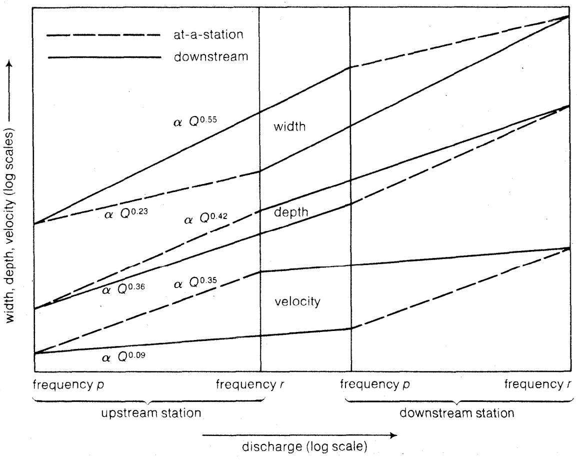


Figure 3: Hydraulic geometry for two frequencies of flow at an upstream and a downstream station, showing variation of width, depth, and velocity with discharge (see text).

Analyses of downstream variation in channel slope with discharge commonly reveal contrasts between field results and the theoretical optima. The discrepancy is probably due in considerable part to the fact that channel slope can vary in concert with channel efficiency, including channel habit, channel size, and channel form. Many past discussions of stream slope are invalidated by their restriction to the two dimensions of height and distance. In any event, the slopes of many natural channels are influenced by some combination of earth movement, change in baselevel, glacial erosion, glacial deposition, and change of discharge and load characteristics that result from change of climate. Consequently, although natural profiles from stream source to stream mouth suggest a tendency toward a smooth concave-upward form, many actually are irregular. Even without a change of baselevel, degradational tendency, or discharge, a change in channel sinuosity can produce a significant change of channel slope.

Down-stream changes of velocity

A marked downstream lessening of slope does not imply a decrease in velocity at a given frequency of discharge; reduction of slope is accompanied, and offset, by an increase in channel efficiency due mainly to an increase in size. The lower Amazon, with a slope of less than 7.6 centimetres per 1.6 kilometres (three inches per mile; less than 0.00005), flows faster at the bankfull stage than many mountain streams, at 2.4 metres (eight feet) per second. According to the assumptions made, an optimal velocity equation in hydraulic geometry can predict a slight increase, constancy, or a slight decrease in velocity downstream, for a given frequency of discharge. On the Mississippi, velocity at mean discharge (not a set frequency) increases downstream; velocity at the overbank stages of the five year and 50-year floods is constant downstream. Constant downstream velocity may well be first attained at the bankfull stage. The fact that relationships are highly disturbed at and near waterfalls and other major breaks of slope (the Paraná just below the Guaira Falls, for instance, runs at nine to 14 metres [30 to 45 feet] per second) has no bearing on the principles of hydraulic geometry, which applies essentially to streams in adjustable channels.

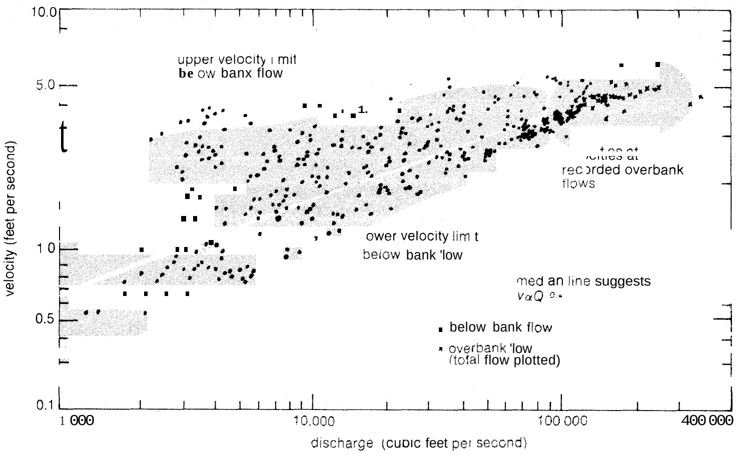


Figure 4: Variation of velocity with water discharge on the Red River at Fulton, Kansas.

The interrelationships and adjustments among width, depth, width–depth ratio, suspended-sediment concentration, sediment transport, deposition, eddy viscosity, bed roughness, bank roughness, channel roughness, and channel slope in their relation to discharge, both at-a-station and in the downstream direction, plus the tendency at many sections on many streams for variation to occur about some modal value, all encourage the conception of rivers as equilibrium systems. The designation quasi-equilibrium systems is usually used since not all variances can be simultaneously minimized, and minimization of some variances (for example, of water-surface slope) can only be secured at the expense of maximizing others (for example, channel depth).

River channel patterns. Distinctive patterns in the plan geometry of streams correspond to distinctive combinations of cross-sectional form, calibre of bed load, downstream slope, and in some cases cross-valley slope, ten-

(Left) Spence Air Photos, (right) V.C. Browne

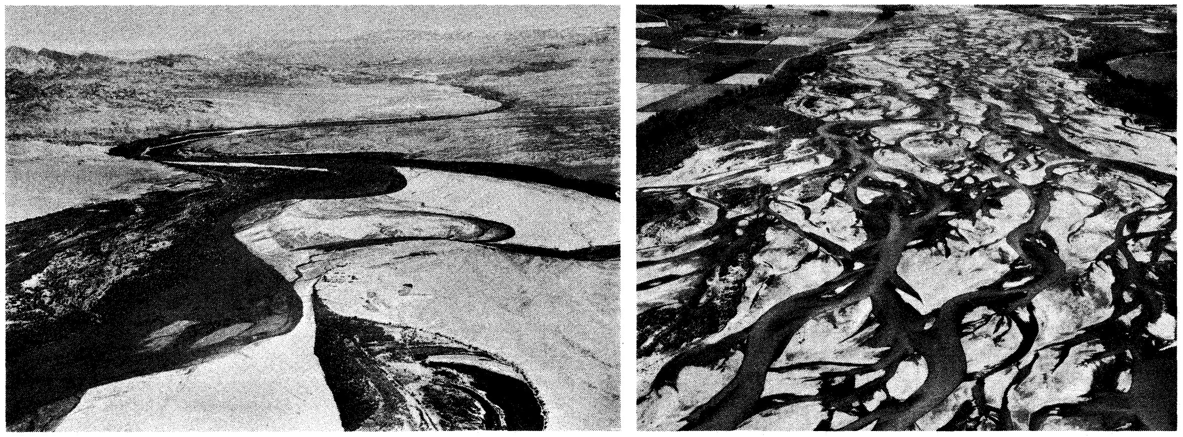


Figure 5: River channel patterns.
(Left) Depositional bars and oxbow cutoffs shown by the meandering reach of the Colorado River. (Right) Anastomosing (braided) reach of the Waimakariri River, New Zealand.

dency to cut or fill, or position within the system. The full range of pattern has not been identified: it includes straight, meandering, braided, reticulate, anabranching, distributary, and irregular patterns. Although individual patterns repeatedly recur in nature and are given separate names, the total range constitutes a series of continua.

Straight channels, mainly unstable, develop along the lines of faults and master joints, on steep slopes where rills closely follow the surface gradient, and in some delta outlets. Flume experiments show that straight channels of uniform cross section rapidly develop pool-and-riffle sequences in their beds. Pools are spaced at about five bedwidths. Lateral shift of alternate pools toward alternate sides produces sinuous channels, and the spacing of pools on each side of the channel is thus about five to seven bedwidths; this relation holds in natural meandering streams.

River meanders and their cause

Meandering channels are single channels that are sinuous in plan, but there is no criterion, except an arbitrary one, of the degree of sinuosity required before a channel is called meandering. The spacing of bends is controlled by flow resistance, which reaches a minimum when the radius of the bend is between two and three times the width of the bed. Accordingly, meander wavelength, the distance between two successive bends on the same side —or four bend radii—tends to concentrate between eight and 12 bedwidths, although variation both within and beyond this range seems to be related to variations in the cross-sectional form of the channel. Because bedwidth is related to discharge, meander wavelength also is related to discharge.

Meandering channels are equilibrium features that represent the most probable channel plan geometry, where single channels deviate from straightness. This deviation, and channel division in general, is related in part to the cohesiveness of channel banks and the abundance and bulk of midstream bars. When single channels are maintained, however, the meandering form is most efficient because it minimizes variance in water-surface slope, in angle of deflection of the current, and in the work done by the river in turning. This least-work property of meander bends is readily illustrated by the trace, identical with that of stream meanders, adopted by a bent band of spring steel. Meander plan geometry is simply describable by a sine function of the relative distance along the channel bend. The least-work and minimum-variance properties of the plan geometry, however, are secured only at the expense of maximizing the variance in depth. The longitudinal profile of the bed of a meandering stream includes pools at (or slightly downstream of) the extremities of bends and riffles at the inflections between bends. Increased tightness of bend, expressed by reduction in radius and increase in total angle of deflection, is accompanied by increased depth of pool. Where riffles are built of fragments larger than sand size, they behave as kinematic waves: that is, the speed of transport of material through a given riffle decreases as the spacing of surface fragments decreases, and the total rate of transport attains a maximum where the spacing is about two particle lengths, as in moving traffic on a highway. Numerous sandbed streams in dry regions, however, fail to develop pool-and-riffle sequences, maintaining approximately uniform cross sections even at channel bends.

Irregularities in meanders developed in alluvium relate primarily to uneven resistance, which is often a function of varying grain size. Variations in total sinuosity are probably due in the main to adjustments of channel slope. The process of cutoff (short-circuiting of individual meanders) is favoured not only by the erosion of outer channel banks and by the tendency of meander trains to sweep downvalley but also by the stacking of meanders upstream of obstacles and by increases of sinuosity that accompany slope reduction.

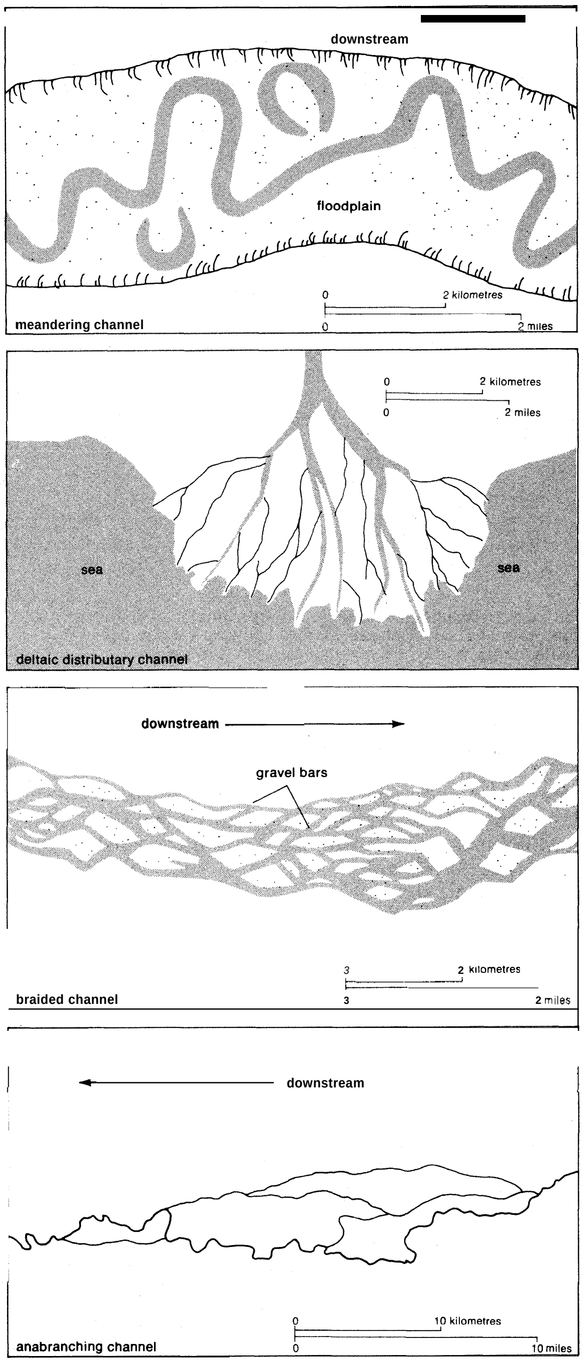


Figure 6: Common types of channel patterns.

Meandering streams that cut deeply into bedrock form entrenched meanders, the terminology of which is highly confused. It seems probable that, in actuality, the sole existing type of entrenched meanders is the ingrown type, where undercut slopes (river cliffs) on the outsides of bends oppose slip-off slopes (meander lobes) on the insides. For reasons not yet understood, lateral enlargement of ingrown meanders seems habitually to outpace downstream sweep, although the trimming of the upstream sides of lobes, and occasional cutoff, are well-known. Many existing trains of ingrown meanders belong to valleys rather than to streams, relating to the traces of former rivers of greater discharge. Reconstruction of the original traces indicates approximate straightness at plateau level, as opposed to the inheritance of the ingrown loops from some former high-level floodplain.

In a broader context, meander phenomena cannot be understood as requiring cohesive banks of the kind usual in rivers. Meanders, with geometry comparable to that of rivers, have been recognized in the oceanic Gulf Stream and in the jet streams (*q.v.*) of the upper atmosphere. In this way, stream meanders are classed with wave phenomena in general.

Braided channels

Braided channels are subdivided at low-water stages by multiple midstream bars of sand or gravel. At high water, many or all bars are submerged, although continuous downcutting or fixation, or both, by plants plus the trapping of sediment may enable some bars to remain above water. A single meandering channel may convert to braiding where one or more bars are constructed, as downstream of a tight bend where coarse material is brought up from the pool bottom. Each of the subdivided channels is less efficient, being smaller, than the original single channel; if its inefficiency is compensated by an increase in slope, that is, by downcutting, the bar dries out and becomes vegetated and stabilized. But many rivers that are largely or wholly braided along their length owe their condition to something more than local accidents. The braided condition involves weak banks, a very high width–depth ratio, powerful shear on the stream bed (implied by the width–depth ratio), and mobile bed material. Thus, braided streams are typically encountered near the edges of land ice, where valleys are being filled with incoherent coarse sediment, and also on outwash plains, as the Canterbury Plains of South Island, New Zealand; width–depth ratios can exceed 1,000:1. Studies on terraced outwash plains demonstrate that braided streams can readily excavate their valley floors: that is, they are by no means solely a response to valley filling.

Distributary patterns, whether on alluvial fans or deltas, pose few problems. A delta pass that lengthens is liable to lateral breaching, whereas continued deposition, both on deltas and on fans, raises the channel bed and promotes sideways spill down the least gradient. The branching rivers of inland eastern Australia, flowing across basin fills that range from thin sedimentary plains to thick riverine (fluvial) accumulations, have affinities with deltaic. distributaries although their patterns are only radial in part. A branch may run for tens of kilometres before joining a trunk stream, whether its own or another. Some channel plans change, notably at waterfalls, where multiple channels with high width–depth ratios above a fall are concentrated below the fall into a single gorge with a very low width–depth ratio. Outstanding examples include the falls of Khon (13 kiiometres, or eight miles wide; Mekong River), Guaira (Paraná River), Iguazu (tributary to Paraná), Victoria (Zambezi), and Niagara; channel width on the Iguazu below the fall is less than 2 percent of width at the fall lip.

STREAMFLOW AND FLUVIAL LANDFORMS

Stage-
discharge
relations

Peak discharge and flooding. Rapid variations of water-surface level in river channels through time, in combination with the occurrence from time to time of overbank flow in flat-bottomed valleys, have promoted intensive study of the discharge relationships and the probability characteristics of peak flow. Stage (depth or height of flow) measurements treat water level: discharge measurements require determinations of velocity through the cross-section. Although records of stage respond to frequency analysis, the analysis of magnitude and frequency is preferable wherever stage is affected by progressive scour or fill, and also where channels have been artificially embanked or enlarged or both. The velocity determinations needed to calculate discharge range from those obtained with portable Venturi flumes on very small streams, through observations with gaging staff or fixed Venturi flumes on streams of modest size, to soundings with current meters at intervals of width and depth at cross sections of large rivers. Frequent velocity observations on large rivers are impracticable. It is standard practice to establish a rating formula, expressed graphically by a rating curve. Such a curve relates height of water surface to the area of and velocity through the cross section, and thus to discharge. Secular changes in rating occur where a stream tends progressively to raise or lower its bed elevation. Short-term changes are common where the bed is mobile, and especially where the bed elevation–discharge relation, and thus the stage–discharge relation, differs between the rising and the falling limb of a single peak discharge curve. In such cases the rating curve describes a hysteresis loop. Rating curves for sandbed streams can include discontinuities, chiefly during rising discharge, that relate to behavioral jumps on the part of the bed.

Floods in hydrology are any peak discharges, regardless of whether or not the valley floor (if present) is inundated. The time-discharge or time–stage characteristics of a given flood peak are graphed in the hydrograph, which tends to assume a set form for a given station in response to a given input of water. The peak flow produced by a single storm is superimposed on the base flow, the water already in the channel and being supplied from the groundwater reservoir. Rise to peak discharge is relatively swift and is absolutely swift in small basins and on torrents, where the duration of the momentary peak is also short; on very large streams, by contrast, peak discharge can be sustained for lengths of days. Recession from peak discharge is usually exponential. The form of the hydrograph for any one station is affected by characteristics of the channel and the drainage net, and also by basin geometry, all of which can be taken as permanent in this context.

As already noticed above, flood-flow prediction that is based on permanent characteristics has hitherto achieved but partial success. Transient influences, also highly and at times overwhelmingly important, include the storage capacity of bedrock and soil, the interrelationships of infiltration, evaporation, and interception and detention (especially by vegetation), plus storm characteristics, which vary widely with respect to amount, duration, intensity, and location of rainfall with respect to the catchment.

Flood-
frequency
analysis
and the
annual
flood

In the longer term, flood-frequency analysis based on recorded past events can nevertheless supply useful predictions of future probabilities and risks. Flood-frequency analysis deals with the incidence of peak discharges, whereas frequency analysis generally provides the statistical basis of hydraulic geometry. Percentage frequency analysis has been much used in engineering: here, the 1 percent and 90 percent discharges, for instance, are those that are equalled or exceeded 1 and 90 percent of time, respectively. General observations of the flashy character of floods in headwater streams, in contrast to the long durations of flood waves far downstream, combine with analytical studies (as yet, restricted) to suggest, however, that percentage frequency is in some respects an unsuitable measure. Magnitude–frequency analysis, setting discharge against time, is directly applicable in studies of hydraulic geometry and flood-probability forecasting, although the best choice among types of probability is much debated. The accompanying example (Figure 7)

Adapted from *American Journal* of Science, vol. 251 (1953)

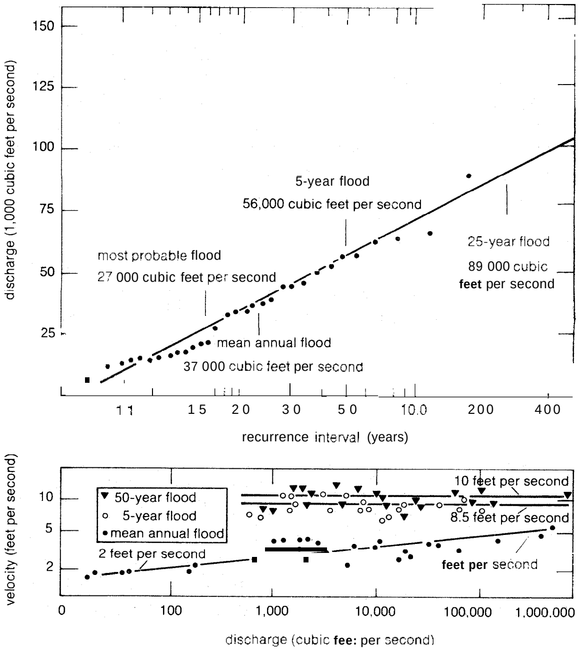


Figure 7: (Top) Magnitude of floods on the Sulphur River, Darden, Texas, and their probable recurrence interval. (Bottom) Relation of stream velocity and water discharge during floods of a given recurrence interval.

uses probability of extreme values to graph the annual series, wherein the peak discharge for each year of record is considered. Each recorded annual peak discharge is plotted against its simply computed recurrence interval. The 50-year flood has an average (but not regular) long-term spacing of twice in a century as an annual maximum; in any given year, it has a 2 percent chance of occurring as the annual maximum. On this scale, the 2.33-year flood is the mean annual flood. The most probable annual flood has a recurrence interval of 1.58 years (annual series); but, when floods lesser than, but independent of, the annual peak discharges are taken into analysis, it can be shown that the 1.58-year flood (annual series) has an average probability of occurrence of once a year.

Regional graphs of magnitude–frequency can be developed, given adequate records, for floods of any desired frequency or magnitude; but predictions for great magnitudes and low frequencies demand records longer than those usually available. Twelve years of record are needed to define the mean annual flood within 25 percent, with an expectation of correct results for 95 percent of time; and in general, a record should be at least twice as long as the greatest recurrence interval for which magnitude is desired.

Predictions of overbank flow, whether or not affected by artificial works, are relevant to floodplain risk and floodplain management. Notably in the conterminous U.S., floodplain zoning is causing risks to be reduced by the withdrawal of installations from the most flood-liable portions of floodplains or risks to be totally accepted by occupiers.

Catastrophic events

In the long geomorphic term, the transmission of sediment through the floodplain storage systems and through stream channels seems to result mainly from the operation of processes of modest magnitude and high frequency. Specifically, analyses suggest that total sediment transport by rivers is normally affected by flows approximating bankfull, over durations ranging down from 25 to 1 percent of total time. Infrequent discharges of great magnitude, such as are expectable on grounds of the probabilities of precipitation, snowmelt, and streamflow, range widely in destructive effect. Severe flooding is normally accompanied by great loss of life and property damage, the mean annual floods along the Huang Ho themselves affecting some 29,800 square kilometres (11,500 square miles) of floodplain, but geomorphic effects may be minimal, even with very large floods. The approximately 100-year floods of eastern England in the spring of 1947, fed bv unusuallv great and deferred snowmelt, scarcely affected either channels or floodplains. The 1955 floods in Connecticut. fed bv rains amounting to 58 centimetres (23 inches) in places, produced but spotty effects of erosion and deposition, even where floodplains were inundated to a depth of six metres (20 feet). For a given valley, there could be a threshold of inundation, river velocity, and sediment load, beyond which drastic changes occur. This is suggested, for example, by the catastrophic alluviation of valleys in eastern Australia and New Zealand during the last 4,000 or 2,000 years. Sudden catastrophes in the historical and geomorphic records relate to special events, mainly nonrecurrent: the 1841 Indus flood, which destroyed an army, the Gohna Lake flood of 1894 on the Ganges, and the 1925 Gros Ventre flood in Wyoming, accompanied the breaching of natural landslide barriers. The Lake Issyk (U.S.S.R.) flood of 1963, which caused widespread erosion and deposition, followed the overtopping of a landslide barrier by waves produced by a mudflow. The Vaiont Dam (Italy), although itself holding, was overtopped in 1963 by 91-metre (300-foot) high waves raised by a landslide: the floods downstream took more than 2,500 lives in 15 minutes. On the Huang Ho, the floods of 1887 took an estimated 900,000 lives. In late Pleistocene time, the overtopping of an erodible natural dam by the then existing Lake Bonneville eventually released nearly 1,666 cubic kilometres (400 cubic miles) of water; the maximum discharge of about 280,000 cubic metres (10,000,000 cubic feet) per second is comparable to the flow of the Amazon, but velocities were very high, perhaps ranging to 7.6 metres (25 feet) per second. The greatest flood peak so far identified is that of the ice-dammed Lake Missoula in Montana, which, on release, discharged 2,085 cubic kilometres (500 cubic miles) of water at an estimated peak flow of 8,500,000 cubic metres (300,000,000 cubic feet) per second, more than nine cubic miles per hour. Iceland is notable for glacier bursts, which are nonrecurrent where they result from subglacial eruptions but recurrent where they involve the sudden failure of ice dams, as with Grimsvotn, which periodically releases 8.3 or more cubic kilometres (two cubic miles) of water in floods that peak at 57,000 cubic metres (2,000,000 cubic feet) per second. Deposition by glacier-burst floods is illustrated by Iceland's Sandur plains.

Ice-dam flooding

Peak discharges that close the range between natural floods of great magnitude and low frequency on noncatastrophic streams and natural catastrophic floods of great magnitude and perceptible frequency include stormwater discharges from expanding urban areas. Because of the progressive spread of impermeable catchment and efficient runoff systems, such floods tend to increase both in frequency and in magnitude.

River floodplains. A floodplain, alluviated land of minor relief traversed by a river channel, is by definition liable to inundation. For convenience, the name is also applied to the alluvium itself. At the simplest, this forms a strip wide enough to accommodate the meander train, extending down to the depth of scour where bedrock is cleanly planed off; along the edges of the floodplain, the migrating meanders trim the valley walls into bluffs. As meander curves enlarge themselves, sweep downstream, and short-circuit themselves or one another, the alluvium is constantly reworked, being eroded from the outer sides of bends but redeposited on the inner sides as point bars. Rates of downstream sweep for alluvial meanders are poorly documented but are known to vary widely from stream to stream, and also in some streams along the length of the valley. Observations on the Rio Grande can be read as indicating sweep through a distance equivalent to one wavelength (distance between two meander extremities) in as little as 20 years; by contrast, some meander trains cut deeply into bedrock have shifted very little downstream in intervals as long as 500,000 years.

The minimum width for a completely developed floodplain equates with meander amplitude, which is to some extent controlled by the tendency of radius of curvature to equal two to three bedwidths, but many individual meanders are distorted by irregularities of composition in the floodplain alluvium, and in special cases the amplitude can be grossly exaggerated by the influence of bedrock structure. Many floodplains are in any case complex, being contained in older alluvium below the present depth of scour. Some, developed on deep and wide valley

By courtesy of the Australian News and Information Bureau

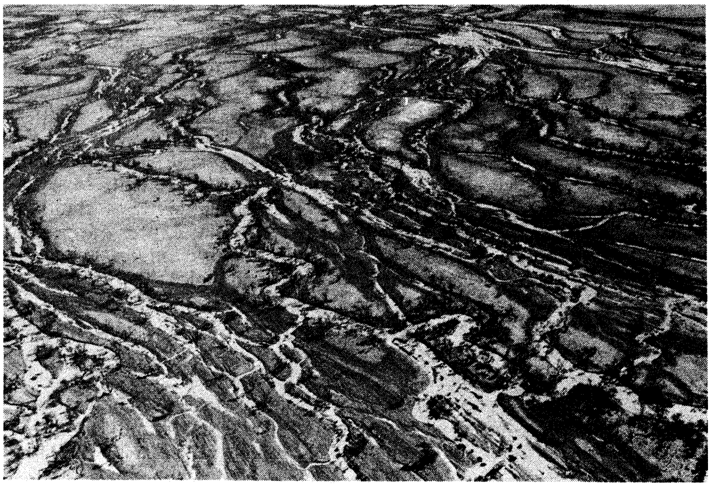


Figure 8: Australian channel country after seasonal rain has caused Cooper Creek to flood. The waters spread over a complicated network of rivers, channels, and swamps to form a natural irrigation area.

fills, are many times wider than the meander belt. Some floodplains possess natural levees, others none: the controlling difference may he a combination of sediment calibre with suspended-sediment concentration at times of overbank discharge. Natural levees tend to impede drainage on the outer parts of floodplains and to produce backswamps. Floodplains have been little studied in relation to other than meandering streams; both on these and on braided streams, floodwater can inundate low terraces on occasion.

Inundation and its effects

Magnitude–frequency analysis suggests that a stream that is neither raising nor lowering its profile at a significant rate will inundate its floodplain in about two years out of three. The geomorphic effects of inundation, including the inundation during high-magnitude, low-frequency floods, varies widely from valley to valley. In some valleys the bulk of the floodplain alluvium consists of point bar deposits, the deposition of which maintains channel width as a meander shifts, although the raising of the point bar surface to the general level of the floodplain may depend on encroachment by plants and their trapping of sediment. In such valleys the geomorphic influence of floods is minimal, despite the characteristic turbidity of floodwater; little sedimentation or valley-bottom erosion occurs during times of flood. In other valleys, botanical studies show that floodplains have been raised by vertical accretion. Trees rooted at former floodplain levels and toppled and covered by sediment during high floods can commence regrowth at the new levels of the raised alluvial surface, supplying markers in the history of buildup.

It is very difficult in general to separate natural from man-induced causes of floodplain alluviation. In some areas clearance, settlement, and cultivation have certainly been accompanied by alluviation of valley bottoms; but man's activities appear elsewhere merely to have accelerated natural processes, in response to which valley filling (including rise in floodplain elevation) alternates with valley clearing (including decrease in floodplain elevation). In some settings the whole matter is complicated by conversions to braiding, with steepening of gradients and much infilling, during ice-marginal or periglacial intervals or both, and reconversions to meandering, with lessening of gradients and excavation of valley fills, during interglacials.

River terraces. In one simple form, a river terrace is the remains of an old floodplain, cut through by the river and left standing above the present floodplain level. In this form the terrace has a flat depositional top and a steep fore-edge; it consists of alluvium. According to context, terrace can connote the upper surface, the combination of top and fore-edge, or the constituent material.

By courtesy of the U.S. Geological Survey: photograph, W.G. Pierce

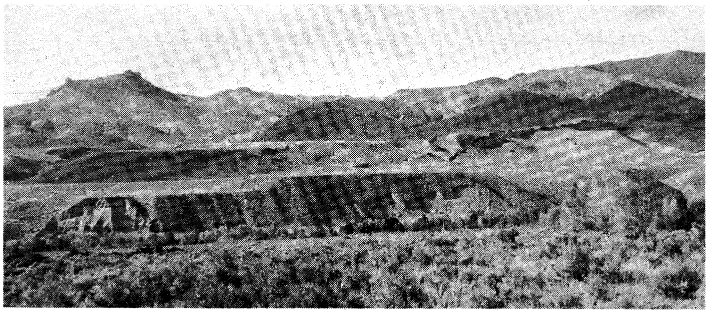


Figure 9: Terraces along the north side of the South Fork of the Shoshone River, Wyoming. The lowest terrace (left) is the post-Wisconsin Cody terrace. Twenty feet above is the pre-Wisconsin Cody terrace.

Paired terraces and their significance

A paired terrace is one where the terrace features (particularly the elevation and slope of the terrace surfaces) correspond across the valley. Sequences of paired terraces are common in the lower valleys of rivers that drain stable unglaciated areas. They relate to the sea-level stands of interglacial episodes, each successive stand being lower than the preceding, in consequence of the secular (cyclic) lowering of sea level on which glacial fluctuations were superimposed. Little is known in general of the rock floors on which such terraces rest; some terrace deposits are presumably shallow, representing simple floodplains, and the rock floors beneath evenly trimmed into planated benches. Because they are related to events of genetic significance, paired terraces have been much studied in connection with valley development; those of northwest Europe possess much archaeologic significance.

Where renewed downcutting has extended a new long profile part way up a stream, the head of the new profile and the remaining tail of the old profile intersect in a knickpoint. As already shown, however, a local increase in downstream slope results from causes other than renewed cutting. Where renewed cutting is responsible, the earlier forms, including floodplain deposits conveıted to terraces, can be used to extend the old profile downstream of the knickpoint; but no great success has attended efforts to fit equations to the curves of old profile plus terrace, and so to reconstruct past strandlines. Older, higher terraces are usually much eroded in lower valleys; allowances for local changes in former gradient are difficult to make.

Unpaired terraces display no cross-valley correspondence. They are most readily produced where a meandering stream is rapidly eroding a valley fill, while shifting its loops both downvalley and cross-valley. The terrace levels (except for the top of the fill) are now determined by the downvalley passage of single meanders, while the heights of terrace scarps depend on the extent of lateral channel swing.

Cut-and-fill sequences

Any climatic or other shift that causes intermittent downcutting, or a pair of shifts that cause filling to be followed by cutting, is likely to produce terraces. Intermittent uplift of the land or intermittent fall of sea level could bring identical suites of terraces into being, always provided that the stillstands (periods of stability) were long enough for floodplains to form. Many trains and sheets of outwash, deposited by the meltwater streams of land ice, have been subjected to intermittent clearance by braided streams, as notably on the Canterbury Plain of New Zealand; each new episode of cutting leaves part of the last wide stream bed as the top of a new terrace, terminating in a scalp where the rıver undercuts at the new, lower, level. A related form is the kame terrace, originally deposited alongside a wasting valley glacier and liable to be pitted by holes called kettles.

In some river systems the downstream relationships of terraces and present floodplain are complex. Where the upper basin was copiously supplied during glacial maxima with rock waste, either from glaciers or by periglacial action, stream gradients were increased by valley filling. Srmultaneously, gradients at and near the shore were increased by valley cutting in response to the glacial lowering of sea level. Many rivers affected in this way adopted braided habits. Reversion to interglacial conditions brought renewed cutting in the upper valley and filling in the lower valley, often with reconversion of the stream to meandering on a much reduced downstream slope. The steep fill surfaces of glacial times plunge seaward beneath existing floodplains; the sequence of cut and fill is inverted from the headward to the seaward valleys; and additional complications occur wherever a stream achieves net cutting between one interglacial and the next. Still further complications apply to streams that, developing a descending sequence of terraces in their upper valleys, experience deltaic subsidence at their mouths. The oldest former floodplain is represented by the highest terrace in the upper valley but is the most deeply buried in the delta.

A special type of terrace forms where a large meltwater stream draining from wasting land ice greatly raises its bed level and floods its tributary valleys with water and sediment. The sediment advances by delta-like progression into the lateral valleys, taking on remarkably horizontal top surfaces. When the trunk river begins to clear its valley fill and the tributaries are reactivated, the fills of the lateral valleys are cut through, their surviving portions forming backfill terraces. These are characteristically paired.

Climatic implications of cutting and filling

Alternations of cut and fill result from climatic shifts independent of, or in some areas additional to, glaciation, periglaciation, and strandline movements. Areas that are currently semiarid to arid have been especially studied in this respect. There is incomplete agreement on the working of the factors involved, except that these include changes in the amount (and probably also regime and intensity) of precipitation, changes in vegetal cover, and changes in temperature that are significant mainly for vegetation and evapotranspiration losses. Valley filling means an increase in stream slope: it implies an increase in sediment load, or at least a marked change in the calibre of load. Increased delivery of load implies increasing instability of the waste mantle on the hillslopes. Reversal to valley cutting requires reversal of some, or all, of the factors that promote filling. Studies in semiarid areas suggest that a shift either toward aridity or toward humidity would reduce total erosion; but the full combined effects of changes in climate, vegetation, and channel form and habit in relation to slope (and thus to cut or fill) may not yet be fully understood. It is clear, however, that quite small climatic changes in unstable environments can result in major changes from cut to fill or fill to cut. In at least some lately settled mountainous areas, the present gully cycle is independent of, although accelerated by, the activities of man in clearing and grazing.

Of necessity, visible terraces are less liable to flood than are floodplains; except in very large valleys, therefore, they usually give preferred locations of settlement, although very low terraces are exposed to occasional flood risk in some valleys. Little work has been done on the frequency of inundation of terraces proper, but recent investigations of valleys too narrow to contain terraces reveal series of small benches that relate to particular intervals or ranges of intervals on the magnitude–frequency scale. The predominant bench may relate to the frequency of the most probable annual flood; the causal connection, if any, between higher benches and channel-side berms (benches or ridges), on the one hand, and points on the magnitude–frequency scale, on the other, cannot yet be fully explained.

River deltas. The term delta originates from the A-like plan view of the exposed sedimentary accumulation at the mouth of the Nile. Deltas result from the outbuilding of alluvial deposits at river mouths; they frequently form seaward extensions of floodplains and in large instances of alluvial plains, which they themselves extend as they build oceanward. Delta plan geometry is simplest in small lacustrine deltas that build out from river channels fixed in bedrock: as such deltas rise, the channels on the delta repeatedly split, spilling down the lowest available slopes. The total effect is that the streams work back and forth across a cone-like surface, tending to produce an arcuate shoreline. Generally similar considerations apply to deltas formed in inland lakes, the Volga Delta at the entry to the Caspian Sea, for instance, attaining a surface area of almost 10,000 square kilometres (4,000 square miles); but the inland delta of the Niger, at its entry from the Fouta Djallon highlands to the Sud Plain, is a giant alluvial fan, correlative to the riverine accumulations of the members of the Murray–Darling system of southeast Australia where these leave the encircling highlands for the interior sedimentary plains.

Characteristics and changes of coastal deltas through time

Large coastal deltas also result from an excess of deposition over removal; and they too are typified by distributary channels. The visible areas—59,600 square kilometres (23,000 square miles) for the Ganges–Brahmaputra, about 25,000 to 35,000 square kilometres (9,500 to 13,500 square miles) for the Lena, Mississippi, and Niger, and a range of 24,600 to 12,200 square kilometres (9,500 to 4,700 square miles) for the Orinoco, Nile, Irrawaddy, and MacKenzie—represent only part of the depositional surface. Individual large deltas vary widely in plan, according to the interplay of deposition, shoreline processes, subsidence, general changes in sea level, and channel habit within the exposed parts. Unless they are advancing very fast, all great deltas locally concentrate enough sediment to depress the underlying crust; but the general tendency is one of progradation, that constructs a lengthening alluvial plain upstream of the delta proper. Long-continued deposition in a single area imparts lenticular cross sections to the sedimentary piles, whereas subsidence exerts a partial check on outbuilding; contemporary subsidence of the Po Delta offers a serious threat to the city of Venice. Submergence independent of subsidence results from deglacial rises in sea level; were it not for the latest rise, between about 15,000 and 5000 BP, coastal deltas would be commoner than they are and the visible deltas more extensive. Delta-building streams that are also levee builders can resist splitting, running far seaward before general breaching of the levees opens a new outlet. In the Mississippi Delta, levee breaching is followed by the formation of a new sedimentary lobe, but this sequence of processes appears untypical of large deltas generally. The Mississiyyi has initiated seven new lobes in the last 5,000 years. High tidal ranges do not inhibit delta building. Wave attack and longshore drift on a delta shoreline of moderate to high energy, on the other hand, can greatly modify the shoreline plan, checking extension of outlets and building spits perpendicularly to them; on some deltas, massive wave-built spits enclose lagoons.

Outbuilding lessens seaward gradients on the delta and also reduces the net gradients between upstream points in the valley and the advancing shore. In consequence, flood liability next upstream of the delta increases, unless the valley bottom is raised by sedimentation; and, within the delta, conditions are set for intermittent major diversions of the chief outlets. The latest 5,000 years in the Mississippi Delta have brought three major displacements of the main channel above the delta head, plus six marked relocations of the chief distributary system. The Huang Ho, artificially embanked for 725 kilometres (450 miles), has raised its bed in places more than ten metres (33 feet) above floodplain level. In some 4,000 years of record, bank failure has averaged one year in three and overtopping one year in ten. Seven complete changes of lower course roughly double the frequency observed on the Mississippi.

Ephemeral streams. Ephemeral streams, constituting the transition from perennial through intermittent to rare surface runoff, command much attention, especially because their effects in erosion, transportation, and deposition can be inordinately great and also because they relate closely to periods and cycles of gullying. Their channels generally have higher width–depth ratios than those of unbraided channels in humid areas; *e.g.*, 150:1 or more on small streams. In extreme cases, ephemeral streamflow merges into sheetflood. Stream beds, usually sandy, are nearly flat in cross section but contain low bars where gravel is available: these behave in many ways like riffles or braid bars elsewhere. Although beds and banks are erodible, the fine-material fraction is usually enough to sustain very steep channel banks and gully walls. Rapid downcutting produces flat-floored trenches, called arroyos, in distinction from the often V-shaped gullies of humid areas.

Geometry of arroyos

Discontinuous vegetation cover, well-packed surface soil, and occasionally intense rainfall promote rapid surface runoff, conversion of overland to channelled flow, and the multiplication of channels. Although reliable comparative data are scarce, it seems likely that ephemeral channel systems develop higher order ranking, area for area, than do perennial streams: channels as high as Ilth order are recorded for basins of about 1,300 square kilometres (500 square miles), whereas the Mississippi is usually placed only in the tenth order. This apart, geometry of ephemeral nets obeys the laws of drainage composition that apply to perennial streams: stream length, stream number, channel width, and water discharge can be expressed as exponential functions of stream order, and drainage area and channel slope as power functions, whereas slope and discharge can be expressed as power functions of width and drainage area.

At-a-station (a particular cross section) variations in width, depth, and velocity with variation in discharge in ephemeral streams resemble the corresponding variations in perennial streams; but differences appear when downstream variations are considered. For a given frequency of discharge, the rate of increase in width differs little between the two groups, but ephemeral streams increase the more slowly in depth, becoming increasingly shallow in proportion in the downstream direction. This effect is compensated by a more rapid downstream increase in velocity, which reflects high concentrations of suspended sediment and a resultant reduction of friction. Ultimately, however, the ephemeral flood may lose so much water by evaporation and percolation that the stream is dissipated in a terminal mudflow.

Trenching, the extension of gullies, and their conversion into arroyo systems, implies valley fills of erodible surficial material. Like streams of humid regions, ephemeral stream systems record complex histories of cut and fill: it is reasonable to expect comparable timing for climatically controlled events. Whatever the effect upon stream erosion of historical settlement in the western U.S., inland eastern Australia, and New Zealand, the present episode of gullying seems merely to have been intensified by man's use of the iand. Accelerated channelling frequently involves three processes not characteristic of humid regions: pipicg, headcutting, and the formation of channel profiles that are discontinuous over short distances.

Piping and discontinuous gullies

In piping, water that has penetrated the topsoil washes out the subsoil where this is exposed in section, forming small tunnels that may attain lengths of many feet. Collapse of tunnel roofs initiates lateral gullying and lengthens existing cuts headward. Headcutting is commonly associated with piping, because headcuts frequently expose the subsoil. A headcut is an abrupt step in the channel profile, some inches to some feet high; it may originate merely as a bare or trampled patch in a vegetated channel bed but will increase in height (like some very large waterfalls) as it works upstream. At the foot of the headcut is a plunge pool, downstream of which occurs a depositional slope of low downstream gradient. Formation of successive headcuts, say at an average spacing of 150 metres (500 feet), snd the construction of depositional slopes below each, causes the profile to become stepped; ephermeral streams with stepped profiles are called discontinuous gullies. Speed of headcut recession varies widely with the incidence and intensity of rainfall; but ultimately, when the whole profile has been worked along, and the bed widened, the original even slope is restored, although at a lower level than before.

THE RIVER SYSTEM THROUGH TIME

Natural river systems can be assumed to have operated throughout the period of geologic record, ever since continental masses first received sufficient precipitation to sustain external surface runoff. The Precambrian portion of the record, prior to 600,000,000 years ago, is complicated by the widely metamorphosed character of the surviving rocks, although even here the typical cross-bedding of shallow-water sands can be recognized in many places. The Cambrian and post-Cambrian succession of the last 600,000,000 years contains multiple instances of deposition of deltaic sandstones, which record intermittent deposition by rivers in many areas in many intervals of past time. The span since the Precambrian is long enough, at present rates of erosion, for rivers to have shifted the equivalent of 25 to 30 times the bulk of the existing continental masses, but the rate of erosion and sedimentation is estimated to have increased with time. Of necessity, river systems now in existence date from times not earlier than the latest emergence of their basins above sea level, but this limitation allows numbers of them to have histories 100,000,000 years or more in length.

Drainage diversion by stream capture

A river system of appreciable size is likely to have undergone considerable changes in drainage area, network pattern, and profile and channel geometry. Adjoining streams compete with one another for territory. Although competition is effectively nil where divides consist of expanses of plateau, or where opposing low-order streams of similar slope flow down the sides of ridges, it frequently happens that fluvial erosion is shifting a divide away from some more powerful trunk stream and toward a weaker competing trunk. In extreme cases, the height difference is so marked that a tributary head from one system can invade, and divert, a channel in the adjoining system: such diversion, usually named stream capture, has already been noted as a principal mechanism in the adjustment of network patterns to structural patterns. Close general adjustment to structure implies multiple individual adjustments, unless the stream network has developed solely by the headward extension of tributaries along lines of structural and lithologic weakness: the network predicated on a single regional slope is dendritic in pattern. By encroachment and capture, a successfully competing stream becomes yet more powerful, the headward extension of its basin increasing the discharge of the trunk channel and permitting reduction of slope; *i.e.*, additional downcutting. Seaward extensions of basins occur where deltas lead the outbuilding of alluvial plains, and where crustal uplift (and also at times strandline movements) result in emergence. Conversely, basin area is reduced along the seaward edge by submergence, in response to crustal depression or rise in sea level. The potential limits to basin size are fixed by available areas of continent with surface moisture surplus, in combination with theoretical optimum shape of basin; but actual basin shapes, for all large rivers, are to some extent affected by crustal deformation.

Glacial damming and isostatic rebound effects

Derangements other than the captures effected in stream competition include those due to nonfluvial invasion and deposition. Regional flooding by basalts, as during the Tertiary Period (65,500,000 to 2,000,000 years ago) in the Deccan of India and the northwestern part of the U.S., obliterates the former landscape and provides a new surface on which new drainage networks form. Major invasions by continental ice displaces fluvial systems for the time being; glacial deposits, especially till sheets, can conceal the preglacial topography and provide initial slope systems for postglacial streams. Individual diversions occur at and near ice fronts, also where preglacial divides in mountain country are breached by the ice of caps or impounded mountain glaciers. The full history of drainage derangement by continental ice is often complex, depending on the particular combinations of preglacial outlet directions, extent of glacial invasion, relation-"ship of regional slope to direction of ice advance, thickness of glacial sedimentation, amount and speed of postglacial isostatic rebound, and self-selection of postglacial outlet directions and drainage lines. The North American Great Lakes and Midwest areas, the Thames Basin in England, and the Eurasiatic plain all record intricate histories of damming during glacial maxima, with postglacial networks and outlets differing markedly from those of preglacial times. Glacial breaching of divides requires the passage cf thick ice through a preglacial notch or gap, with erosion severe enough to provide a new drainage line when the ice melts. The spinal divide of Scandinavia was breached by the ice cap centred over the Gulf of Bothnia, just as the highland rim of Greenland is being breached by effluent glaciers today. After deglaciation, areas of divide breaching display streams with anomalous courses through gaps in major relief barriers. Morphologically related to glacial breaching, especially with respect to indeterminate present-day divides, are the disordered drainage nets of formerly glaciated terrains where bedrock is widely exposed and where relief is subdued.

Changes through time in channel slope have already been partly treated in connection with terraces. In the long view, streams must tend to reduce their slopes as the basin relief is lowered, although isostatic (balancing) compensation for erosional reduction of load largely offsets the reduction of slope. The effects involved here are independent of, although necessarily associated with, glacial–deglacial changes in the strandline level, crustal warping, and isostatic rebound from glacial reduction of load. It can be argued that large river systems, removing large quantities of sediment and dumping them offshore, should promote intermittent isostatic uplift when yield thresholds are passed and, in consequence, promote the generation of new waves of erosion that, working upstream, are recorded in sequences of cyclic knickpoints. The implications of this conceptual view have been applied especially to the unglaciated shield areas (central and oldest part of continents, generally) of tropical latitudes and extratropical parts of the Southern Hemisphere, in all of which rivers descend in high falls or lengthy cascades across the edges of major erosional platforms. In the shorter term, severe and rapid erosion of a trunk channel can leave a tributary valley stranded at height. Channel geometry demands that tributary glacier troughs should hang above the floors of main troughs, while tributary stream valleys often hang above trunk valleys formerly occupied by long glacier tongues. Hanging valleys on shorelines are correspondingly due to the outpacing of channel erosion by cliffing.

Effects of climatic change

Climatic shifts are known to be capable of effecting fill or clearance of channels and valleys: they can also change channel habit. In addition to the alternation in some near-glacial areas between braiding during maximum cold and meandering during interglacial warmth, the record includes conversions of channel width and meander pattern. On numerous midlatitude streams, existing channels have been much reduced from their earlier dimensions; and on many, but by no means all streams, existing floodplains are contained in the floors of meandering valleys where the wavelength is determined by the plan of the floodplain as opposed to the existing channel. Valley meanders were cut by streams 20 to 100 times as voluminous as existing streams, at the bankfull stage. They illustrate only one variant, although a widespread one, of the underfit stream, which combines a former large with an existing reduced channel. Reduction to the underfit state is commonly, although not invariably, accompanied by the infilling of former large channels both laterally and from below, so that existing floodplains are contained in valley-bottom fills. Accidents of capture and glacial diversion apart, the underfit condition results generally from climatic shift. The last major shift responsible for channel shrinkage appears to have occurred in the interval 12,000 to 9,000 BP, or later in areas that were still ice-covered 9,000 years ago (see further CLIMATIC CHANGE). Involving a reduction of bedwidth to as much as one-tenth of earlier values, and in meander wavelength by similar proportions, channel shrinkage is known to have been succeeded in well-studied areas by lesser fluctuations that are recorded in episodes of partial clearance followed by renewed fill. Significant alternations between cut and fill during the last 10,000 to 20,000 years have perhaps averaged a periodicity of 1,000 to 2,000 years: there is no a priori reason to suppose that the corresponding periodicity differed from this value during the whole Pleistocene, 2,000,000 to 3,000,000 years in duration so far. Inferences about pre-Pleistocene fluctuations await detailed analysis of rates of deposition of graded beds, coral growth, and the like.

On account of the temporal–dynamic qualities that have been discussed, stream channels and river networks are to be regarded as open systems (those open to additions or subtractions of materials or energy through time), whether in relation to short-term adjustments to individual peak discharges, in relation to accommodation to the constraints of climate, vegetal cover, characteristics of infiltration and overland flow, or in relation to the long-term influences of crustal movement, interbasin competition, and land wastage. Channels and networks experience inputs and outputs of matter and energy. Some, but probably a small minority at any one time, and for a minor duration of total time in any one channel or network, act as open systems in disequilibrium. The general tendency seems to be for channel and river systems to attain steady-state conditions, wherein negative feedback tends to counter individual disequilibrium tendencies, and counteracting effects ensure variations about recurrent norms of form and behaviour. (G.H.D.)

LIFE IN RIVER SYSTEMS

The river as a biological environment. A distinction is often made between a zone of erosion and a zone of deposition in rivers, but it must be remarked that, whatever the size of the particle that is moved by a flood, it will be deposited when the flow is less. It is preferable to make the distinction between zones where the bottom has stones large enough for the biggest invertebrate animals to cling to and where the bottom is of sand or mud whose particles are so small that the larger invertebrates must burrow. If a river with a sandy or muddy bottom is shallow enough, rooted plants will grow in it. Between the two zones there will be an intermediate one where fine particles settle during low flow and are washed away by a flood, and the slope, or river bottom grade, may be such that this intermediate zone is longer than either of the other two. A river from source to mouth is a continuous whole, and attempts to divide it into zones must be arbitrary. Any classification system may prove useful for a particular purpose, without being a concept with fundamental significance for biologists.

Water emerging from deep underground layers is generally cold, often with but a small range of temperature change during the course of a year, but it may be warm if it drains shallow soil. The smaller the volume of water the warmer it will become in the sun and the more heat it will lose at night. Larger water volumes have less daily temperature fluctuation and they are never far from average air temperature. If there is not a big difference in altitude, headstreams may therefore reach a higher maximum temperature than the main river. This condition, of course, has importance regarding the type of life found in the various parts of a stream.

Oxygenation of streams

A torrential stream is nearly always well oxygenated. A slow river is often well oxygenated by day because of the activities of photosynthetic plants, but it may be depleted of oxygen by night, when only oxygen consuming and decomposition processes are active. Turbidity, the muddiness and cloudy condition of the water, generally increases steadily from source to mouth. Accumulated organic matter washed from the land is believed to be the main base of the food chain in running water, and therefore in hilly or mountainous regions the amount of material entering the stream that may become a source of food tends to increase as the area drained increases with distance from the source. The amount may decrease farther downstream where the carrying capacity of the river drops with reduced flow.

Biological classification of streams. An old and well-known method of dividing rivers into zones is based on the presence of one of four fish species commonly found in western Europe: *Salmo* (trout), *Thymallus* (grayling), *Barbus* (barbel), and *Abramis* (bream). Each zone as defined by these fish coincides with a given slope, though for any zone the wider the stream the more gentle the slope. Even in western Europe there is a certain inconvenience in designating the zones by species of fish, for some, notably grayling, are not widespread. Elsewhere other species must be sought. Running water systems have also been classified into zones called the rhithron, which is roughly the stony upstream region, having well-oxygenated water and with temperature rising to no more than 20" C (68° F), and the potamon, which is the downstream portion having a sandy or muddy bottom and temperature exceeding 20° C at the warmest time of year. Both rhithron and potamon are subdivided. Within any one subdivision, however, the composition of the living community may be changed by some factor unconnected with flow or temperature, and the terms therefore have no precise biological significance though they may be useful in a general sense.

Organisms in streams. The slower the streamflow, the more the composition of the living community resembles that of still water and, therefore, any discussion of peculiar features of running-water organisms must centre largely round those that have colonized the swifter flowing waters.

Stagnant waters fill and disappear but a watercourse must flow as long as there is precipitation to supply it. It is, therefore, not surprising to find in running water certain groups of organisms that have changed little during the millions of years in which others have evolved

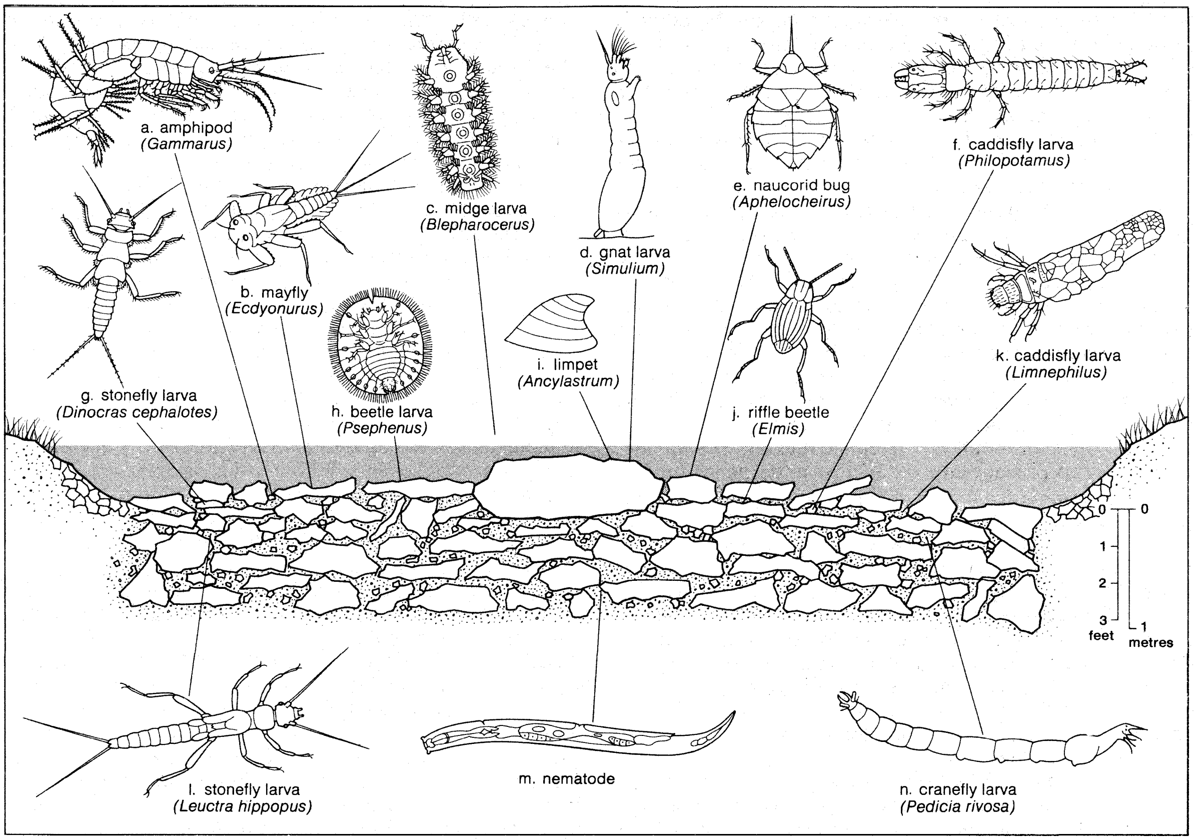

a. amphipod (Gammarus)
c. midge larva (Blepharocerus)
d. gnat larva (Simulium)
e. naucorid bug (Aphelocheirus)
f. caddisfly larva (Philopotamus)
b. mayfly (Ecdyonurus)
g. stonefly larva (Dinocras cephalotes)
h. beetle larva (Psephenus)
i. limpet (Ancylastrum)
j. riffle beetle (Elmis)
k. caddisfly larva (Limnephilus)
0
1
2
3
feet
0
1
metres
l. stonefly larva (Leuctra hippopus)
m. nematode
n. cranefly larva (Pedicia rivosa)

Figure 10: Riverbed inhabitants.

From (a, d–g, i, k–n) T. Macan, *A Guide to Freshwater invertebrate Animals;* by permission of the author and Longman Group Ltd., (b) G. Pleskot, *Der Stand der Biologischen Fliesswasserforschung,* (c, h, j) C. Wesenberg-Lund, *Biologie der Susswasserinsekten*

and adapted. Most members of the insect order Plecoptera (stoneflies) and many of the order Ephemeroptera (mayflies) inhabit running water, and other primitive groups confined to it include the crustacean ***Anaspides*** and the caseless free-ranging caddis larvae of the insect family Rhyacophilidae. Moreover there are few groups of freshwater organisms that are not represented in running water. The exceptions are extreme specialists such as the phyllopod crustaceans (fairy shrimps and related forms), which must swim to feed and have survived only in temporary water where they are free from the predation to which this mode of life exposes them.

Hazards of the environment

Rapidly flowing water is well-oxygenated and a moving medium brings to its inhabitants a constant supply of oxygen and also salts, at the same time removing their waste products. It also carries a constant supply of food. The disadvantage of life in a moving medium is that any accidental displacement must always be in one direction. Even the organisms that live in the substratum (various layers of the stream bottom) are occasionally subject to this hazard when exceptional flow causes the substratum to shift. A further problem is the accomplishment by organisms of even colonization of a biotope (uniform habitat occupied by a uniform community of organisms) that may be very long and very narrow.

Plant communities. Stones and rock are covered by algae, many of which are single cells, though some are filamentous and traii in the water in tresslike tufts. There are many species of these and they are often scoured off their supports by sand and gravel when the rate of streamflow increases; there are marked changes in the communities with the seasons even when this does not happen. Algae tend to occur in irregular patterns varying considerably in species composition and abundance. It has not, therefore, so far proved possible to discern definite algae communities associated with given stream conditions, though it is known that temperature, light, substratum, and dissolved substances in the water are factors that influence the occurrence of certain species.

A stone or boulder that survives for a long time without being overturned, and that is not scoured by particles of sand and gravel, will become densely covered by mosses and liverworts.

Rooted plants grow thickly where substratum and water depth are suitable, but attempts to define communities comparable with those in still water have been unsucessful. This is partly due to the instability of the flowing water system. Once established, a tuft of vegetation impedes flow and facilitates the deposition of silt, which may alter the substratum until it becomes favourable for some other species. Eventually the vegetation may present so much resistance to streamflow that the whole mass is washed away. Many species of plants in temperate latitudes die down in winter and no longer protect silt that settled round them in summer. Copious plant growth interferes with water flow to such an extent that it often becomes necessary to cut and remove it from some streams to prevent flooding.

Animal adaptations. ***Life in the streambed.*** There is a zonation of animals in the substratum as may be seen in Figure 10. If bottom particles are not too fine, there is a rich and varied animal community down to a metre (3.3 feet) or more below the surface of the streambed. Certain crustaceans, mites, and nematodes are restricted to these deep habitats but also abundant may be tiny nymphs or larvae, the immature insects that, when they are larger, inhabit higher regions where the particles are coarser.

Stratification of life in the streambed

Nearer the surface, in sand and gravel, may be found various animals, notably segmented worms (of class Oligochaeta) and the larvae of certain families of flies (of the insect order Diptera) that, having come in from a similar habitat elsewhere, have been able to colonize running water with little change of form and habit. Nearer the surface still, but protected from the current by the large stones above, are many animals that show no structural adaptation to life in running water, though obviously without some modification of behaviour they would not survive. Most of the stoneflies (order Plecoptera), various caddis larvae (order Trichoptera) that make cases, and amphipods (shrimplike crustaceans) such as ***Gammarus*** are examples that are likely to be familiar over most of the world. Some of the stonefly nymphs are long and thin and consequently are well designed to live among the spaces between small stones and pieces of gravel. There is no rigid line of distinction between the inhabitants of the three zones mentioned.

Life at the surface of the streambed. Finally come the animals that colonize the large stones forming the surface of the streambed. Many are structurally modified for swift water environments. The freshwater limpets (family Ancylidae) have a simple conical shell that offers little resistance to the current and a foot that forms an efficient sucker. Larvae of the family Blepharoceridae, the net-winged midges, have six suckers provided with muscles that can create a powerful suction. Another similar adaptation is seen in several families of beetles and reaches its highest development in the water pennies, which belong to the family Psephenidae. The body is flat and wide, and the margin is pliable and beset with spines and hairs. It can be adjusted to fit the irregularities of the surface sufficiently to prevent any water flow underneath the larva, and the pressure on the rounded upper surface tends to press it against the substratum even though there are no suction-producing muscles. In certain Ephemeroptera (mayflies) modification of the nymph's gills produces the same effect. Such nymphs are frequently confined to running water because they cannot use their gills to create a current if necessary. For example, the mayfly genus ***Rithrogena*** has modified gills that respire less rapidly as current speed falls, whereas related genera that can move their gills consume as much oxygen in still as in flowing water. They occur on stony lakeshores and species of ***Rithrogena*** do not.

Nymphs of the mayfly family Ecdyonuridae (or Heptageniidae) have broad flat heads with veiy thin margins, and upper leg segments are of a similar shape though less acute at the margins. The body and legs are pressed against the surface, and, retaining its hold by means of sharp claws, the nymph can scuttle across a flat hard surface with great nimbleness. In the Southern Hemisphere this adaptation is found in members of the mayfly family Leptophlebiidae, which replace the Ecdyonuridae.

Adaptation to the running-water environment

Larvae of the fly family Simuliidae may perhaps be said to be the most highly adapted of all running-water animals because they not only anchor themselves in swift currents but make use of the flow to bring food. At the tip of the abdomen and on a proleg near the head are pads armed with concentric rows of hooks, of which the points can be directed outwards by pressure of the body fluid and retracted by muscles. The larvae can produce threads of silk, and these it lays over the surface of a stone, often where the current is swift. At intervals, tangled masses of silk are laid down. The larva can move across its silken lattice with a looping action, taking hold with fore and aft pads of hooks alternately. Most of the time it trails in the water with the rear hooks embedded in one of the tangled blobs of silk. It extends a rakelike apparatus, and when this has strained a mass of fine particles from the water the appendage is retracted and the mass is eaten.

Three families of caddis (of the insect order Trichoptera) utilize the current to bring food. None makes a case in the fashion of most caddis larvae but all spin nets. Members of two of these families feed on drifting organic matter, but the third is carnivorous and lives a life not unlike that of a spider on land. Members of this family occur in still water.

Adults of most beetles (Coleoptera) and bugs (Heteroptera) must come to the surface periodically to renew their store of air, a necessity that is plainly inconvenient in a medium flowing one way. It is not unexpected, therefore, that in swifter regions of rivers these two orders are represented mainly by families that have developed plastron respiration and can live permanently on the bottom (Elminthidae of the beetles and Aphelochiridae of the bugs).

Factors affecting communities in streams. Several factors influence the composition of the communities. It is impossible to separate the effects of streamflow and substratum. When current falls to a speed at which fine particles settle, the animals that inhabit the spaces between small stones and pieces of gravel are replaced by burrowing forms, of which the main groups are oligochaetes (worms), lamellibranchs (clams, mussels), and chironomid larvae (flies). Various mayfly nymphs such as the genera ***Ephemera*** and ***Caenis*** also burrow in sand and mud, and some dragonfly nymphs have also taken to this way of life, particularly in Africa.

Effects of flow velocity

A varied animal community is found in vegetation, if this develops. Animals with adaptations for life on a hard flat surface such as the mayflies of the family Ecdyonuridae or limpet-like mollusks do not occur, but less specialized forms such as the shrimplike ***Gammarus*** may continue in great numbers. ***Simulium*** (gnats and blackflies) and ***Baetis*** (mayflies) are two genera that are also abundant in both stony and weedy zones, though the species are generally different in the two.

Within one zone, rate of flow is probably not of great importance for most species since they live among stones, avoiding places exposed to the full force of the current. It is important to the net-spinning caddis larvae, which must find a place where current speed lies within a narrow range. It has been shown experimentally that a change in the pattern of flow is quickly followed by a change in the distribution of their nets.

Different kinds of rock break up into stones that differ in shape and size and texture, but whether this affects the communities that colonize them is not known. That the distribution of certain species that inhabit streambeds of fine particles is related to the size of the particles has been clearly established. In Europe the mayfly ***Ephemera danica*** inhabits sand, ***E. vulgata*** mud, and several similar examples have been described in North America. Stones that have been rounded by abrasion during transport by ice or water provide a difficult substratum colonized only by agile animals. Flat stones that move less easily harbour a more varied community, and it is richer still if, as generally happens, the stones become covered with moss.

Springs and the stretch immediately below them are often inhabited by species not found elsewhere, and some of these are confined there by their requirement of low temperature. One of the few species that has been thoroughly investigated, both in the field and in the laboratory, is the flatworm, ***Crenobia (Planaria) alpina.*** It is commonly confined to water whose maximum temperature does not exceed 15° C (59" F), though it can tolerate 25° C (77" F), provided it is not exposed to it for too long. Reproduction generally takes place only in water colder than about 12" C (54" F). Other animal species are confined to cold water not by physiological limitations but by their inability to compete with a rival in warmer water. Britain for example lacks the amphipods ***Gammarus fossarum*** and ***G. roeselii*** and instead ***G. pulex*** ranges in British streams from source to mouth. In Europe ***G. pulex*** is confined to the middle reaches of flowing waters and is apparently kept out of the upper ones by ***G. fossarum*** and the lower ones by ***G. roeselii.***

Effects of water pollution

Animal populations tend to be densest in those places where the configuration of the bottom produces areas of still water with consequent settling of debris. In addition some of the most notable changes in the composition of the population can be related to food supply. Below lakes the water is rich in floating algae and tiny drifting animals as well as suspended organic matter, and here the numbers of net-spinning Trichoptera and *Simulium* are much higher than elsewhere. They appear to exclude other species found under similar conditions of flow and substratum but without the augmented food supply. Slight enrichment by sewage leads to an increase in the number of flatworms and a decrease in the number of certain other species. It has been suggested but not proved that the flatworms eat them. In rivers enriched by the decomposition of sewage, great numbers of *Asellus* (a freshwater crustacean), leeches, mollusks and ***Sialis*** (alderflies) are found. Below the zone of enrichment these animals become progressively scarcer and are replaced by the community associated with similar conditions of flow and substratum in unenriched water. Again the full interrelationships have not been worked out. Decomposition of sewage is accompanied by lowered oxygen concentration in the polluted waters, and the typical inhabitants of stony rivers require a high concentration. There is also evidence that they are adversely affected by the bacteria that coat'the stones and indeed some of the animals themselves. In lakes, where water movement, in this case wave action, and substratum are more uniform, the most plansible explanation at present is that predation by the species favoured by enrichment leads to the elimination of the species typical of areas where there is no enrichment, and it is reasonable to postulate that this factor operates in running water too.

Life cycles of running-water animals are unexpectedly diverse and offer a rich field for experimental investigation. Nearly all the observations have been made in temperate regions. A number of animals, notably the larger ones such as leeches, snails, bivalve mollusks, and crayfish, take more than a year to reach maturity. Many are univoltine; *i.e.,* they take one year to complete development. Some species in the genera ***Baetis*** (Ephemeroptera) and *Simulium* (Diptera) and in the Trichoptera pass through two generations in a year. Quicker development is rare, ***Gammarus pulex*** being an outstanding example, completing its life cycle in less than two months.

Many of the univoltine species have a resting stage, which generally tides them over the warm period of the year. It is often the egg stage, sometimes the early nymphal stages, and sometimes, chiefly in Trichoptera, the last larval molting stage. The common European mayfly, ***Ephemerella ignita,*** apparently spends some ten months in the egg stage, at least in parts of its range; nymphs appear around midsummer, grow rapidly, and emerge in August or early September. In other species, the mayfly ***Baetis rhodani*** for example, some eggs remain unhatched for many months whereas others hatch .after a few weeks.

The phenomenon of drift

A net set so that water flows through it catches representatives of many of the species inhabiting the stream. Catches of most are much higher by night than by day and are often greatest just after dusk and sometimes again before dawn. This phenomenon, commonly referred to as drift, has been much studied and has provoked several controversies. The number of specimens drifting seems to bear some relation to the population in the stream and is at its highest during the period of most rapid growth. Therefore it probably represents to some extent a removal of surplus population. Drifting animals are likely to fall a prey to fish, but there is evidence that those that escape this fate do not travel far before regaining the safety of the bottom. Some of the larger adult caddisflies have been observed flying upstream to lay their eggs, but there is evidence that other insects do not maintain the population in the upper reaches in this way. Many have been shown to move against the current during the aquatic stage of their life cycles, particularly when small, and it is likely that uniform colonization of all suitable stretches of a watercourse is maintained in this way.

Detritus, drifting organic matter originating from dead leaves and other vegetable fragments that are blown or washed into the water from the land, is the main primary source of food in rivers and streams. The limpets, the flat nymphs of the mayfly family Ecdyonuridae, and larvae of various caddis species graze upon the algal felt that covers the upper surface of stones but probably subsist extensively on the detritus trapped in this. Detritus may also stick to the mucous trails left by flatworms or lodge in the irregularities of a rough surface. Little is known about the digestive powers of the various detritus feeders. That vegetable remains are not an easy source of food is indicated by the frequency with which they are passed twice through the same intestinal tract. It is likely that the actual source of food is the fungi and bacteria that are breaking down the vegetable tissue. The flatworms, the large stoneflies, some of the caddisflies and many fish are carnivores.

Biological productivity. In one study of total production of flesh by trout per year in the Horokiwi Stream in New Zealand, there were about 50 grams per square metre (*i.e.,* 500 kilograms per hectare, which is approximately 500 pounds per acre). An important point that came out of this investigation was the large contribution to production made by young fish, whose size was very small but whose numbers were large. Fish are comparatively easy animals with which to work, mainly because of their large size and the fact that the eggs hatch within a short period. Moreover the young do not live in the substratum. The trout establishes territories whose sizes depend on the configuration of the bottom and not on food supply. Fish, however, secure efficient exploitation of the resources of the environment by means of an indeterminate size at maturity. If food supply is poor the fish remain small, if it is good they grow much more rapidly, but sexual maturity is attained after a period that does not vary greatly with size, and the small fish reproduce just the same as normal adults. Furthermore, that calculations of production by populations of wild fish are of the right order can be confirmed by comparison with the extensive data obtained by commercial cultivators. Consequently there are reliable figures for the production by several species in the wild state.

Allen's paradox

Information about invertebrates is much less satisfactory. It was discovered how much the trout in the Horokiwi ate and it turned out to be some 17 times greater than the number of invertebrates apparently available as food. This observation, known as Allen's paradox, has been confirmed also in other parts of the world. In other words, production is underestimated. It is possible that unaccounted-for populations among running-water animals are to be found well below the surface of the substratum, where it is known that tiny nymphs abound and where most sampling instruments do not penetrate. They may also be washed down from the smallest tributaries, which are too shallow or too swift for fish to enter. This explanation of Allen's paradox is speculation, but it is plausible. What can be stated with confidence is that until an explanation has been provided and verified, calculations of production of invertebrates are premature.

Measurement of primary production by algae is difficult because any enclosing of them in containers, the technique that has been used extensively in lakes, alters the environment drastically by cutting off the current. Measurements of changes in the concentration of oxygen in a natural stream have been made, but the difficulty of ascertaining the amount of exchange between air and water has not been wholly overcome. Results suggest that primary production by algae is generally low. In contrast, that by rooted vegetation may be high. (T.T.M.)

BIBLIOGRAPHY. The following books provide information of both the general and technical nature of rivers, fluvial processes, and the action of running water on the landscape: R.J. CHORLEY (ed.), ***Water, Earth, and Man*** (1969); G.H. DURY (ed.), ***Rivers and River Terraces*** (1970); L.B. LEOPOLD, M.G. WOLMAN, and J.P. MILLER, ***Fluvial Processes in Geomorphology*** (1964).

Basic data on the biota in flowing freshwater and riverine ecology is available in: T.T. MACAN, ***Freshwater Ecology*** (1963); and H.B.N. HYNES, ***The Ecology of Running Water*** (1970).

The following papers are of a more technical nature but are worthy of investigation by the interested reader who wishes to pursue further the several aspects of river systems discussed in this article: W.B. LANGBEIN and L.B. LEOPOLD, "River Meanders, Theory of Minimum Variance," ***Prof. Pap. U.S. Geol. Surv. 422-H*** (1966); M.A. MELTON, "Methods for Measuring the Effect of Environmental Factors on Channel Properties," ***J. Geophys. Res.,*** 67:1485–1490 (1962); N.A. RZHANITSYN, ***Morphological and Hydrological Regularities of the River Net*** (1964; orig. pub. in Russian, 1960); S.A. SCHUMM, "Speculations Concerning Paleohydrologic Controls of Terrestrial Sedimentation," ***Bull. Geol. Soc. Am.,*** 79:1573–1588 (1968); R.S. SIGAFOOS, "Botanical Evidence of Floods and Flood-Plain Deposition," ***Prof. Pap. U.S. Geol. Surv. 485-A*** (1964); H. SIOLI, "Principal Biotopes of Primary Production in the Waters of Amazonia," in R. MISRA and B. GOPAL (eds.), ***Symposium on Recent Advances in Tropical Ecology, Proceedings,*** 2:591–600 (1968); L.D. STAMP (ed.), ***A History of Land Use in Arid Regions*** (1961); A.N. STRAHLER, "Quantitative Analysis of Watershed Geomorphology," ***Trans. Am. Geophys. Un.,*** 38:913–920 (1957); and V.I. ZHADIN and S.V. GERD, ***Fauna and Flora of the Rivers, Lakes and Reservoirs of the USSR*** (1963; orig. pub. in Russian, 1961).

(G.H.D./T.T.M.)

Roads and Highways

The terms road and highway define those travelled ways on which wheeled vehicles, carriage animals, and men on foot have moved throughout recorded history. The most ancient name for these arteries of travel seems to be the antecedent of the modern "way." Way stems from the Middle English wey, which in turn branches from the Latin veho ("I carry"), derived from the Sanskrit *vaḥ* ("carry," "go," or "move"). The word highway goes back to the elevated Roman roads that had a mound or hill formed by earth from the side ditches thrown toward the centre, thus "high" "way." The more recent road derived from the Anglo-Saxon rad ("to ride") and the Middle English rode or rade ("a riding or mounted journey"). In modern usage highway refers to a rural travelled way as contrasted with the urban "street" derived from the Latin strata via ("a way paved with stones"). The word road is more generally used today to describe lesser travelled ways in rural areas, primarily those carrying small amounts of traffic or being of minor importance. In more recent years the terms freeway, expressway, and motorway and similar terms in other languages (autobahn, azrtostrada) have come into use to describe highways in both urban and rural areas for which there is full control of access. On these facilities points of entrance and exit for traffic are limited and strictly controlled.

This article is concerned with the development of major road and highway systems of the world. It reviews the major elements of highway building from financing to final construction and including the activities essential to the operation and maintenance of the system.

I. History and origin of roads and highways

ROADS OF ANTIQUITY

The first road builders probably practiced their art in southwestern Asia in the area bounded by the Black and Caspian seas, the Mediterranean Sea, and the Persian Gulf. People migrated east, west, north, and south from this area; presumably in their earliest travels they recognized the necessity of improving their paths and trails to facilitate the movement of their draft animals; this made the beginnings of trade possible. The first artificial roadways may have been constructed by levelling the high ground, filling the hollows, and transferring earth from the edges of the pathway to the centre, thus forming side ditches and providing for drainage.

Wheeled vehicles were probably first developed in a broad, roughly trapezoidal area with its longer base extending from north of the Black Sea to the Caspian and its shorter base the northern end of the Persian Gulf, with Lake Van in eastern Asia Minor as the centre. The earliest wheeled vehicles have been found within 600 miles (966 km) of the lake. The oldest archaeological evidence indicates that the wheeled vehicle came into existence somewhat earlier than 3000 BC. The earliest of these were probably two-wheeled wooden carts built by the Sumerians in the forested regions south of the Caucasus and Tarsus mountains. Four-wheeled vehicles with draft poles recently found north of the Caucasus Mountains in the U.S.S.R. date from about 2400 BC. The wheeled vehicle apparently was taken westward into Europe by people who travelled up the Danube River and north to the Balkans where there is evidence for wheeled vehicles going back beyond 2000 BC (see WAGONS AND CARRIAGES).

The role of trade in growth of roads

During the Bronze Age, the development of agriculture and trade, facilitated by the domestication of the horse, marked the beginning of civilization. Trade required better roads. The first serious road builders probably were the Mesopotamians, who developed a travel route from the Babylonian Empire west and southwest to Egypt. Processional loads (700–600 BC) connecting the temples and palaces (Figure 1A) of the ancient cities of Assur, Babylon, and Tall al-Asmar were paved roadways in which burnt brick and stone were lain in bituminous mortar. Such roads, while they did not serve the normal needs of caravan traffic, may have been the forerunners of the Roman system.

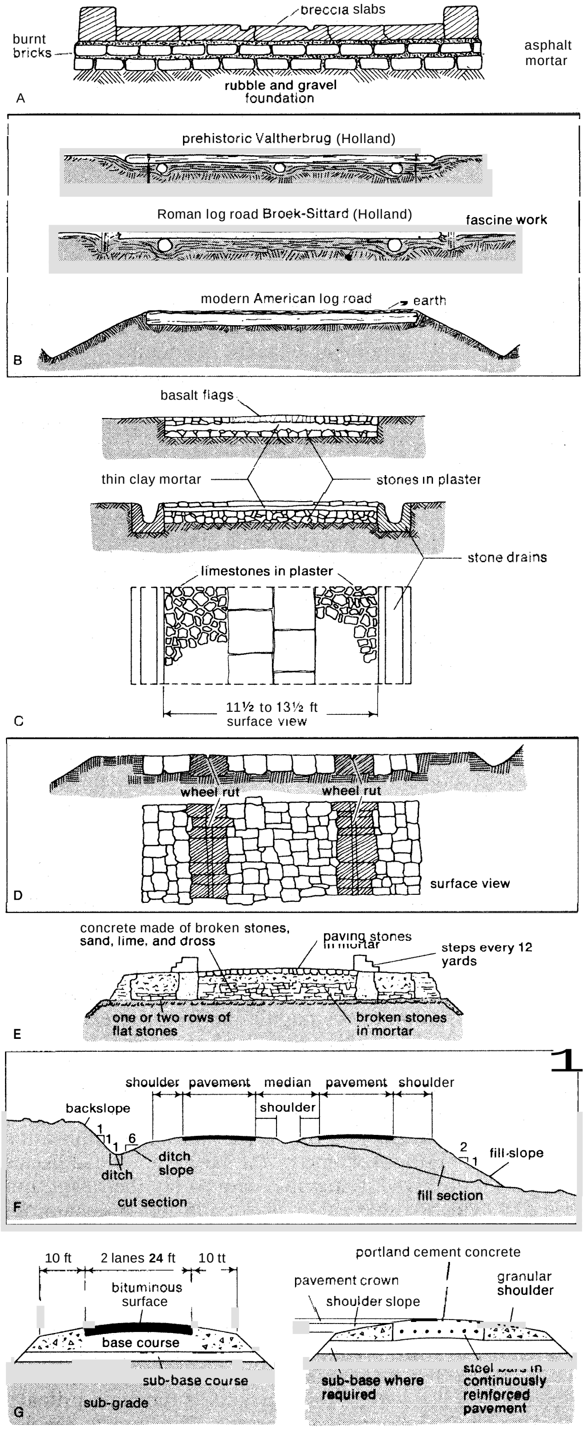


Figure 1: Cross sections of representative ancient and modern road and highway constructions (see text).
(A) Processional road in the Temple of Ishtar. (B) Log roads. (C) Ancient Cretan stone road. (D) Greek highway with wheel ruts. (E) Typical Roman road. (F) Modern concrete highway. (G) Types of modern pavement.

From (A,B,D) R J Forbes, Notes on the *History of Ancient Roads and Their Construction* (1934); N.V. Noord-Hollandsche Uitgevers-Mij., (C,E) H. Schreiber, *Sinfonie der Strasse*; Econ Verlag GmbH, Dusseldorf, West Germany

The oldest road. The modern highway systems are a natural outgrowth of the ancient road systems. The earliest long-distance road, in use from approximately 3500 to 300 BC was the Persian Royal Road, which began at Susa near the Persian Gulf, wound northwestward to Arbela and thence westward through Nineveh to Harran, a major road junction and caravan centre. The main road then continued northwestward to Samosata (modern Samsat) where it crossed the Euphrates River, and westward to Boghaz-Koi, the capital of the Hittite kings. From there travellers journeyed westward to Ancyra (modern Ankara) and Sardis where the road divided to twin termini at Smyrna and Ephesus. A branch went south from Harran through Palmyra, Damascus, Tyre, and Jerusalem to Memphis (Cairo) and another went westward from Harran to Tarsus, thence south to Caesarea and Tyre. From Susa an alternate route went westward to Ur and thence north through Babylon and Assur to Nineveh. From Boghaz-Koi one could travel northward to the Black Sea. The overland distance from Susa to Smyrna was 1,775 miles (2,857 kilometres), and Herodotus, writing in about 475 BC, put the time for the journey at 93 days.

"Amber Routes" of Europe. The earliest roads in Europe were the "Amber Routes" probably used between 1900 and 300 BC by Etruscan and Greek traders to transport amber and tin from the north of Europe to points on the Mediterranean and Adriatic. Four routes have been identified, the first from modem Hamburg southwestward by dual routes through Cologne and Frankfurt to Lyons and Marseilles. The second also passed from Hamburg south to Passau on the Danube and then through the Brenner Pass to Venice. The third began at Samland on the East Prussian Coast (where amber is still found) crossed the Vistula River at Thorn and thence continued southeastward through the Moravian Gate to Aquileia on the Adriatic. The fourth, the Baltic-Pontus road, followed the main eastern rivers, the Vistula, Saw, Sereth, Prut, Bug, and Dnieper. While these were not roads in the modern sense, they were improved at river crossings and over the mountain passes. In the same time period, evidence indicates, log roads (Figure 1B) were constructed extensively in northern Europe (modern The Netherlands, Germany, Poland, Latvia, Sweden, and White Russia) to carry traffic across wet and swampy areas. These roads were constructed by laying two or three strings of logs in the direction of the road on a bed of branches and boughs up to 20 feet (6 metres) wide and covering them with transverse logs 9 to 12 feet (2.7–3.6 metres) in length laid side by side. In the best log roads, every fifth or sixth log was fastened to the underlying subsoil with pegs. There is evidence that the older log roads were built prior to 1500 BC. The roads were maintained in a level state by covering with sand and gravel or sod, and the Romans used side ditches to reduce the moisture content and increase the carrying capacity.

Log roads

Imperial roads of China. The ancient road system of China, which paralleled in time the Royal Road, was a substantial and remarkable system. The Imperial roads played the same role in southeastern Asia as the Roman roads in Europe and Asia Minor. Many of the Chinese roads were wide, well built, and surfaced with stone; rivers were crossed by bridges or well-managed ferries; steep mountains were traversed by stone-paved stairways with broad treads and low steps. The Imperial road system, about 2,000 miles (3,200 kilometres) in total length, radiated from Sianfu, Nanking, and Ch'eng-tu. Except for short periods of time, however, the roads of China were not adequately maintained and fell into disrepair; it was said that in China a road was good for seven years and then bad for 4,000 years. Chinese roads differed markedly from the Roman roads in their crookedness, particularly in hilly areas.

Early roads of Malta and Crete. Unique "rut" roads were built on the island of Malta probably about 2000 to 1500 BC. These consisted of two V-shaped grooves about 4% feet (1.35 metres) apart cut into the coral sandstone of the island. The roads apparently were traversed by carts, drawn by human power, with the wheels running in the grooves. During the Minoan civilization on the island of Crete (3000 to 1100 BC), a road (Figure 1C) was built from Gortyna on the south coast over the mountains at an elevation of about 4,300 feet (1,300 metres) to Knossos on the north coast. Constructed of layers of stone, the roadway took account of the necessity of drainage by a crown throughout its length and even gutters along certain sections. The pavement was about 12 feet (3.6 metres) wide and the central portion consisted of two rows of basalt slabs two inches (50 millimetres) thick. The centre of the roadway seems to have been used for foot traffic and the edges for animals and carts. Guard houses were located at frequent intervals along the road.

India. The Indus civilization in Sindh, Baluchistan, and the Punjab probably flourished in the period 3250–2750 BC. Excavations indicate that the cities of this civilization paved their major streets with burned bricks cemented with bitumen. Great attention was devoted to drainage. The houses had drain pipes that carried the water to a street drain in the centre of the street, two to four feet (about a metre) deep and covered with slabs or bricks.

Evidence from archaeological and historical sources indicates that by AD 75 several methods of road construction were known in India. These included the brick pavement, the stone slab pavement, a kind of concrete as a foundation course or as an actual road surface, and the principles of grouting (filling crevices) with gypsum, lime, or bituminous mortar. Street paving seems to have been common in the towns in India at the beginning of the Christian era and the principles of drainage were well known. The crowning of the roadway and the use of ditches and gutters was common in the towns. Northern and western India in the period 300 to 150 BC had a network of well-built roads. The rulers of the Maurya Empire (4th century BC), which stretched from the Indus to the Brahmaputra and from the Himalayas to the Vindhya Range, generally recognized that the unity of a great empire depended on the quality of their roads. The Great Royal Road of the Mauryans began at the Himalayan border, ran through Taxila (near modern Rāwalpindi), crossed the five streams of the Punjab, and continued by way of Jumna to Prayag (now Allahābād). A "Ministry of Public Works" was responsible for construction, marking, and maintenance of the roads and rest houses and for the smooth running of the many ferries that carried the Royal Road across the wide rivers.

Egypt and Greece. Although Herodotus credits the Egyptians with building the first roads to provide a solid track upon which to haul the immense limestone blocks used in the pyramids, archaeological evidence indicates that road-building technology travelled southwest from Asia toward Egypt. The wheel arrived in Egypt at the relatively late date of about 1600 BC. There is little evidence of street surfacing in ancient Egyptian towns, though there is evidence of the use of paved processional roads leading to the temples. The ancient travel routes of Egypt ran from Thebes and Coptos on the central Nile east to the Red Sea and from Memphis (Cairo) across the land bridge to Asia Minor.

The early Greeks depended primarily on sea travel. There is evidence of the building of special roads for religious purposes and transport about 800 BC, but there is little evidence of substantial road building for travel and transport prior to the Roman system. The Greeks built a few ceremonial, or "sacred," roads, paved with shaped stone and containing wheel ruts (Figure 1D) about 55 inches (1.4 metres) apart, very similar to those built in Malta at a much earlier date.

Roman roads. The first scientific road builders were the Romans. By the peak of the Empire the Romans had built nearly 53,000 miles (85,000 kilometres) of road connecting the capital with the frontiers of the far-flung empire. Twenty-nine great military roads, the *viae militares,* radiated from Rome. The most famous of these was the Appian Way, started in 312 BC, following the Mediterranean coast south to Capua, then turning eastward to Beneventum, where it divided into two branches, both reaching Brundisium (Brindisi). From Brundisium the Appian Way traversed the Adriatic coast to Hydruntum, a total of 410 miles (660 kilometres) from Rome.

There are differences of opinion concerning the origin of the Roman road-building methods, but the consensus credits the Etruscans of northern Italy as Rome's principal teachers, though the Cretans, Carthaginians, Phoenicians, and Egyptians probably also contributed.

Roman road-building practice

Roman roads were remarkable for preserving a straight line from point to point regardless of obstacles. They were carried over marshes, lakes, ravines and mountains, and by their bold conception they have excited the admiration of modern engineers. In its highest stage of development the Appian Way was constructed by excavating parallel trenches about 40 feet (12 metres) apart to mark its exact location and to indicate the nature of the subsoil (Figure 1E). The foundation was then covered with a light bedding of sand or mortar on which four main courses were constructed; (1) a *statumen* layer of large flat stones 10 to 24 inches (250–600 millimetres) in thickness; (2) a *rudus* course of smaller stones mixed with lime about 9 inches (225 millimetres) thick; (3) the *nucleus* layer, about one foot (300 millirnetres) thick, consisting of small gravel and coarse sand mixed with hot lime; and (4) on this fresh mortar a *summa crusta*, or wearing surface, of flint-like lava about six inches (150 millimetres) deep. The total thickness thus varied from 3 to 5 feet (0.9 to 1.5 metres). The width of the Appian Way in its ultimate development was 36 Roman feet (35 English feet or 10.5 metres). The two-way central lane, heavily crowned, was 15% feet (4.7 metres) wide flanked by curbs 2 feet (0.6 metre) wide and 18 inches (0.45 metre) high on each side and paralleled by one-way side lanes 7% feet (2.3 metres) wide. This massive Roman road section adopted about 300 BC set the standard of practice for the next 2,000 years.

Classes of Roman roads

The public transport of the Roman Empire was divided into two classes: (1) *Cursus rapidi,* the express service and (2) *agnarie*, the freight service. In addition there was an enormous amount of travel by private individuals. The most widely used vehicles were two-wheeled chariots drawn by two or four horses and its companion, the two-wheeled cart used in rural areas. A four-wheeled *raeda* in its passenger version corresponded to the stage coaches of a later period and in its cargo version to the freight wagons. Fast freight *raedae* were drawn by eight horses in summer and ten in winter and, by law, could not haul in excess of 1,000 Roman pounds or 330 kilograms. Speed of travel ranged from a low of about 15 miles (24 kilometres) per day for freight vehicles to 75 miles (120 kilometres) per day by speedy post drivers.

The Silk Road. The trade route from China to Asia Minor and India, known as the Silk Road, had been in existence for 1,400 years at the time of Marco Polo's travels (*c.* AD 1270–90). but during this entire period incessant warfare and raiding by the Mongols and other nomads had kept it closed to traffic for all but 400 years. At its zenith in AD 200 this road and its western connections over the Roman system constituted the longest road on earth. The extreme western terminal of the Silk Road was at Gades (modern Cadiz, Spain) on the Atlantic Ocean; thence it ran northeastward across the Pyrennees north of Tarraco (modern Tarrasa) and around the shore of the Mediterranean through Genoa to Rome; from Rome south the road followed the Appian Way to Brundisium, where merchants took ship across the Adriatic, picking up the road again to cross the Balkan Peninsula to Byzantium (Istanbul) and continue southeastward through Ancyra (Ankara), with an alternate route through Antioch to Rhagae (near modern Tehrān) thence westward through Meshed, Bokhara, and Samarkand to Ferghana (Ush, the stone tower). From Ferghana the road traversed the valley between the Tien Shan and Kunlun Mountains through Kashgar where it divided and skirted both sides of the Takla Makan Desert to join again at Ansi. The road then wound eastward to Chia-yii-kwan (Su-chou) where it passed through the westernmost (Jade Gate or Yumen) gateway of the Great Wall of China. It then went southwest on the Imperial Highway to Sian and eastward to Shanghai on the Pacific Ocean. The Silk Road and its western connections bounds a globular rectangle 20" of latitude in height and 128" of longitude in breadth, a travel distance of 8,000 miles (12,800 kilometres) from Cadiz to Shanghai, for more than 2,000 years the longest road on earth. From Ferghana trade routes to the South passed over the mountains to the great trading centre of Eactria and to northern Kashniir.

Decline of roads: AD 200–1800. At the zenith of the Roman Empire overland trade joined the cultures of Europe, North Africa, Asia Minor, China, and India. But the system of road transport was dependent on the Roman, Chinese, and Mauryan empires, and as these great empires declined in the early Christian era the trade routes became routes of invasion. The road networks nearly everywhere fell into centuries of disrepair. Transport wagons gave way to pack trains, which could negotiate the badly maintained roads acd sufficed to carry the reduced stream of commerce. Eventually a commercial revival set in; by the 12th century old cities were reviving and new ones were being built, especially in western Europe. Some of the larger towns paved their principal streets. There was an awakened interest in better overland travel, better protection of merchants and other travellers, and the improvement of roads. Public funds, chiefly derived from tolls, were committed to road upkeep. The corvée, or road-labor tax, made an even more substantial contribution. Long-distance overland commerce increased rapidly and included a restoration of the trade route between Europe and China through Central Asia that Marco Polo travelled in the late 13th century.

Early paved town streets

During the 14th century the Black Death and other disasters brought a slowdown. During the 15th and 16th centuries street paving became more popular and wheeled vehicles increased in number and quality.

Inca roads of South America. Across the Atlantic,, the period witnessed the rise of another notable road-building empire, that of the Incas. The Inca road system, extending from Quito, Ecuador, to points south of Cuzco, Peru, consisted of two parallel roadways, one along the coast about 2,250 miles (3,600 kilometres) in length, the other following the Andes about 1,650 miles (2,640 kilometres) in length with a number of cross connections. At its zenith when the Spaniards arrived early in the 16th century, it served an area of about 750,000 square miles in which lived nearly 10,000,000 people. Some of the original Inca system was still in use in the 1970s. The Andes route was remarkable. The roadway was 25 feet (7.5 metres) wide and traversed the loftiest ranges with cutbacks and easy gradients. It included galleries cut into solid rock and retaining walls built up for hundreds of feet to support the roadway. Ravines and chasms were filled with solid masonry and suspension bridges with wool or fibre cables crossed the wider mountain streams. The surface was of stone in most areas and asphaltic materials were used extensively. The steeper gradients were surmounted by steps cut in the rocks. Traffic consisted entirely of pack animals (llamas) and people on foot; the Incas lacked the wheel. Yet they operated a swift foot courier system and a visual signalling system along the roadway from watchtower to watchtower. Interestingly, Inca roads resemble those of ancient China, which, it has been suggested, may indicate a direct cultural influence.

The Andes route

THE BIRTH OF MODERN ROAD BUILDING

The 17th and 18th centuries saw crude carts and wagons operating over rough, unimproved roads give way to regularly scheduled common-carrier stagecoaches and freight wagons running on stone-surfaced toll roads. The first engineering school in Europe, École des Ponts et Chaussées (the School of Bridges and Highways), was founded in Paris in 1747. Late in the 18th century Adam Smith in discussing conditions in England wrote,

> Good roads, canals, and navigable rivers, by diminishing the expense of carriage, put the remote parts of the country mote nearly upon a level with those in the neighbourhood of a town. They are upon that account the greatest of all improvements.

In the last half of the 18th century the fathers of modern road building appeared in France and England.

Up to this time the roads built had utilized, with minor modifications, the very heavy Roman cross section. In France in 1764, Pierre-Marie-Jérôme Trésaguet, an engineer from an engineering family, became engineer of bridges and roads at Limogs and, in 1775, inspector general of roads and bridges for France. In that year he developed an entirely new type of relatively light road surface, based on the theory that the subsoil, rather than the surface should support the load. His standard section, 10 inches (250 millimetres) thick, consisted of a course of uniform stones laid edgewise covered by a layer of walnut-sized broken stone. The roadway crown rose six inches (150 millimetres) in its 18-foot (5.4 metres) width and had a uniform cross section.

John Metcalf, the first of England's pioneer road builders, was a contemporary of Trésaguet. Born in 1717, he was blinded by smallpox at the age of six, but became an expert climber, horseman, and swimmer. About 1754 he launched a stagecoach route between Knaresborough and York and in 1765 built a portion of the turnpike authorized between Harrowgate and Boroughbridge. Over the next 37 years he built more than 180 miles (290 kilometres) of English turnpike roads and bridges. In his road building he emphasized the use of ditches for adequate drainage and special precautions for distributing the load by using baled brush as a subbase in marshy areas.

Thomas Telford, born of poor parents in Dumfriesshire, Scotland, in 1757, was apprenticed to a stone mason; intelligent and ambitious, he progressed to designing bridges and building roads. His Carlisle-Glasgow road was considered the finest road ever built up to that time (1816). Telford placed great emphasis on two features: (1) maintaining a level roadway with a maximum gradient of 1 foot in 30 feet and (2) building a stone-surfaced roadway capable of carrying the heaviest anticipated loads. His roadways were 18 feet (5.4 metres) wide, crowned only four inches (100 millimetres) and built in two courses, a lower course of 7 inches (175 millimetres) consisting of good quality stone carefully placed by hand (Telford base) and a second layer consisting of 7 inches of broken hardstone of 2% inch (62.5 millimetres) maximum size and a layer of gravel one inch (25 millimetres) thick.

The work of McAdam

John Loudon McAdam, born in 1756 at Ayr, Scotland, began his road-building career in Bristol, then the second city in England. The roads surrounding Bristol were in poor condition and in 1816 McAdam, chosen general surveyor of the Bristol municipality, had an opportunity to test his theory that road building could be reduced to a science based on fundamental principles. Like Trésaguet, he believed that a well-drained, compacted subgrade should support all the load while the stone surfacing should act only as a wearing surface and a roof to shed water. There is evidence that McAdam was indebted to others for many of his ideas. McAdam insisted on a roadway with adequate side ditches and a subgrade elevated above the surrounding ground surface and compacted with a crown of 3 inches (75 millimetres) in 18 feet (5.4 metres) to drain surface water rapidly. He believed that 10 inches (250 millimetres) of surfacing was adequate for any load of his time. His stone surfacing utilized two inch (50 millimetres) maximum size stone laid in loose layers and compacted under traffic. Weak spots were detected and replaced before the next layer was placed. Compaction under the wheeled traffic of the early 1800s proved to be very effective. McAdam's remarkable success was, however, due in large measure to his efficient system of administration.

By 1820 Britain had 125,000 miles (200,000 kilometres) of road, of which 20,000 miles (32,000 kilometres) were turnpikes. By 1836, 3,000 coaches operated on these roads; the rapid development of the railroads, however, brought road building virtually to a halt. Roadway improvements for the next 60 years were essentially confined to city streets.

Continental Europe and the U.S. had road-building histories virtually identical to Britain's, with a period of rapid construction of the new lightweight Trésaguet-McAdam type of roads followed by a sudden halt as railroads intruded on the transportation scene. The first engineered and planned road built in the United States was a privately constructed toll turnpike from Philadelphia to Lancaster, Pennsylvania, built between 1793 and 1794 at a cost of $465,000. Its 62-mile (100 kilometres) length with 9 toll-gates was surfaced with broken stone and gravel with maximum grades of 7 percent. The Cumberland Road, also known as the National Pike, was an even more notable road-building feat. It opened for traffic between Cumberland and Wheeling, West Virginia in 1818, and to Springfield, Ohio and part of the way to Vandalia, Illinois in 1838; total cost was about $6,825,000. The road was maintained by government appropriations and by tolls collected by the states. Specification requirements called for 2 66-foot (20 metres) right-of-way completely cleared. The roadway was to be covered 20 feet (6 metres) in width with stone 18 inches (450 millimetres) deep at the centre and 12 inches (300 millimetres) deep at the edge, the upper 6 inches (150 millimetres) was to consist of broken stone of 3-inch maximum size and the lower stratum of stone of 7-inch maximum size.

During the period 1840 to 1910, the era of railroad building all over the world, country roads everywhere remained virtually impassable in wet weather. The initial stimulus for a renewal of road building came not from the automobile, whose impact was scarcely felt before 1900, but from the bicycle, for whose benefit road improvement began in many countries during the 1880s and 1890s. Though the requirements of the lightweight, low-speed bicycle were satisfied by the old "macadamized" surfaces as the world entered the 20th century, the horseless carriage rather quickly rendered this type of road obsolete.

Road administration and financing

The responsibility for financing and building roads and highways has been both a local and a national responsibility of the nations of the world during many centuries. It is notable that this responsibility has changed along with political attitudes toward road building. English road building, for example, for centuries remained entirely local despite clear evidence that local responsibility was not providing adequate roads. Local authorities and private turnpike trusts dominated both British road building and maintenance throughout the 19th century, though the national government edged into the picture through increasing grants of funds, climaxed by the establishment in 1909 of a national Road Board authorized to construct and maintain new roads and to make advances to highway authorities to build new or improve old roads.

Except for the National Pike, early highway building in the United States was also carried on by local government. Toll roads, variously surfaced, were constructed in the first half of the 19th century under charters granted by the states. The Congress made a number of land grants for the opening of wagon roads but exercised no control over the expenditure of funds and, as in Britain, little road building was accomplished. Road labour to satisfy taxes was both unpopular and unproductive.

II. The automobile road

BASIC PROBLEMS

The automobile, and a little later the heavy truck, introduced totally new requirements for road and highway construction. Vehicle speeds increased rapidly; roadway alignment suitable for horse and buggy travel was completely inadequate, as were road surfaces, whose stones were torn loose by the heavily loaded tires. The early trucks, built with solid rubber tires, carried gross loads of 12,000 to 14,000 pounds (5,400–6,300 kilograms), which by the close of World War I had risen to 28,000 pounds (12,600 kilograms). The development of the pneumatic tire substantially reduced the destructive effect of truck loading on thin pavements intended for horse and buggy, but it was obvious that much stronger surfacing was required. Loads continued to increase, to gross weights of 40,000 pounds (18,000 kilograms) on tandem axles.

In western Europe and the United States, the world's principal automobile areas in the 1960s, approximately half of the total vehicle mileage travelled was travelled in urban areas. There was substantial amount of long distance travel both for business and pleasure, and the tonnage of intercity freight carried by trucks was increasing steadily, bringing a powerful demand for direct, long-distance express routes, including suitable links and bypasses in the metropolitan areas. In the late 1960s and early '70s increasing attention was being turned to the problem of safety in design of highways and their anxiliary equipment.

Highway planners and designers have had to take into account the speed and operating characteristics of the motor vehicles, wheel or axle loads, and the density and composition of vehicles in the traffic stream, as well as the safety, comfort, and convenience of the travelling public.

Classification and design standards

Depending upon the volume of traffic, composition of traffic, and major purpose, roads and highways may be divided into four functional classifications: (1) local roads and city streets; (2) collector and feeder roads, and secondary rural highways; (3) primary highways that carry relatively high volumes of traffic between population centres; and (4) expressways that serve major traffic flows.

In order to have reasonable uniformity in a given jurisdiction, design standards are usually established for each functional classification, taking into account the type of terrain in which the road or highway will be built (or rebuilt) classified as flat, rolling, or mountainous. Design standards are usually established on the basis of average daily traffic volumes, and it is common practice to set both a minimum standard and a desirable standard (Figure 1F). Design standards commonly provide for a right of way width, the speeds for which the roadway is to be designed, maximum permissible sharpness of horizontal cuives, maximum permissible vertical gradient in feet rise per 100 feet of horizontal distance, minimum width of roadway and surfacing (pavement), minimum nonpassing sight distance (the distance a driver a normal distance abcve the roadway can see an object six inches [150 millimetres] high on the roadway ahead), and the clearance and capacity of bridges. For the two higher functional classifications most nations have governmental or quasi-governmental agencies that establish design standards.

TYPES OF PAVEMENT

Pavements were first developed for use on city streets. The earliest city pavements were stone block, wood block, vitrified brick, and bitumen (*e.g.*, natural asphalt). The first bituminous pavement was laid in Paris in 1854 using a natural rock asphalt from Switzerland. The first portland cement concrete pavement was built in Inverness, Scotland, in 1865. At the beginning of the automobile era rural road surfacing, where it existed, consisted of broken stone or gravel. Such roads were too rough and dusty, and inadequate in strength for automobile traffic.

Highway pavement may be defined as the portion of the highway cross section above the natural earth or subgrade. Figure 1G illustrates a typical pavement cross section for a divided highway in a rural area and the elements (surfacing, base, and subbase) that make up the cross section. In some pavement sections not all of the three elements will need to be used and in others a given element may consist of several layers of differing materials.

Flexible pavement. Pavements are divided into two types: flexible and rigid. A flexible pavement consists of base and subbase layers of natural aggregate materials (sand and gravel), or crushed stone, and a surfacing of aggregates mixed with a bituminous material, commonly an asphalt extracted from petroleum or coal tar. In some areas of the world natural mixtures of asphalt and rock, called "rock asphalts," occur; these make excellent surfacing materials. Base and subbases for flexible pavements may also be made of "stabilized materials," in which natural soils or poor quality sand and gravel are mixed with such materials as lime, portland cement, chemicals, asphaltic oils, and coal tars in order to increase the strength of these materials and reduce their susceptibility to loss of strength with increasing moisture contents. The terms soil-lime or lime stabilization, soil-cement or cement stabilization, soil-asphalt or asphalt stabilization are used to refer to mixtures of this type. Beneficial effects can be obtained in many cases by mixing two soils, or a soil and an aggregate, together to instill the good qualities of both in the finished mixture.

The surfacing portion of a flexible pavement may be produced by a range of processes that yields surfaces of varying texture, thickness, strength, and quality. The major processes described below, are the surface treatment, macadam, mixed-in-place, and plant-mix types. The latter three processes are used for both the base and surfacing portions of the pavement structure.

Surface preparation

Surface treatment. In this installation the pavement is placed over a completed, compacted base course. The first step is to cover the base course with a thin asphaltic oil or tar sprayed on in sufficient quantity to fill cracks and crevices in the base without leaving excess oil or tar on the surface. After this prime coat has penetrated, the area is sprayed with a harder asphaltic oil or tar and covered with a layer of uniform-size gravel or stone chips, which is rolled to seat it in the bituminous material. A second, third, and even fourth layer of oil and stone may be applied to increase the pavement thickness. Surfacings so constructed are called single, double, triple, and quadruple surface treatments. Such surfaces are adequate for low volumes of traffic particularly in relatively arid climates.

Macadam construction. In this type of construction the surfacing is constructed by placing a layer of uniform size crushed stone or gravel in the size range of one inch (25 millimetres) to three inches (75 millimetres) on the completed and compacted base course. The layer thickness is somewhat greater than the maximum-size aggregate used. The stone layer is thoroughly compacted and is then bound together by one of several processes. In water-bound macadam a layer of stone screenings is worked into the surface by rolling, after which the surface is sprinkled with water and rolling is continued; the stone dust-water mixture forms a natural cement to bind the stone together. Water-bound macadam will not stand the abrasive effects of modern traffic and is seldom used today except for base courses. In penetration macadam, also referred to as asphalt or tar macadam (tarmac), the stone is impregnated with a substantial quantity of semisolid asphalt cement or tar heated to fluid temperaure and sprayed into the compacted stone. A layer of stone of such size as to fill the interstices in the first course is then rolled on. If the penetration macadam is the surfacing, it is completed with the application of a single surface treatment. In cement-bound macadam a cement-sand slurry is worked into the interstices of the compacted stone to provide the necessary cementing action.

Mixed-in-place surfacing. Mixed-in-place (road-mix) surfacing involves mixing aggregates in the roadway cross section with bituminous or other cementitious materials in order to obtain watertight and stronger surfaces. This kind of processing is used extensively for base courses and to a limited extent for subbase courses. In the 1920s and 1930s when many of the highways in western sections of the United States were surfaced in this manner, the natural gravel or stone surfaces were loosened and thoroughly mixed, after which asphaltic oils or liquid tars were added by spraying and worked into the loose aggregate with graders until uniformity was obtained. The resultant mixture was then compacted. Surfaces of this type last several years under light to moderate traffic and are easily repaired by loosening, adding material and recompacting.

Plant mix. Plant mix surfacings possess the necessary strength and waterproofing to carry the highest volumes of traffic and the heaviest wheel loads under severe climatic conditions. They also provide a high quality riding surface. The process involves the assembly of excellent aggregates in several sizes, from walnut size to dust, drying these aggregates, heating to temperatures of 300–400" F (150–200" C) and mixing, in a central plant, with the proper quantity of asphalt cement or semisolid tar also at elevated temperature. The resulting compound is hauled to the roadway where it is placed by a laying machine or paver and thoroughly rolled before the mixture cools. Such mixtures are placed in thicknesses varying from one to four inches (25–100 millimetres) and a given surfacing may consist of two or more layers. The resulting surface is smooth and if properly designed provides good frictional resistance for vehicles operating

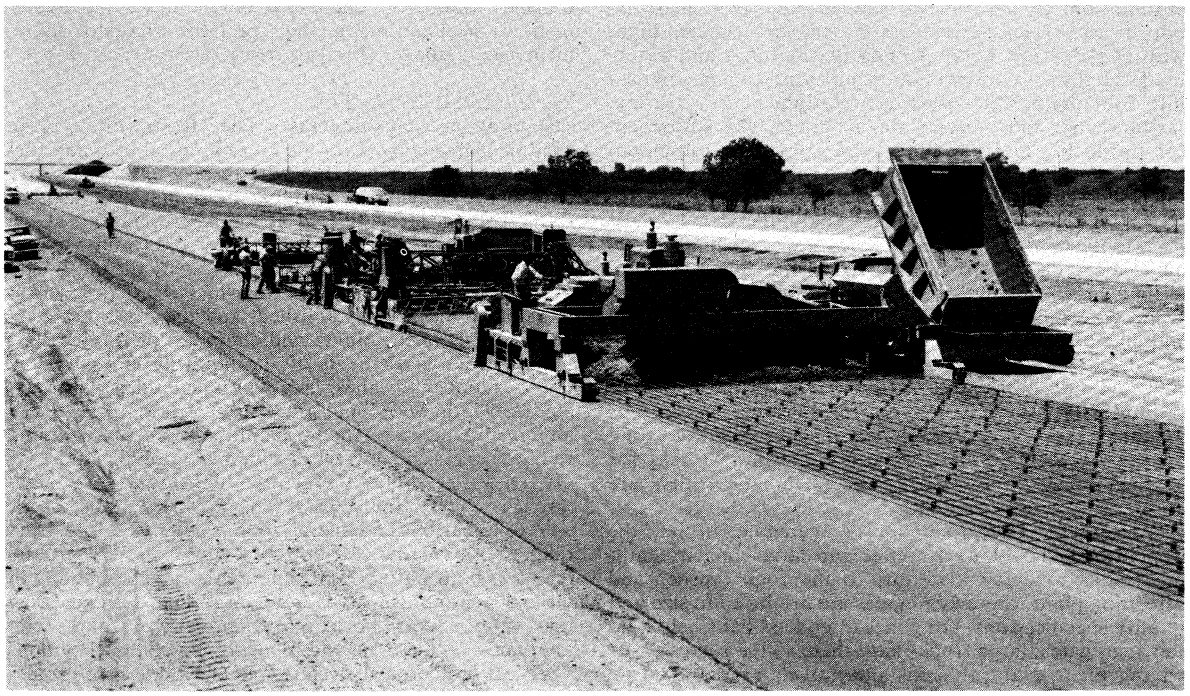


Figure 2: Three stages of concrete paving.
(Left) The gravel subbase is laid, compacted, and stabilized; (centre) the base is covered with concrete; (right) reinforcing steel mesh is overlaid and (not shown) a final layer of concrete is added.

BY courtesy of Rex Chainbelt Inc.

over it. Its repair is simple and the surface can be upgraded easily by adding another layer mixed and placed in the same manner.

Rigid pavement. Rigid pavement is portland cement concrete surfacing placed directly on the subgrade or on a subbase course or base course. The pavement thickness is in the range of 6 to 12 inches (150–300 millimetres) and is dependent on the volume and weight of truck traffic using the highway. This type of surfacing is produced by assembling graded aggregates and cement at a central location where they are carefully proportioned on a batch basis and then mixed with water to form concrete that will harden and yield adequate strength.

As the concrete changes from a plastic mass at placement to a hardened surfacing it undergoes a decrease in volume referred to as shrinkage. This shrinkage is accompanied by a tendency for the concrete to be pulled across the underlying layer, thus developing tensile stresses. In a continuous concrete mass the tension so developed exceeds the strength of the concrete and cracking occurs. The hardened concrete is also subject to volume changes and warping due to daily and seasonal temperature and moisture variations. As the temperature rises (or moisture content increases) there is an expansion of the surfacing while lower temperatures (or a decrease in moisture) cause contraction. In order to control these volume changes and the consequent cracking of the portland cement concrete pavement, three types of planned joints may be introduced. Contraction joints, spaced at intervals across the roadway, consist of grooves that extend partway through the pavement to insure that, when the concrete contracts, cracking will occur at these locations. Expansion joints, also transversely laid, consist of narrow openings through the pavement to provide room for expansion of the concrete. Elaborate measures are taken to keep these expansion joints water tight, and steel-rod dowels are provided across the joint to transfer part of the wheel loads from one slab end to the other. The third joint type is a longitudinal-tied joint used at the edge of the lanes and made by parting the pavement, as for contraction joints, and inserting steel tie bars under the parting strip to prevent the joints from opening. Such joints relieve the stresses caused by warping of the pavement slab due to temperature and moisture variations between the top and bottom. In the morning of a summer day the top of an 8-inch (200-millimetre) concrete pavement may be 25° F (14" C) hotter than the bottom. The weakened planes for the contraction and warping joints of concrete pavements were originally formed by hand methods, using a metal or fibre strip in the fresh concrete to produce the cut. However in recent years these joints have been formed after the concrete hardens by cutting the concrete to the desired depth and width with an abrasive circular saw blade.

Rigid pavement jointing

Maintenance problems of concrete pavements

Many of the maintenance problems in portland cement concrete pavements are associated with the jointing systems. It is difficult to keep the joints adequately sealed against water, dirt, and dust, and it is difficult to place dowel bars so that they permit expansion joints to open and close readily. For these reasons the continuously reinforced concrete pavement that does not require a jointing system has become increasingly popular. In this pavement (Figure 2) a continuous layer of steel bars is placed longitudinally at mid-depth of the pavement slab. The bars provide a steel area of about 0.5 percent to 0.7 percent of the cross section of the pavement. Shrinkage and temperature and moisture volume changes cause the concrete slab to crack transversely at intervals of 3 to 10 feet (1 to 3 metres) but the longitudinal steel absorbs the tension in the concrete so that the cracks are held tightly closed and no surface water can pass through. A small amount of transverse steel is used to hold the longitudinal bars in place and control longitudinal cracks. This method of construction permits the use of continuous slabs, several miles or kilometres long, with no jointing. Expansion and contraction of several inches may occur at the terminal ends of the continuously reinforced pavement and must be provided for in the design.

In deciding whether to use flexible or rigid pavement, engineers must take into account initial cost, probable life, probability of traffic disruptions for maintenance, riding characteristics, ease of repair, climatic conditions with their probable effects, and service to the travelling public.

MACHINERY AND EQUIPMENT

Highway construction in the developed nations has benefitted from rapid and spectacular developments in construction equipment technology. These developments have been marked by substantial increases in capacity per unit and decreases in manpower requirements through automation. In the underdeveloped countries where labour is less expensive and less skilled, sophisticated highway construction equipment is uneconomic and rarely used. Highway construction equipment can be divided into four major categories; (1) equipment for clearing, earthmoving, and building the subgrade; (2) equipment for producing and handling aggregates; (3) equipment for mixing and placing pavements, and (4) equipment for bridge construction.

Clearing and earth moving. The bulldozer is most commonly used for clearing vegetation and undesirable materials from the roadway. Earth moving is accomplished with bulldozers, hauling scrapers, and motorgraders. Compaction is accomplished with large tamping and pneumatic-tired rollers; sprinkling trucks are used to adjust the moisture content for compaction. Rock cuts and fills utilize the mobile wagon drill capable of drilling 200 feet (60 metres) for blasting, shovels, or draglines (very large excavating machines) for loading the excavated material, and very large dump trucks for hauling. Grid, steel-wheeled, and pneumatic rollers are used for rock compaction.

Aggregates. The production of aggregates utilizes the wagon drill, shovel, or drag line and large-capacity hauling trucks to carry materials to the rock crusher and screening plant where aggregates are produced to size and quality specifications. Draglines or endless belts help load the aggregates; large trucks haul them to the roadway, or paving plant. Natural sands and gravels are handled by draglines, washed if necessary, and screened to the desired size gradation. Compaction of base and subbase aggregates is accomplished by steel-wheeled rollers and pneumatic rollers.

Mixing pavements. Surface treatment and macadam pavements utilize a bituminous distributor that sprays the bituminous material, hot or cold, at the proper rate. Both steel-wheel and pneumatic-tire rollers are used for compaction. The roadway is swept clean by a rotary broom. Mixed-in-place pavements utilize truck-water distributors, bituminous distributors, blade graders, and special mixers. Alternately, road-mix machines excavate the roadway surface, apply the bituminous material or other liquid stabilizer, mix the materials, and return the mixture to the roadway. Special hauling units to transport and spread lime and cement are used in stabilization.

Bituminous paving

Bituminous paving operations involve equipment to assemble aggregates, storage tanks for asphalt and tar, cold bins for aggregate loaded by dragline or belt, a mechanical cold-materials feeder, a cylindrical drier to dry and heat the aggregates, hot bins to proportion the hot aggregates, an asphalt or tar metering system, and a mill *to* perform the mixing operation. Recent developments have eliminated hot bins in many plants. The mixed material is hauled to the roadway in dump trucks and placed in a paving machine that distributes it uniformly on the base course. Compaction is accomplished by steel-wheeled and pneumatic rollers.

In portland cement concrete-paving plants, aggregates are assembled, placed in bins by size, and proportioned by weighing; the cement is separately weighed and added to the batch after which the materials are loaded into a stationary or truck-hauled mixer where water is added and fresh concrete produced. If a central mixer is used the completed mixture is hauled to the roadway in agitator trucks or in ordinary trucks with a bathtub-type body. The concrete mixture is distributed on the compacted roadway and placed by a paver that forms and smooths the concrete. Steel side forms that were necessary a few years ago have been eliminated by the formless paver.

The steel in continuously reinforced pavements is made up in sections by tying the longitudinal bars to the transverse bars. It may be placed on the compacted roadway ahead of placement of concrete or continuously with the fresh concrete. Joints are cut with heavy duty circular saws. Dowel bars for contraction and expansion joints in plain concrete pavements are held in place by heavy metal supports.

***Bridge** construction.* Bridge building or viaduct building is a highly specialized operation using pile drivers, cranes, forming for concrete, riveting and welding equipment for steel and many other specialized types of equipment (see BRIDGES, CONSTRUCTION AND HISTORY OF).

DESIGN AND CONSTRUCTION

Highway location **and plans.** The design of a given road or highway involves the consideration of many factors and numerous decisions with respect to the necessary criteria for design. The designer must first establish the traffic volume to be carried at the beginning and at the end of the roadway's probable life. The number and loaded weight of trucks using the highway during its life must be determined. A speed for which the highway is to be designed must be established, and the maximum gradient decided. The volume and character of traffic determine the design elements of the highway cross section, that is, number of lanes, lane width, shoulder width and median width. Speed and gradient determine the vertical alignment, and speed the horizontal alignment, including radii of curves and degrees of superelevation on curves.

It is next necessary to match the highway to the terrain through which it must pass. The first step, thus, is the establishment of the general route. When this has been refined to a narrow corridor a map showing the ground features, both natural and man-made, and variations in ground contours must be prepared. The designer then lays out the exact horizontal alignment or route, with alternates, on the map, and by using the contours obtains a vertical profile on the centre line and transverse profiles at desired intervals. A grade line is then established and earth work quantities determined by comparing the finished roadway cross section with the ground cross section. The grade line is adjusted to balance the earth to be excavated with the fills to be made, to provide for moving the least amount of earth, and to satisfy the requirements for maximum gradient and minimum sight distances for vehicle operations. In setting the horizontal and vertical alignment the designer attempts to make the highway flow with the terrain so as to minimize the sense of tension with the driving environment.

New equipment, including a stereoscopic plotter utilizing aerial photographs for plotting points on a map and the computer, has taken much of the drudgery out of the work of highway design by permitting the automatic plotting of highway cross sections and the automatic computation of earthwork volumes. When the exact horizontal and vertical alignment has been established for the route selected, a set of plans is prepared showing the alignment details and the elements to be constructed. The required right-of-way is then purchased using location maps supplemented by ground surveys to establish precise ownership boundaries.

Soil analyses

Earthwork. In order to design the pavement it is next necessary to establish the characteristics of the subgrade soils over which the pavement is to be constructed. Certain soil features can be determined by expert study of the aerial photographs. Detailed soil information must, however, be determined on the ground and in the laboratory. Ground subsurface explorations are carried out by means of auger borings in which the soil strata are identified, classified, and samples obtained for laboratory analysis. The engineering properties of the soil encountered, including its strength, susceptibility of its strength to moisture increase, and amount of shrinkage and swell with moisture change are established. Soils completely unsuitable in the final roadway section are identified for removal; embankment and cut slopes are established, and the degree of compaction to be achieved in the field determined. The probable strength or capacity of the materials to resist the traffic loads is also determined.

After the construction contract has been let and surveys have established the exact location of the finished roadway in plan and elevation on the ground, with suitable line and elevation indicators, construction of the subgrade begins. The first step is to remove all vegetation from the roadway section, an operation in which the bulldozer plays a large part. Heavy earth-moving machinery then moves materials from cut sections into fill sections, where the material is placed in layers, brought to the proper moisture content, and compacted to the required density. In 1970, the unit cost of earthwork, including excavation, hauling, placement and compaction, was not substantially greater in Europe, the United States, and Japan than it had been a generation earlier, in spite of huge increases in other construction costs. Higher capacity equipment operating at greater speeds has prevented increases in costs per unit. It is generally considered desirable to perform the earthwork operations. including installation of the drainage pipes and culverts, as the first element of construction for most highways, followed next by the bridges, and finally the pavements. If it is possible to permit some traffic to operate over the completed earthwork, it is desirable to do so in order to detect weak spots in the grade prior to placing the finished pavement.

Drainage. Highway drainage includes those elements of the roadway that remove surface water (rain) from the roadway and carry flowing streams across the roadway as well as those used to control ground water. It has often been emphasized that adequate drainage is the most important element in road and highway construction; John McAdam was one of the first to grasp the fact. The major drainage is carried in the streams that continuously or intermittently cross the highway route. Surface water from the roadway, as well as from the surrounding lands, is carried in these streams.

"Design storm"

The highway designer must estimate the amount of water that will be carried in the stream at the highway location for the "design storm," which is either the most severe flood expected in a hundred years for a major stream, or merely that expected once in five years for a minor drainage channel on a low traffic-volume route. The elevation of bridges and the size of other drainage structures are fixed to carry this design storm without flooding the roadway. Hydraulic considerations are taken into account in the design of bridges and culverts, and construction may include work to increase the carrying capacity of the stream channel adjacent to the highway through channel changes, bank lining, and similar methods. In areas where land use is changing rapidly, particularly from agricultural to residential or business, attention must be given to the fact that ground water runoff and stream flow will materially increase as the area is covered with houses, drives, streets, business buildings, and parking lots.

The drainage of the roadway itself is an important consideration. Surface water drainage is insured by the crown in the pavement and the slope of the shoulders. Modern designs provide for surfacing the shoulders so that a complete waterproof surface is provided. It is important that all elements of the cross section, subgrade, subbase and base, as well as the surface, be crowned so that surface water entering the pavement system can drain outward to the ditch. Flat side slopes in the ditch section have two advantages. First they provide a better opportunity for vehicles leaving the roadway to recover without a serious accident occurring, and second, the ditch is further removed from the roadway laterally so that there is less opportunity for water in the ditch to penetrate and soften the subgrade. Both rounded V-bottom and trapezoidal ditch sections are commonly used. They must be sized to carry the maximum quantities at water elevations well below the elevation of the roadway shoulder.

In urban areas and at necessary locations in rural highways, particularly at intersections, the drainage of the highway pavement is accomplished by carrying the water laterally to shallow gutters at the edge of the shoulder, then along the gutter to storm-sewer inlets at frequent intervals. The water passes through these inlets into a pipe drainage system that carries it to a natural water course for discharge. For multilane freeways in areas of high rainfall the storm-drain system required is extensive and represents a substantial portion of the project cost.

Capillary water held in the pavement by surface tension is not subject to drainage. Where such water rises into the pavement and comes into contact with an overlying impervious layer, condensation occurs and may produce moisture problems. The use of impervious surfacing layers of appreciable thickness and free draining base and subbase courses minimizes this. If the ground water table is quite high relative to the elevation of the top of the pavement, capillary water problems can be severe; it may be essential to lower the level of the water table under the highway cross section by placing perforated drainage pipes, surrounded by free draining filters of gravel and sand, well below the water table. Ground water flows into the pipes, thus lowering the ground water table and reducing capillary moisture at the pavement elevation. The same procedure can be used to intercept ground water flowing under the highway cross section in pervious strata.

Drainage problems also occur when ground water can drain down the back slopes of cut sections and the fill slopes of higher fills. If these slopes are steep, severe erosion may occur. The usual solution is to flatten slopes so as to permit control through vegetstive cover. When this type of control is not feasible, the surface water is collected in shallow ditches or gutters at the top of the slope and carried along the roadway to a point at which it may be discharged into the ditch system or, alternately, carried down the slope in pipe drains to a suitable discharge point.

Design and construction of pavement elements. In order to design the elements of the pavement cross section the strength of the subgrade material must be determined in its condition of lowest probable strength or highest probable saturation. Strength is determined either by estimates based on experience with similar materials using routinely determined soil characteristics as an aid to judgment, or by means of laboratory test measurements. Materials tested are normally permitted to absorb water by capillarity prior to testing.

Designing roads for future needs

Where construction is to be in stages, the pavement section may be designed for the traffic to be carried during the early years, at the end of which added pavement material will be provided to increase the strength to that required for the next design period. Materials used for subbase and base courses must normally be located near the construction site to avoid high hauling costs. In some cases it may be advantageous to improve the strength of local materials by stabilizing with lime, cement, or bituminous materials. The subbase may also be constructed by stabilizing the top 6 to 12 inches (15–30 centimetres) of the subgrade. Operations involved in constructing the subbase and base courses consist of locating the materials to be used, processing these materials at the pit or quarry as necessary, hauling to the roadway, depositing in the proper quantities, mixing to provide uniformity, grading to proper elevation, adding water where necessary and compacting to the proper density.

Though, for many years concretz pavements were placed directly on the subgrade, it is now recognized that the destructive ejection of water through joints and cracks and along the pavement edge, can be avoided by providing granular bases or subbases or by providing stabilized bases under these rigid pavements. Most portland cement concrete pavements are now constructed over base courses stabilized with cement or bituminous materials. The thickness of the concrete pavement is determined an the basis of the strength of the concrete and the stresses induced by the heavier loads expected to use the pavement.

NOTABLE REBUILDING ACHIEVEMENTS 1920–1945

The parkway concept, forerunner of modern high-volume, high-speed, limited-access highways, was proposed first by William Niles White of New York as a part of the Bronx River protection program of New York City and Westchester County. The 15-mile (24-kilometre), four-lane drive known as the Bronx River Parkway was completed in 1925. Protected on both sides by broad bands of park land that limited access, the highway was located and designed so as to cause minimum disturbance to the landscape, and its use was restricted to passenger cars. The success of the concept led to the creation of the Westchester County parkway system and the Long Island State Park Commission. More parkways and expressways were built in the New York area, including the Merritt Parkway (1934–40), which continued the Westchester Parkway System across Connecticut as a toll road providing divided roadways and limited access.

The Italian autostrada system was started under Mussolini in the 1920s, beginning with an expressway from Milan to Varese. A national road board (Azienda Autonoma Statale della Strada) was given responsibility for the construction, maintenance, and repair of the state roads with advisory supervision over provincial and local roads. This body was assigned the funds from motor-vehicle taxes, and an annual matching grant of approximately the same amount from general tax funds. Superhighways were built primarily as toll roads under government supervision. The first of the major autostrada crossed northern Italy from Venice to Turin. The autostrada were built generally as undivided three-lane roadways with shoulders. All highway and railway grades were separated, access was limited, and there were restrictions on use of the highways by commercial vehicles.

The Inter-American (Pan American) Highway was conceived at the Fifth Conference of American States in 1923 and formalized by cooyerative agreements between the nations involved in 1928. Its route joined the highway system of the United States at Laredo, Texas, and went due south to Mexico City; thence southeast to Guatemala City, San Salvador, Managua, San José, and Panama City, a distance of 3,356 miles (about 5,400 kilometres). The U.S. Congress appropriated $1,000,000 for the highway in June, 1934, and construction began in 1935. Work on the highway in Mexico progressed rapidly but lagged on most of the remainder of the route until World War II, when a large U.S. appropriation permitted construction of a usable pioneer trail for the entire route.

The autobahn

The first fully modem highway system was the German autobahn network consisting of dual roadways separated by a substantial median area and providing for limitation of access. The idea of the motorway (expressway) was first conceived in Germany in 1926 and incorporated in the Cologne-Bonn roadway started in 1929 and opened to traffic in 1932. When Hitler came to power he implemented a plan for the construction of about 7,000 kilometres (4,350 miles) of an integrated highway network known as the "Reich Motor Roads". Construction began in 1934 on the Frankfurt-Mannheim-Heidelburg Autobahn. The entire system included three north–south routes and three east–west routes. The highway provided separate dual-lane roadways with a median strip of five metres, and one-metre shoulders. Clearly military in intent, the roads were designed for large traffic volumes and speeds in excess of 100 miles per hour, bypassing cities, and providing limited access. The whole system, of about 2,500 miles, was rushed to completion in a few years.

The Pennsylvania Turnpike Commission, established in 1937 to raise funds and build a toll road across the Appalachian Mountains, found an unusually favourable situation in the form of an abandoned railroad right-of-way, with many tunnels and excellent grades over much of the route. The turnpike provided a divided dual-lane highway with no cross traffic at grade and with complete control of access and egress at 11 traffic interchanges. Its alignment and grades were designed for high volumes of high-speed traffic and its pavement to accommodate the heaviest trucks. The favourable public reaction to this new type of highway provided the impetus for the post-World War II toll road boom in the United States, advanced the start of a major interstate highway program, and influenced highway developments elsewhere. The Pennsylvania Turnpike, originally running from Harrisburg to Pittsburgh, was later extended 100 miles (160 kilometres) east to Philadelphia and 67 miles (107 kilometres) west to the Ohio border, making it 327 miles (523 kilometres). An original feature of the turnpike, later widely copied, was the provision of restaurant and fuelling facilities.

The Alaska Highway was formalized in 1930 by a joint agreement between Canada and the United States. No funds were appropriated until the beginning of World War II when Alaska became an area of primary strategic importance, and an all-weather road link was needed to Fairbanks. In 1942 a pioneer road was cut from the Dawson Creek railhead northwest of Edmonton, Alberta, to Fairbanks. When complete, the motor-vehicle road was 20 to 24 feet wide and over 1,500 miles long. The road has been improved and has remained in continuous use since its pioneer days.

III. National highway and expressway systems

HISTORY

The Romans realized that a coordinated system of road ways connecting the major areas of their empire would be of prime significance for both commercial and military purposes. In the modern era the nations of Europe first introduced the concept of highway systems. In France, the State Department of Road and Bridges was organized in 1716 and by the middle of the 18th century the country was covered by an extensive network of roads built and maintained primarily by the national government. In 1797 the road system was divided into three classes of descending importance: (1) roads leading from Paris to the frontiers; (2) roads leading from frontier to frontier but not passing through Paris; (3) roads connecting towns. In the early 1920s this general plan remained essentially the same except that a gradual change in class and responsibility had taken place. At that time the road system was divided into four classes: (1) National highways (routes national), improved and maintained by the national government; (2) regional highways (routes *départementales*), improved and maintained by the Department under a road service bureau appointed by the Department Commission; (3) main local roads (*chemins* des *grandes* communications and chemins *d'intérêt commun*) connecting smaller cities and villages, built and maintained from funds of the communes supplemented by grants from the Department; and (4) township roads (chemins vicinaux ordinaires), built and maintained by the communities alone.

While the British recognized the necessity for national support of highways and a national system as early as 1878, the Ministry of Transport Act of 1919 first classified the roadway system and provided for 23,230 miles (37,168 kilometres) of Class I roads and 14,737 miles (23,579 km) of Class II roads with 50 percent of the cost of Class I roads and 25 percent of Class II roads to be borne by the national government. The need for a national through-traffic system was recognized in the middle 1930s and the Trunk Roads Act of 1939 followed by the Trunk Roads Act of 1944 created a system of roadways for through traffic. The Special Roads Act of 1949 authorized existing or new roads to be classified as "motorways" that could be reserved for special classes of traffic. The Highways Act of 1959 swept away all previous highway legislation in England and Wales and replaced it with a comprehensive set of new laws.

The U.S. road system

In the United States, New Jersey in 1891 enacted a law providing for state aid to the counties and established procedures for raising money at township and county level for road building. In 1893 Massachusetts established a state highway commission. By 1913 most of the states had adopted similar legislation but there was little coordination among the states. The Federal Aid Road Act of 1916 established federal aid for highways as a national policy, implemented by an appropriation of $5,000,000. The Bureau of Public Roads, established in the Department of Agriculture in 1893 to make "inquiries with regard to road management," was given responsibility for the program; and an apportionment formula based on area, population, and mileage of post roads in each state was adopted. Funds were allocated for construction costs up to $10,000 per mile and the states were required to bear all maintenance costs. The location and character of roads to be improved was left to the states, an arrangement that had some shortcomings. A national Good Roads Movement that had developed in the later years of the 19th century had long lobbied for a system of national roads joining the major population centres. This point of view was recognized by the Federal Aid Highway Act of 1921, which required each state to designate a system of state highways not to exceed seven percent of the total highway mileage in each state, and federal-aid funding was limited to this federal aid system. The system was divided into interstate roads, not to exceed 3/7 of the total mileage with the balance to be intercounty highway. Bureau of Public Roads approval of the system was lequired, and federal aid was limited to 50 percent of the estimated cost. The first map of the federal-aid system of 168,881 miles (270,210 kilometres) was published in 1923. The Federal Aid Highway Act and the Highway Revenue Act of 1956 provided funding for an accelerated program of construction on the Interstate System. A federal gasoline tax was established, the funds from which, with other highway-user payments, were placed in a Highway Trust Fund. The federal-state ratio for funding construction of the Interstate System was changed to 90 percent federal and ten percent state. It was expected that the system would be completed no later than 1971, but cost increases extended this time to about 1974–75. The system connects nearly all of the major cities in the United States and when completed will carry 20 percent of the nation's traffic on slightly more than one percent of the total road and street system.

An important element in the United States highway system, overlapping the Interstate network, is toll road mileage, most of which was built in the years immediately following World War II. A total of 3,500 miles (5,600 kilometres) of toll road has been constructed in the United States, most of it in the 1950s.

Canada and China are other nations that have formulated national highway policies. The Canadian Highway Act of 1919 provided for a system of 25,000 miles (40,000 kilometres) of highways and provided for a federal allotment for construction not to exceed 40 percent of the cost. The Trans-Canada Highway jointly financed by the federal government and the provinces has made good progress since World War II. China passed "The Regulation of Highway Improvement in China," establishing a National Highway Commission and a system of highways and village highways in 1920, but little has been done to implement the ambitious plan.

The Soviet road system

The Soviet Union has a 1,000,000-mile rural highway system that consists primarily of two-lane roadways. The cities have adequate street systems with very wide paved sections. Most highways radiate from the major cities and while some intercity routes are planned most are formed when two city systems intersect. In order to overcome this problem the Soviet Union is now planning and building several intercity highways, some four-lane divided and some with controlled accesses. It is estimated that 30% of the rural highway system was paved as the 1970 decade began. Funding is provided by an annual tax on cars and trucks, 2% of annual income from industrial plants and collective farms, and some state appropriations. In the future the profits from transportation enterprise are expected to provide much of the cost for new roads. The relatively small automobile population of the Soviet Union in proportion to its large land area generates little demand for highway improvements. As motor vehicle numbers increase strong pressures for more road building will undoubtedly develop.

Japan has a national expressway system of 2,480 miles (4,696 kilometres) most of which is on the island of Honshu. Three major toll roads, comprising about 400 miles, are major links in the system. The typical toll road section consists of two or three lanes in each direction separated by a 15-foot (4.5-metre) median. Because of land shortage in Tokyo and other major cities, urban expressway facilities are frequently double decked over existing streets or over rivers. In 1965 about one-fifth of Japan's national system was paved but good progress has been made since that time. The 117-mile (190-kilometre) Meishin toll expressway, Kōbe to Nagoya, was opened in 1969. Typical of expressway construction in Japanese cities is the 933-million-dollar construction program for the city of Osaka planned and partially completed to support Expo 70. The Osaka system includes a 35-mile (56-kilometre) ring highway, costing 135 million dollars, with 3 lanes in each direction and a 105-foot (32-metre) median. Almost one-third of the Osaka system is over water.

ADMINISTRATION AND FINANCING OF NATIONAL HIGHWAY SYSTEMS

Early road building was administered and financed on a local basis. The advent of the automobile created the necessity for integrated highway systems. Local roads and local streets are still generally administered by cities, townships, and counties or similar units of local government.

The major highway and expressway systems of a nation must, of necessity, have a national administrative base to guarantee continuity of routes and reasonable uniformity in design and construction. This responsibility may be shared with local units of government. In the United States current national highway policies are established by the Federal Highway Administration, an agency of the Department of Transportation. The local point of view is represented by advisory bodies of national policies and practices. notably the American Association of State Highway Officials. in England and Wales the Ministry of Transport, a cabinet-level department of national government, has authority over the motorways. the highest classification of highways, and the truck roads, the next highest, although this latter authority is often delegated to county or municipal authorities. In Scotland the secretary of state is the principal highway authority.

Financing of national roads

The financing of national 'highway systems is a national obligation. Before the automobile age, funds for highway systems were raised by general taxation in the form of labour taxes, by the issuance of bonds guaranteed by general tax revenues and to a limited extent by tolis charged against the users of highways and bridges. Since about 1920 the financing of highways has been almost entirely transferred to the highway user. A broad variety of taxes is employed, with motor-fuel taxes providing the largest single source of revenue. Vehicle licensing is common and trucks are usually licensed on a weight basis. Taxes on tires, rubber used in tires, lubricating oil, and on other motor-vehicle equipment items are widely applied for highway purposes. Excise and sales taxes on new car purchases more commonly accrue to general tax revenues. In periods of urgent demand for highways, for example, after World War II, governments employ credit financing for highway improvements, but the bonds are generally financed by highway-user tax revenues so the real source of funds does not change. Toll roads also become popular in periods of high demand, particularly over heavily travelled routes. The toll system permits rapid construction of a segment of the highway system through bond financing supported by tolls charged to the highway user. The toll, then, is in reality a special highway-user tax that assigns the tax to the particular highway. The toll system has proved effective in funding high-volume elements of national expressway systems.

SYSTEM PLANNING

Overall system study and planning. The planning of national expressway systems is an orderly, continuous process of assessing highway needs, determining the scope and requirements for a system, evaluating the required financing, and finally dealing with the complex relationships that occur among the various governmental units concerned with the system. The plan must provide for future extensions of the system, maintenance, and necessary rebuilding. System planning today is influenced by two major factors: the popularity of the automobile, which has provided new levels of mobility for many of the world's people; and the mass movement of people from rural areas to cities and from central cities to suburbs in the past quarter century, which has put great strains on urban expressway systems.

The need for transportation of both people and goods is closely associated with the physical location of most of the people. Consequently the major rural routes are easy to establish since they must join the major centres of population. The volume of traffic carried over such intercity routes is generally a function of the sizes of the terminal cities and the distances between them. Traffic volumes on intercity routes are established on the basis of current volumes and projections based on increasing population and vehicle ownership. Such volumes are customarily expressed in terms of the average daily traffic. For design purposes it is also essential that the traffic be classified by type with the heavy vehicles assigned to weight classification groups.

Determination of urban traffic volume

In urban areas the determination of traffic volumes assignable to various elements of the expressway systems is much more difficult. The automobile travel of individuals in urban areas is quite complex. Surveys of the origin and destination of present traffic are used to determine and project travel demands. In this procedure the travel of a percentage of the urban households and businesses is carefully studied with regard to trips made and their points of origin and destination. Future traffic is estimated on the basis of added demands as the city grows and on increasing use of the automobile. The result is a picture of the major travel-desire lines in the urban area and an estimate of the average daily traffic on each. The system must be designed to accommodate this traffic. While traffic-study techniques are constantly being refined and improved, the experience of the past 20 years is that urban traffic demands are chronically underestimated. In urban areas consideration must also be given to the parking of automobiles at their destinations. This problem involves both the quantity of parking area required and the methods for accepting and discharging the vehicles.

Rural expressways, urban expressways, and motorways are the most modern and expensive elements in the highway system of a country. Consideration must be given to the effect of this system on the lesser street and highway systems. Traffic must move to and from the expressways and motorways on these lesser systems; thus the proper design and operation of the whole requires careful attention to the needs of each of the system elements and to the proper connections between these elements. When the design and construction of a highway system element is the responsibility of several governmental agencies it is difficult to obtain proper coordination and proper consideration of the entire problem of traffic movement on a system basis. Only by the most effective use of all levels of highway facilities can the motoring public be served well. The result of a systems study for a nation or an urban area should be a network of roadways of varying complexity and traffic-service capability that will adequately serve the needs of the country or the urban area. For example, the Ministry of Transport plan for Great Britain's principal national routes calls for 720 miles (1,159 kilometres) of motorway (expressway) and 1,500 miles (2,400 kilometres) of other national routes. In September, 1950, most of the European countries adopted a system of major highways, totalling about 26,000 miles, (41,600 kilometres), to be improved and known as the "E" system. Completion of the system prior to the turn of the century is unlikely. Only Germany has completed a substantial portion of the routes to expressway standards. Many of the world's urban areas have comprehensive plans for highway development in the area. All of these plans must be continuously reviewed to accommodate changes in traffic needs.

Design engineering and testing. The highway designer starts with information on traffic volumes; types of traffic; vehicle, axle, and wheel loadings; and maximum speeds anticipated. His task is to design the highway cross section and to fix its alignment both horizontally and vertically. He has the further obligation of producing a design that can later be modified or expanded to meet increasing traffic demands. The design process is simplified and uniformity in the system is insured by the use of "standards" that are established for highways of various classifications. These standards are fixed by the highway building agencies. Most agencies with direct responsibility for the construction and maintenance of roads, streets, or highways also have standards for design, construction, and materials. Loadings to be used in the design of expressway systems are based primarily on actual loading measurements of vehicles using the facility. While vehicle weights and axle loads are usually prescribed by law, these legal weight limits are frequently exceeded, hence design loadings are based on actual wheel and axle load measurements and reasonable estimates of trends.

Because they carry large volumes of heavy traffic, expressways justify more sophisticated studies of the probable loading during the life of the pavement. Materials to be used in their construction, as well as for other highway systems, are specified and tested for conformity in accordance with standard specification requirements and test procedures. In addition, most major agencies constructing highways have special specifications and testing methods that apply to their special conditions. The multiplicity of specifications for materials used in highway construction causes problems for the materials-supplying industries and increases construction costs. A given aggregate supplier may furnish the national government, two states or provinces, and a number of local jurisdictions all having different gradation specifications for essentially similar materials. This condition forces the supplier to make and store several materials that have only insignificant differences in size.

Materials selection. A most important element in the design of the highway system is the selection of materials to be used in its various segments. Major factors considered in materials selection are the initial cost in place, estimated annual maintenance costs, service life, and suitability for the driving public. The latter factor is an important consideration for expressways, particularly in urban areas. The high volumes of traffic that use these expressways on a nearly continuous basis create difficulty in maintenance and renovation and for the drivers using the expressway during periods of repair or change. For this reason materials and methods of construction for expressways are selected to minimize maintenance and renovation; *i.e.*, the highest types of pavement produced with high quality materials.

Limited access and special requirements. A prominent feature of an expressway system is the basic premise that the expressway has the function of moving through traffic safely at reasonable speeds and with the maximum feasible limitation of points of access to and egress from the system. For expressways the land service function is relegated to a very minor role. All points at which traffic enters or leaves the traffic stream are points of traffic turbulence and particular attention must be given to designing these entrances and exits to facilitate traffic movement. Vehicles entering the expressway must ihcrease speed to that of expressway traffic and enter a gap in the traffic stream; this is accomplished by added lanes, called acceleration lanes, at points of entry. Similarly, deceleration lanes are provided at exit points.

The use of grade separation

Where two major expressways intersect, traffic flow is maintained uninterrupted by grade separation of the through traffic and the provision of separate lanes for each of the traffic movements from one expressway to the other, of which there are eight in all. The resulting pattern is commonly called a cloverleaf. Intersections of this type in urban areas often involve very complex problems, because of the unavailability of land, requiring compact intersections, and the need to accommodate traffic movements between the expressway and the street system. In such cases it is usually necessary to provide grade separations for some of the turning movements; three or four level grade separations may be provided by elevated bridge structures.

Urban expressways are also faced with the problem of clear separation of expressway traffic from that of the street system. This is accomplished either by depressing the expressway below ground and carrying the streets across the expressway on bridge structures or by elevating the expressway above the street system. The depressed expressway is generally better from the point of view of appearance but poses difficulty in maintenance of the back slopes, handling of storm water drainage, and construction in areas of high water table. The elevated expressway is expensive and often considered aesthetically objectionable, but its maintenance cost is lower and in areas of high water table or where large numbers of heavily travelled streets must be crossed it offers the most feasible solution to the traffic-separation problem. Expressway exits and access points in urban areas present difficult design problems for both the depressed and elevated freeways but are generally somewhat more difficult for the elevated system.

Expressways in urban areas require extensive lighting because of the movement of large volumes of traffic during the hours of darkness and the necessity for illumination to provide proper vision for the entrance, exit, and route-change movements. In addition to continuous lighting of the through lanes, special lighting and sign lighting is provided to accommodate entering and departing vehicles.

IMPACT OF THE NEW EXPRESSWAYS

Benefits. The new expressways and motorways have been very popular with the travelling public. Carrying large volumes of traffic at high speeds, they have excellent safety records due to the directional separation of traffic, absence of intersections, minimum interference from entering and leaving vehicles, more uniform traffic speed, and excellent visibility. Full control of access, as compared to no control, other conditions being equal, is responsible for an approximate 60 percent reduction in accident rates and 45 percent reduction in fatalities based on travel mileage. The greatest reduction occurs in suburban areas. Urban and rural expressways have only one-fifth the accident rate experienced on city streets and rural highways.

There is also a substantial saving in time and operating costs for all types of vehicles on the expressway in comparison to operating costs on normal rural highways or city streets. Studies made in Los Angeles in the early 1960s indicated operating and accident cost savings of 3% cents per mile for passenger cars, 10 cents per mile for trucks. Expressways, furthermore, provide much better driving conditions for the motoring public, so that trips are completed with less physical wear and tear. This has led to increased recreational and cultural travel.

Controversies and problems associated with expressways. The expressway, because of its greater traffic capacity produces problems of air pollution, noise, and, in the eyes of some, visual pollution. Air pollution due to vehicle operation has become a substantial problem in the larger urban areas of the world, particularly those in which common atmospheric conditions prevent rapid dispersion of the pollutants. Though this is not alone an expressway problem, the heavy concentration of vehicles on expressways makes them a major source of pollutants. The answer to the problem presumably lies in new technical developments that either will drastically reduce the volume of pollutants emanating from the internal-combustion engine or will facilitate the substitution of alternate, nonpolluting power sources for automobiles. Another possibility being explored in the early 1970s in Rome, New York City, and elsewhere in the restriction of vehicle use in the central city.

Urban expressways

The urban expressway and to a lesser extent the rural expressway are considered by many to be poor neighbours. A major intersection of urban expressways consumes tens of acres of land and the construction of thousands of linear feet of bridge structure. Such an intersection in a built-up area means a substantial disturbance of rhe neighborhood. The same is true to a lesser extent along the entire route in built-up areas. Strong objections are often raised to expressway locations in urban areas; public opposition has sometimes halted construction. Conservationists have objected to the location of expressways that utilize lands now in parks, golf courses, and other green areas, or which run adjacent to recreation areas such as beaches. The destruction of historic buildings and landmarks by expressway construction has drawn much criticism in Europe and the United States.

In rural areas the problems of expressway location are much less severe because of the availability of alternate locations and the smaller numbers of businesses and residences involved. Rural expressway intersections are generally designed to provide for turning movements at grade, thus eliminating the need for long bridge structures, although requiring large land areas. There have nevertheless been protests, not only from conservationists but from rural property owners whose lands and homes were affected. It has been alleged that highway designers give insufficient attention to aesthetics and create ugly scars on the landscape. Expressway designers of the 1970s are sensitive to these criticisms and are devoting much effort to design, location, and landscaping that fits the roadway to the countryside. In the United States the Highway Beautification Act of 1965 authorized the partial withholding of federal funds from states that do not take proper action to build and maintain aesthetically pleasing highways. Funds have been authorized for the control of billboards and junkyards adjacent to expressways. The appearance of expressways can be materially improved by good planning and design to fit the roadway to the terrain, the retention of native trees and shrubs and landscaping by judicious planting. Wide median areas of natural vegetation or landscaping are popular. The wider rights-of-way required are considered to be justified by the improvement in driving conditions as well as appearance.

Rivalry among highway users

Another area of controversy, principally in the United States, is the rivalry that exists among different elements of the transportation industry, particularly between the trucking and railroad industries. Unquestionably the expressways, which permit truck operations at high speeds and with few stops, substantially reduce the cost of operating these vehicles. In addition the trucking industry has argued for increases in allowable gross vehicle weights and dimensions, increases that are opposed by highway administrators and designers on the grounds that existing expressways are not designed to accommodate these heavier vehicles. Railroad interests and some highway administrators contend that the commercial trucks and buses do not pay their fair portion of the highway-user taxes and are, therefore, being subsidized by tax and highway-user funds from the general public. The question has never been settled, but it is generally conceded that higher taxes on commercial vehicles approximately cover the incremental costs involved in constructing the expressway. Highway cost allocation remains a thorny fiscal and philosophical problem that has occupied highway economists for many years; the approaches presented do not have the approval of all interests involved.

IV. Operation and maintenance

The glamorous phase of highway engineering is that of system planning, design, and construction of new highway facilities. Service to the driving public, however, is very much dependent upon operation and maintenance activities. The life of a highway is a function of the quality of maintenance and minor renovation, and long life provides the most important single element in highway economics.

The operation of traffic on roads and highways is subject to four types of control: (1) legal control, or the laws setting forth the rules of the road; (2) roadway signs and markings that provide instructions and information; (3) traffic light signals; (4) police action. Expressways are specifically designed to avoid the necessity for traffic signals, except on access or exit lanes, and police traffic control except in the case of traffic congestion due to accidents.

Maintenance consists of activities concerned with the condition of the pavement and shoulders, including surface conditions, structural integrity and adequacy, and the condition of the right-of-way areas outside the travelled way. It is also concerned with snow removal, debris removal, and the installation and care of pavement markings, signs, and signals.

OPERATION

Markings, signs, and signals. The marking of roadway surfaces with painted lines or types of permanent markers is standard practice throughout the world. While there are some disadvantages of pavement surface marking, notably high maintenance costs and problems in night visibility, it is generally accepted that the advantages far exceed the disadvantages. Interior lane lines and centre lines are marked with broken lines, either yellow or white, and dangerous conditions such as restricted sight distance and pavement edges are indicated with solid lines.

Signs are used to advise the driver of special regulations that apply at specific times and places and to provide information with regard to routes, directions, destinations, hazards, and points of interest. Expressway sign planning is particularly important because the driver who makes a mistake and misses an entrance or exit cannot recover quickly because of the limited access characteristic, which means that the next entrance or exit may be miles away. Signs are classified as: (1) regulatory signs, which provide notice of traffic laws and regulations such as speed-limit, stop and yield signs, and signs regulating traffic movement; (2) warning signs, which call attention to conditions in or adjacent to the roadway representing a potential traffic hazard such as turns, steep grades, low vertical clearance, or slippery pavement surface; (3) guide signs, which show route designations, destinations, distances, points of interest, and other similar information. In most nations signs have standard shapes and colours, with one shape used for the stop sign, another for warning signs, etc. Expressway directional signs, commonly mounted over the roadway on a sign bridge, are large in size for easy reading at high speeds and often have white letters and symbols on a green background. Special shapes and colours are used for route markers. A United Nations commission studied highway signs in 1951 and made recommendations for standard sign procedures that have been adopted by many nations. The commission found that the use of symbols is preferable to words; the advantage of wordless signs in Europe, with its large international traffic, is evident. The commission also found that three-colour signs are quite visible. European danger signs traditionally were triangular, but the commission recommended adoptions of the American diamond shape on the basis of better legibility and comprehension. Conferences in Bangkok (1967), Montevideo (1967), and Vienna (1968) were devoted to adoption of a convention for road signs and signals acceptable to most nations of the world.

Types of signs

The traffic signal has its primary use in traffic control in city street systems, but it also has a place on rural highways. At highway grade intersections accommodating large volumes of traffic the traffic signal is used to allocate the right-of-way to the various traffic streams. Systems known as traffic-actuated signals automatically monitor the demand in the traffic streams and allocate the green time or right-of-way correspondingly. Preference can be given to particular traffic directions by such signals. A recent development in expressway operation is the traffic-signal system to meter traffic entering on access lanes. These signals provide a red indication to entering traffic until a gap sufficient for entry occurs in the exterior expressway lane at which time the signal gives a green indication and the entering vehicle moves into the traffic stream. Signals are also used on expressways to indicate lanes open and closed ahead of the driver.

Rules of the road and speed limits. Legal rules governing the movement of traffic are an essential part of orderly movement on the highway. Such regulations may be nationwide, state- or province-wide, or local. In general the rules for operation of vehicles on the highway may be divided into three main categories. First are the rules applying to the vehicle and the driver such as vehicle and driver registration, vehicle equipment, accident reporting, and financial liability. Second are the general rules for drivers and pedestrians such as speed limits, right-of-way, and turn requirements known as the rules of the road. Third are those regulations which apply to limited roadway sections such as speed zones, one-way operations, and turn controls.

The important rules of the road are reasonably uniform throughout the world, except in two important aspects. First, the right-of-way is allocated to vehicles on the right-hand side of the highway in most nations, but on the left-hand side in a few, notably the U.K. This influences vehicle design since it is important that the driver be on the side of the vehicle adjacent to opposing traffic; *i.e.*, on the left side of the vehicle for the right-hand rule. The second major variation is in the regulation of vehicle speed by the use of speed limits. Speed limits on open rural highways and motor ways are not specified in many European countries. Police control judgments are made as to whether or not a driver's speed is compatible with driving conditions. In the United States and Canada speed limits are widely used. The limits set vary from 30 miles (48 kilometres) per hour for local service highways in built-up areas to 55–70 miles (88–112 kilometres) per hour on rural highways and expressways. Higher speed limits of 75–80 miles (120–128 kilometres) per hour are used in a few jurisdictions and on toll highways. An important aspect of the speed question is traffic speed differential, which should be minimal. There is much less interference in a nearly uniform traffic stream and consequently greater safety. For this reason minimum speed limits are established on many expressways, particularly in urban areas, with the lower limits set at 10 to 20 miles per hour below the upper limits.

Special regulations are important to the efficient movement of traffic in specific segments of the street and highway system. One-way street systems in congested urban areas provide safer driving conditions and increase the traffic-carrying capacity of the system. The prohibition of turns at intersections contributes to safety and reduces conflicts. Such regulations, however, may adversely affect some businesses. Speed zoning is used to establish localized speed limits in special zones. Reduced speed limits are commonly used on highways approaching built-up areas and on dangerous highway sections where lower speed limits are justified. Higher than normal speed limits may also be established on particularly safe sections of highway.

Highway policing. Highway patrols or highway police were inaugurated to help in solving the highway accident problem by enforcing driving regulations on the major highway systems. In urban areas traffic patrol and enforcement of traffic regulations is a responsibility of the police department. Most large cities have a traffic division in the police department. On heavily travelled expressways in rural areas adequate patrol requires one patrolman for each 4 to 5 miles (6.4–8 kilometres). In addition to patrolling to insure compliance with speed and other regulations, the patrolmen also investigate accidents, render assistance to disabled vehicles, and attempt to apprehend criminals.

Speed control methods

An important aspect of traffic regulations and accident prevention is the control of excessive speed. Speed is commonly measured by radar devices or by pacing with a patrol car. Speed traps that involve travel time measurements over a fixed distance are ordinarily not permitted. In accident investigation speed is determined by skid marks. Another important factor in highway accidents is the driver who is under the influence of alcohol or drugs. Tests for intoxication are now widely used. The determination of the alcohol content in the blood or other bodily fluids is probably the most conclusive indication of excessive use of alcohol, but such tests are not practical for police identification of excessive drinking. The most widely used test for police identification of the state of intoxication from alcohol is the breath test in which the driver blows up a balloon and his breath is run through a series of chemical contacts in an analyzer. The results indicate to the detaining officer whether or not the driver's condition is due to alcohol and the approximate blood alcohol content. The alcohol blood concentration that is presumptive of poor driving ability is not firmly established. Maximum levels of 0.15 percent have been widely used to prove alcoholic influence, but many authorities believe that a limit of 0.05 percent is more realistic and that 0.10 percent is the maximum reasonable upper limit.

In addition to the traffic control and policing function, highway patrols regulate traffic at the scene of accidents, provide information and are in this and other ways very helpful to the driving public. The primary work of various expressway emergency patrols is to assist in emergencies and to carry on activities to insure efficient movement of high volumes of traffic. New developments in recent years include the use of light airplanes and helicopters for patrol purposes. They are particularly effective in locating trouble spots causing traffic jams and relaying this information to patrol officers on the ground for corrective action.

MAINTENANCE

The proper and adequate maintenance of the highway after construction is one of the most important elements for proper functioning of the roads and highway system. Proper maintenance keeps the roadway safe, provides good driving conditions, and prolongs the life of the pavement, thus protecting the highway investment. Maintenance operations are carried on in all governmental jurisdictions by crews of men organized, trained, and equipped to carry out the maintenance function. Maintenance activities can be grouped into three major areas: (1) maintenance of pavements and shoulders; (2) maintenance of ditches, slopes. right-of-way areas, and drainage structures; and (3) maintenance of signs and markings and operations aiding traffic movement.

The three major maintenance activities

Maintenance of pavements and shoulders involves operations of the same type and involves use of much of the same equipment as for new construction. Maintenance of the ditches, slopes and right-of-way areas involves mowing operations to control vegetation and work to control and eliminate soil erosion. A major maintenance expenditure is involved in picking up and removing trash thrown or dumped in the ditch and right-of-way areas by the travelling public. Roadside containers for such trash provide only a partial solution to the problem. In the more rigorous winter climates substantial maintenance expenditures are required to remove snow and ice from the pavement including snow plowing, the spreading of salt for snow and ice removal, and spreading sand to provide traction.

V. Future highway trends

One of the major questions for the future concerns the desirability of continued construction of expressways in urban areas. In many high-population-density urban areas the transportation problem can probably be solved more economically and with lesser requirement for land by the use of mass transit, both bus and fixed rail. The operation of a Bay Area Rapid Transit System, in the San Francisco Bay Area, will provide an indication of the potential success of well-designed mass transit in attracting riders and reducing automobile traffic. At the same time it seems reasonable to predict that the automobile and truck will have a large place in the world's transportation for many years. In most parts of the world automobile use is steadily increasing. In urban areas continued development of more adequate expressway facilities in limited land area requires new approaches. One possibility is to take the expressway system underground in tunnels. The primary difficulties in this type of construction are the high cost of tunnelling and the ventilation problem. Another possibility is to move the expressway system to the second story level with direct access to buildings and parking garages at this level and the allocation of the street level to pedestrians and a few local service vehicles. If the automobile and bus are to continue to be the primary source of transportation of the people in urban areas, new approaches will be needed to integrate expressway systems into the total facilities of the central city.

Highway safety will continue to demand attention, much of which will be given to the design of roads and highways. More four-lane divided highways will be built to replace existing two- and three-lane roads. Recent developments of easy-breaking signs and light poles, along with impact attenuators to protect the vehicle that strikes rigid fixed objects such as piers and expressway gores, have reduced the severity of accidents involving vehicles striking fixed objects. Eetter roadway surfacing, alignments, and signing and marking will improve driving conditions and make highways safer.

A major safety feature under consideration is the development of an electronic highway in which vehicles entering the highway will be locked into a guidance system that will control its speed and path of travel. To be acceptable such a system will have to be absolutely reliable and provide a quality of travel that will insure the driver's willingness to give up his individual driving freedom.

Truck highways

Another potential development is the truck highway. There have been many examples of roads and highways on which trucks were prohibited, in particular the parkways around New York. Increases in transportation of goods by truck in the developed countries have led to truck traffic congestion on many highways. In some areas serious consideration will probably be given to the planning and construction of highways to be limited to truck traffic. Such highways would be financed by the trucks using them, thus eliminating the criticism that commercial vehicles are subsidized by automobile user taxes on the public systems.

One of the major question marks in the highway-building future is the outlook for the internal-combustion engine. Widespread problems of air pollution in major urban areas have led to the suggestion that the engine be legally regulated nut of use or that very rigid controls on emissions be established. If this trend continues and new types of engines or other devices to power the automobile and truck are adopted. the changes may require substantial changes in the design of highway systems. As an example major changes in acceleration rates would have a serious effect on the design of the highway system for passing and merging operations. Changes in other vehicle operating characteristics will have similar impact for other elements of the system.

BIBLIOGRAPHY

Major studies: T.R. AGG and J. E. BRINDLEY, *Highway Administration and Finance* (1927), a comprehensive discussion of highway administration, state highway organizations, highway finance, and the development of highway systems in the U.S. during the period 1800–1925; E. DAVIES (ed.), *Roads and Their Traffic* (1960), a discussion of major traffic arteries primarily those in urban areas from the British point of view; R.J. FORBEs, *Notes on the History of Ancient Roads and Their Construction* (1934), a history based on archaeological explorations of early road building in Europe, Asia Minor, and India from about 3500 B.C. to the Fall of Rome; A.C. ROSE, *Public Roads of the Past,* 2 vol. (1952–53), a world history of road building from the dawn of recorded history to the beginning of the development of the U.S. Interstate system in the late 1940s; HERMANN SCHREIBER, *The History of Roads: Frotn Amber Route to Motorway* (Eng. trans. 1961), a history of the ancient roads of Asia, Europe, Egypt, and the Inca roads in South America; L.J. RITTER and R.J. PAQUETTE, *Highway Engineering,* 3rd ed. (1967), a general text on highway engineering covering practice in the United States up to 1967: WILBUR SMITH AND ASSOCIATES, *Future Highways and Urban Growth* (1967), a comprehensive study of the National System of Interstate and Defense Highways of the United States as it relates to future travel requirements and the changing shape of urban areas; K.B. WOODS·(ed.), *Highway Engineering Handbook* (1960), a handbook covering all elements of engineering as it applies to the design, construction, operation, maintenance, financing, and administration of highway systems in the United States. See also the publications of the NATIONAL ACADEMY OF SCIENCE, NATIONAL RESEARCH COUNCIL, HIGHWAY RESEARCH BOARD: *Highway Research Record,* formerly *Highway Research Bulletins* (irreg.); *Highway Research Abstracts* (monthly); and *Highway Research Review* and *Highway Research News* (both irreg.).

Other references: T. AITKEN, *Road Making and Maintenance*, 2nd ed., pp. 1–27 (1907), a general discussion of the construction and maintenance of roads from the British point of view at the beginning of the 20th century; S.G. CHAPMAN (ed.), *Police Patrol Readings* (1964), a treatise on police patrol operations in the U.S. and England; R.J. FORBES, *Bitumen and Petroleum in Antiquity*, pp. 69–75 (1936), a discussion of the sources of natural bitumen from natural springs and bitumen impregnated limestones in the ancient world and its use in building: J.W. GREGORY, *The Storv of the Road,* 2nd rev. ed. (1938), a history of road and highway development, with emphasis on British roads, ancient and modern, prehistoric trade routes, Roman roads, the roads of China, and

the Inca roads of Peru; L.I. HEWES, *American Highway Practice,* 2 vol. (1942), a comprehensive portrayal of American highway practice just prior to World War II; J.L. MCADAM, *Remarks on the Present System of Road Making* (1821), a discussion of the condition of English roads in the early years of the 19th century; STUART PIGGOTT, "The Beginning of Wheeled Transport," *Scient. Am.,* 219:82–90 (1968), a discussion of the earliest development of wheeled transport based on archaeological studies in Mesopotamia, Georgia, and Armenia in Southern Russia and in Europe; W. BREWSTER SNOW (ed.), *The Highway and the Landscape* (1959), a discussion of the design of highways to best fit the landscape and methods for their beautification. See also the report of the U.S. BUREAU OF PUBLIC ROADS, *Highway Statistics* (annual); and *Traffic Quarterly,* a journal devoted to traffic and traffic control.

(F.J.B.)

Robert I the Bruce, of Scotland

Robert "the Bruce," the 14th-century champion of Scottish independence from England, became king of Scots in 1306 and led the forces that freed Scotland from English suzerainty in 1328. Among the legends that later became attached to his name was the story that, when outlawed and hard pressed by the English, with his fortunes at low ebb, he derived hope and patience from watching a spider perseveringly weave its web.

The Anglo-Norman family of Bruce, which had come to Scotland in the early 12th century, was related by marriage to the Scottish royal family, and hence the sixth Robert de Bruce (died 1295), grandfather of the future king, claimed the throne when it was left vacant in 1290. The English king Edward I claimed feudal superiority over the Scots, and awarded the crown to John de Balliol instead.

The eighth Robert de Bruce was born on July 11, 1274. His father, the seventh Robert de Bruce (died 1304), resigned the title of earl of Carrick in his favour in 1292; but little else is known of his career until 1306. In the confused period of rebellions against English rule from 1295 to 1304 he appears at one time among the supporters of the rebel leader William Wallace, but he was later apparently restored to Edward I's confidence. There is nothing at this period to suggest that he was later to be the Scottish leader in a war of independence against Edward's attempt to govern Scotland directly.

The decisive event was the murder of John ("the Red") Comyn in the Franciscan church at Dumfries on February 10, 1306, either by Bruce or his followers. Comyn, a nephew of John de Balliol, was a possible rival for the crown, and Bruce's actions suggest that he had already decided to seize the throne. He hastened to Scone and was crowned on March 25.

Coronation on March 25, 1306

The new king's position was very difficult. Edward I, whose garrisons held many of the important castles in Scotland, regarded him as a traitor and made every effort to crush a movement that he treated as a rebellion. King Robert was twice defeated in 1306, at Methven, near Perth, on June 19, and at Dalry, near Tyndrum, Perthshire, on August 11. His wife and many of his supporters were captured, and three of his brothers executed. The King himself became a fugitive, hiding on the remote island of Rathlin off the north Irish coast.

In February 1307 he returned to Ayrshire. His main supporter at first was his only surviving brother, Edward, but in the next few years he attracted a number of others. The King himself defeated John Comyn, earl of Buchan, and in 1313 captured Perth, which had been in the hands of an English garrison; but much of the fighting was done by his supporters, who progressively conquered Galloway, Douglasdale, the forest of Selkirk and most of the eastern borders, and finally, in 1314, Edinburgh. During these years the King was helped by the support of some of the leading Scottish churchmen and also by the death of Edward I in 1307 and the ineptness of his successor, Edward II. The test came in 1314 when a large English army attempted to relieve the garrison of Stirling. Its defeat at Bannockburn on June 24 marked the triumph of Robert I.

Victory at Bannockburn

Almost the whole of the rest of his reign had passed before he forced the English government to recognize his position. Berwick was captured in 1318, and there were repeated raids into the North of England, which inflicted great damage. Eventually, after the deposition of Edward II (1327), Edward III's regency government decided to make peace by the Treaty of Northampton (1328) on terms that included the recognition of Robert I's title as king of Scots and the abandonment of all English claims to overlordship.

The King's main energies in the years after 1314, however, were devoted to settling the affairs of his kingdom. Until the birth of the future king David II in 1324 he had no male heir, and two statutes, in 1315 and 1318, were concerned with the succession. In addition, a parliament in 1314 decreed that any who remained in the allegiance of the English should forfeit their lands; this decree provided the means to reward supporters, and there are many charters regranting the lands so forfeited. Sometimes these grants proved dangerous, for the King's chief supporters became enormously powerful. James Douglas, knighted at Bannockburn, acquired important lands in the counties of Selkirk and Roxburgh that became the nucleus of the later power of the Douglas family on the borders. Robert I also had to restart the processes of royal government, for administration had been more or less in abeyance since 1296. By the end of the reign the system of exchequer audits was again functioning, and to this period belongs the earliest surviving roll of the register of the great seal.

In the last years of his life, Robert I suffered from ill health and spent most of this time at Cardross, Dumbartonshire, where he died on June 7, 1329, possibly of leprosy. His body was buried in Dunfermline Abbey, but the heart was removed on his instructions and taken by Sir James Douglas on a pilgrimage to the Holy Land. Douglas was killed on the way (1330), but, according to one tradition of uncertain value, the heart was recovered and brought back to Melrose Abbey. In later times Robert I came to be revered as one of the heroes of Scottish national sentiment and legend.

BIBLIOGRAPHY. The authoritative biography is G.W.S. BARROW, *Robert Bruce and the Community of the Realm of Scotland* (1965). This gives references to the sources and a full bibliography and supersedes all earlier works. The most important original authority for the life of Robert I is *The Bruce,* a poem in Scots by JOHN BARBOUR, probably completed in 1376; the most recent edition is by W.M. MACKENZIE (1909).

(B.We.)

Robert Guiscard

A commander and statesman of exceptional ability, Robert de Hauteville, better known as Robert Guiscard (the Astute), was the remarkable leader of the Norman invasions of southern Italy during the last two-thirds of the 11th century.

Robert was born in Normandy around 1015, into a family of knights. Arriving in Apulia, in southern Italy, around 1047 to join his half brother Drogo, he found that it and Campania, though they were southern Italy's most flourishing regions, were plagued by political disturbances. These regions attracted hordes of fortune-seeking Norman immigrants, who were to transform the political role of both regions in the following decades.

Arrival in Apulia

In Campania, the Lombards of Capua were launching wars against the Byzantine dukes of Naples in order to gain possession of that important seaport. In Apulia, William (Iron Arm) de Hauteville, Robert's eldest half brother, having successfully defeated the Byzantine Greeks who controlled that region, had been elected count of Apulia in 1042. In 1046 he had been succeeded by his brother Drogo.

When Robert joined his brothers, they sent him to Calabria to attack Byzantine territory. He began his campaign by pillaging the countryside and ransoming its people. In 1053, at the head of the combined forces of Normans from Apulia and Campania, he defeated the haphazardly led forces of the Byzantines, the Lombards, and the Pope at Civitate. Guiscard now consolidated the riches of the Hautevilles. Because of the deaths of William and Drogo, and of his third half brother, Count Humphrey, in 1057, Robert returned to Apulia to seize control from Humphrey's sons and save the region from disgregating internal conflicts. After becoming the recognized leader of the Apulian Normans, Robert resumed his campaign in Calabria. His brother Roger's arrival from Normandy enabled him to extend and solidify his conquests in Apulia.

In his progression from gang leader to commander of mercenary troops to conqueror, Robert emerged as a shrewd and perspicacious political figure. In 1059 he entered into a concordat at Melfi with Pope Nicholas II. Until that time the papacy had been hostile toward the Normans, considering them an anarchist force that upset the political structure in southern Italy—a structure based on a balance of power between the Byzantines and the Lombards of northern Italy. The schism between the Greek and Latin churches in 1054 worsened the relations between the Byzantine emperors and the papacy, and eventually the papacy realized that Norman conquests over the Byzantines could work to their advantage. Guiscard's plan to expel the Arabs from Sicily and restore Christianity to the island also found favour in Nicholas' eyes. This expedition into Sicily, accompanied by an ever increasing religious fervour, got under way in 1060, as soon as the conquest of Calabria was completed. Guiscard entrusted the command of the expedition to his brother Roger; but on particularly difficult occasions—*e.g.*, the siege of Palermo in 1071—he came to his brother's aid.

Until this time, Robert's relations with Roger had not always been amicable, since Roger, aware of both his own talent and Robert's dependency on him, would not settle for the subordinate role allotted him. Their differences were resolved when Robert invested Roger, after he had recognized Guiscard's supreme authority, with "the County of Sicily and Calabria" along with the right to govern and tax both counties.

Expansion of the duchy

Robert continued to expand the small county left by Humphrey into a duchy, extending from the Adriatic to the Tyrrhenian seas. The capture of Bari in April 1071 resulted in the end of Byzantine rule in southern Italy. Guiscard turned next to the neighbouring territories of Salerno, controlled by the Lombards. Instead of fighting them, he dissolved his first marriage and in 1058 married the sister of Salerno's last Lombard prince, Gisulf II. Hostilities broke out between the two rulers, however, and Gisulf naively tried to bring about a Byzantine counteroffensive against Guiscard. Fearing that the Norman advances into Campania, Molise, and Abruzzi would threaten the papal dominions, Pope Gregory VII excommunicated Guiscard and gave Gisulf considerable military aid. The struggle came to a head when Gisulf, determined to display his power, advanced toward the prosperous city of Amalfi. Guiscard responded to the city's plea for help in 1073 and successfully defended it; in December 1076 he took Salerno from Gisulf and made it the capital of his duchy.

Robert was now at the height of his power. During his rise he repressed with an iron hand not only the claims of Humphrey's sons but also the uprisings of towns and lords fretting under the restraints imposed upon them. The harshness with which Robert dealt with these rebels was intended to mold a heterogeneous population into a strong, sovereign state.

When, in 1080, the conflict between church and state over the right to make ecclesiastical investitures had become more intense, Robert chose to reconcile himself with Gregory VII, entering into the Concordat of Ceprano, which confirmed the commitments of the earlier Council of Melfi, Even the Byzantine court drew closer to him and went as far as trying to establish a familial relationship with Guiscard. The Byzantine emperor Michael VII, in need of Robert's help to uphold his unstable throne, married his son, Constantine, to one of Guiscard's daughters, Helen. The opposition party, however, deposed Michael and confined Helen in a monastery. To guarantee Apulia against attack from the new rulers of Byzantium, Guiscard wanted the territories on the Adriatic coast of the Balkan Peninsula, and he began to build a large navy. Michael's ouster and Helen's confinement reawakened his unappeased spirit of adventure and hastened his long-considered expedition. Now his goal was even more ambitious: to march to Byzantium and crown himself emperor in place of the deposed Michael.

In 1083 Guiscard landed in Epirus with a well-trained army and immediately succeeded in defeating the Byzantines and their Venetian allies. The Pope, however, suddenly recalled him to Italy to help him oust the German king Henry IV, who was marching on Rome en route to claiming southern Italy for the Holy Roman Empire. Having returned home and suppressed the revolts of the lords hostile to himself and to Pope Gregory VII, Guiscard moved toward Rome, defeated the Pope's enemies, and escorted him to Salerno in the summer of 1084. Following this success, he returned to his campaign on the Adriatic coast. He died during the siege of Cephalonia on July 17, 1085.

Achievements

Physically attractive, endowed with an acute and unscrupulous intelligence, a brilliant strategist and competent statesman, Guiscard had begun to organize a state composed of diverse ethnic and civil groups: Latin and Germanic in Lombard territories and Greek in Byzantine domains. The new political structure was built on a monarchial–feudal framework characteristic of the time, but it was controlled by the energetic and uncompromising Guiscard, who tried to use his ducal power to create a powerful and prosperous state. The other base on which he built was Roman Catholicism, the religion of the conquerors and most of the conquered, which he used to reconcile the subjected peoples. An extremely religious man, Guiscard was distrustful of the Greek clergy because of their ties with Byzantium. On the other hand, his generosity toward the Latin Church was bountiful. He endowed it with territories and clerical immunities in order to tie it firmly to the feudal system. Splendid cathedrals and Benedictine abbeys were built in the hope that they would consolidate and diffuse Latin language and culture among the heterogeneous people and tie them into a new, unified state. Guiscard was kept from realizing this political vision only by his death.

BIBLIOGRAPHY. Additional information on Guiscard may be found in the following sources (all with extensive bibliographies): GUILLAUME DE POUILLE, *La geste de Robert Guiscard,* ed. by M. MATHIEU (1961); F. CHALANDON'S classic work, *Histoire de la Domination Normande en Ztalie et en Sicile*, 2 vol. (1907); and E. PONTIERI, *Tra i Normanni nell'Italia meridionale,* 2nd ed. (1964).

(E.Po.)

Robespierre

Maximilien-François-Marie-Isidore de Robespierre, who was known to contemporaries as "the Incorruptible," played a leading part in the French Revolution, particularly the period of the Jacobin republic of 1793–94. A democrat, he thought, like Rousseau, that moral virtue was inseparable from the exercise of sovereignty. The strictness with which he adhered to his principles won him the approval of the French people, but eventually his realism led him to the belief that to save the country and the republic the government had to impose its will by means of coercion and "the Terror."

Early life. Maximilien was born at Arras on May 6, 1758, the son of a lawyer. After his mother's death, his father left home, and Maximilien, his brother, and his sisters were brought up by their maternal grandparents.

From 1765 Maximilien attended the college of the Oratorians at Arras, and in 1769 he was awarded a scholarship to the famous college of Louis-le-Grand in Paris. A brilliant student, he obtained his bachelor's degree in 1780 and, after taking his law degree the following year, became a lawyer at Arras, where he set up house with his sister Charlotte. He soon made a name for himself and was appointed a judge at the Salle Épiscopale, a court with jurisdiction over the provostship of the diocese. His private practice provided him with a comfortable income.

Lawyer in Arras

He was admitted to the Arras Academy for the advancement of the arts and sciences in 1783 and soon became its

R b pi e it by a n artist.
Ir the Musée t, ri
J.E. Bulloz

chancellor and later its president. Contrary to the long-held belief that Robespierre led an isolated and withdrawn life, he often visited local notables and mingled with the young people of the district. He entered academic competitions, and his "Mémoire sur les peines infamantes" ("Report on Degrading Punishments") won first prize at the Academy of Metz. He belonged to a literary society in Arras and wrote elegies in the fashion of the time. Though he was rumoured to be engaged to his cousin Antoinette Deshorties, he never married.

By 1788 Robespierre was already well-known for his altruism. As a lawyer representing poor people, he had alarmed the privileged classes by his protests in his "Mémoire pour le Sieur Dupond" against royal absolutism and arbitrary justice. When the summoning of the States General (a national assembly that had not been called since 1614) was announced, he issued an appeal entitled "To the people of Artois on the necessity of reforming the Estates of Artois." In March 1789 the citizens of Arras chose him as one of their representatives, and the Third Estate (the commons) of the bailiwick elected him fifth of the eight deputies from Artois. Thus he began his political career at the age of 30.

Membership in the National Assembly and Leadership of the Jacobins. Robespierre preserved his frugal way of life, his careful dress and grooming, and his simple manners both at Versailles and later in Paris. Despite his youth he quickly attracted attention in an assembly that included some distinguished names. He probably made his maiden speech on May 18, 1789, and was to speak more than 500 times during the life of the National Assembly. He succeeded in making himself heard despite the weak carrying power of his voice and the opposition he aroused, and his motions were usually applauded. Proofs of his growing popularity were the ferocious attacks made by the royalist press on this "Demosthenes," "who believes everything he says," this "monkey of Mirabeau's" (the Comte de Mirabeau, a politician who wanted to create a constitutional assembly). Robespierre was kept out of the committees and from the presidency of the National Assembly; 'only once, in June 1790, was he elected secretary of the National Assembly. In April he had presided over the Jacobins, a political club promoting the ideas of the French Revolution, of which he had been a member since it was first organized. In October he was appointed a judge of the Versailles district tribunal.

Robespierre, nevertheless, decided to devote himself fully to his work in the National Assembly, where the constitution was being drawn up. Grounded in ancient history and the works of the French philosophers of the Enlightenment, he welcomed the Declaration of the Rights of Man and of the Citizen, which formed the preamble of the French constitution of September 3, 1791, and insisted on observing it. He fought for universal suffrage, for unrestricted admission to the National Guard, to public offices, and to the commissioned ranks of the army, and for the right to petition. He opposed the royal veto, the abuses of ministerial power, and religious and racial discrimination. He defended actors, Jews, and black slaves and supported the reunion of Avigncn, formerly a papal possession, with France in September 1791. In May he had successfully proposed that all new deputies be elected to the next legislature so that as a completely new body, it would better express the people's sovereign will.

His passionate fight for liberty won him more enemies, who called him a dangerous individual—and worse: robber, murderer, spy, dictator. After the flight of Louis XVI (June 20–21, 1791), for which Robespierre vainly demanded his trial, the slanders of the revolutionary deputy became twice as violent. He hastened the vote on the constitution so as to attract "as many of the democratic party as possible," inviting in his ***Adresse aux Français*** (July 1791) the patriots to join forces. Martial law was proclaimed, and at the Champ-de-Mars (July 17) the National Guard, under the Marquis de Lafayette, a moderate who wanted to save the monarchy (and who had played an important role in the American Revolution), opened fire on a group demanding the abdication of the King. Robespierre, his life threatened, went to live with the family of the cabinetmaker Maurice Duplay. He managed to keep the Jacobin Club alive after all of its moderate members had joined a rival club. When the National Assembly dissolved itself, the people of Paris organized a triumphal procession for Robespierre, an honour repeated in Artois during a short stay there (October 1791).

The
Jacobin
Club

Although he had excluded himself and all his colleagues from the new Legislative Assembly, Robespierre continued to be politically active, giving up the lucrative post of public prosecutor of Paris, to which he had been elected in June 1791. Henceforth, he spoke only at the Jacobin Club, where he was to be heard about 100 times until August 1792. There he opposed the European war that his fellow revolutionary Jacques-Pierre Brissot was advocating as a means of spreading the aims of the revolution. He denounced the secret intrigues of the court and of the royalists, their collusion with Austria, the unpreparedness of the army, and the possible treason of aristocratic officers whose dismissal he demanded in February 1792. He also defended patriotic soldiers, such as those of the Châteauvieux regiment, who had been imprisoned after their mutiny at Nancy. When Brissot's supporters stirred up opinion against him, Robespierre founded, in May, a newspaper, ***Le De'fenseur de la Constitution,*** which strengthened his hand. He violently attacked Lafayette, who had become the commander of the French Army, whom he suspected of wanting to set up a military dictatorship, but failed to obtain his dismissal and arrest.

The reverses suffered by the French Army after France had declared war on Austria and Prussia had been foreseen by Robespierre, and, when invasion threatened, the people rallied to him. Although he had defined the aims of insurrection, he hesitated to advocate it: "Fight the common enemy," he told the *fédérés* (provincial volunteers) assembled in Paris, "only with the sword of law." When the insurrection nevertheless broke out on August 10, 1792, Robespierre took no part in the attack on the Tuileries Palace. But that same afternoon his ***section*** (an administrative subdivision of Paris), Les Piques, nominated him to the insurrectional Commune. As a member of the electoral assembly of Paris he heard the news of the September Massacres of imprisoned nobles and clergy by Parisian crowds. He exonerated the mob, and on September 5 the people of Paris elected him to head the delegation to the National Convention.

Work in the National Convention. The Girondins—the Revolutionary group that favoured political but not social democracy and that controlled the government and the civil service—accused Robespierre of dictatorship from the first sessions of the National Convention, but the Legislative Assembly did not concur. At the King's trial, which began in December 1792, Robespierre spoke 11 times and called for death. His speech on December 3 rallied the hesitant. With the help of of his new journal, ***Les Lettres à ses commettants*** ("Letters to His Constituents"), he kept the provinces informed. The King's execution did not, however, resolve the struggle between the Girondins and the Montagnards—the deputies of the extreme left. At the same time the scarcity of food and rising prices created a revolutionary mood. The treason of Gen. Charles Dumouriez, who went over to the Austrians, precipitated the crisis. A kind of "popular front" was formed between the Parisian sans-culottes, the poor, ultraleft republicans, and the Montagnards. On May 26, 1793, Robespierre called on the people "to rise in insurrection." Five days later he supported a decree of the National Convention indicting the Girondin leaders and Dumouriez's accomplices. On June 2 the decree was passed against 29 of them.

The democratic constitution of June 1793 contained several articles drafted by Robespierre. But his proposals to limit property rights, to proclaim the right to work for all, to establish a graduated tax and a league of nations seemed too bold. This constitution was never applied.

Role in the Committee of Public Safety and the Reign of Terror. After the fall of the Girondins, the Montagnards were left to deal with the country's desperate position. Threatened from within by the movement for federalism and by the civil war in the Vendee in the northwest, and threatened at the frontiers by the anti-French coalition, the revolution mobilized all its resources for victory. In his diary, Robespierre noted that what was needed was "une volonté une" ("one single will"), and this dictatorial power was to characterize the revolutionary government. Its essential organs had already been created, and now he set himself to the task of making them work.

On July 27, 1793, Robespierre took his place on the Committee of Public Safety, which had first been set up in April. While some of his colleagues were away on missions and others were preoccupied with special assignments, he strove to prevent division among the revolutionaries by relying on the Jacobin societies and the vigilance committees. Henceforward his actions were to be inseparable from those of the government as a whole. As president of the Jacobin Club and then of the National Convention, he denounced the schemes of the Parisian radicals known as the Enragés, who were using the food shortage to stir up the Paris ***sections.*** Robespierre answered the demonstrators on September 5 by promising maximum prices for all foodstuffs and a revolutionary militia for use in the interior against counterrevolutionaries and grain hoarders.

The Reign of Terror

In order to bring about a mass conscription, economic dictatorship, and total war, he asked to intensify the Reign of Terror. "It is our leniency towards traitors that is ruining us," he declared, and asserted the need for a "swift, harsh, and inflexible justice" that should be administered in accordance with laws against suspects. But he objected to pointless executions, protecting those deputies who had protested the arrest of the Girondins and of the King's sister. He was sickened by the massacres condoned by the ***représentants en mission*** (members of the National Convention sent to break the opposition in the provinces) and demanded their recall for "dishonoring the Revolution."

Robespierre devoted his report of 5 Nivôse, year I (December 25, 1793 [the French Republican calendar had been introduced in September 1793]), to justifying the collective dictatorship of the National Convention, administrative centralization, and the purging of local authorities. He protested against the various factions that threatened the government. The Hébertists, the Cordeliers, and the popular militants all called for more radical measures and encouraged de-Christianization and the prosecution of food hoarders. Their excesses frightened the peasants, who could not have been pleased by the decrees of 8 and 13 Ventôse, year II (February 26 and March 3, 1794), which provided for the distribution among the poor of the property of suspects. Reappearing at the Jacobin Club after a month's illness, Robespierre denounced the radical Revolutionist Jacques-René HBbert and his adherents, who together with some foreign agents were executed in March. Those who wanted, like Georges Danton, to halt the Reign of Terror and the war attacked the policies of the Committee of Public Safety with increasing violence. Robespierre, although still hesitant, led the National Convention against these so-called Indulgents. The Dantonist leaders and the deputies who were compromised in the liquidation of the French East India Company were guillotined in April.

A deist in the style of Rousseau, Robespierre disapproved of the anti-Christian movement and the "masquerades" of the cult of reason, In a report to the National Convention in May, he affirmed the existence of God and the immortality of the soul and strove to rally the revolutionaries around a civic religion and the cult of the Supreme Being. That he remained extremely popular is shown by the public ovations he received after Henri Admirat's unsuccessful attempt on his life on May 22. The National Convention elected him president, on June 4, by a vote of 216 out of 220. In this capacity he led the festival of the Supreme Being in the Tuileries Gardens on June 8, which was to provide his enemies with another weapon against him.

Declining Influence and authority

After the law of 22 Prairial (June 10) reorganizing the Revolutionary Tribunal, which had been formed in March 1793 to condemn all enemies of the regime, opposition to Robespierre grew; it was led by those ***représentants en mission*** whom he had threatened. His influence was challenged in the Committee of Public Safety itself, and the Committee of General Security, which felt slighted by the General Police Bureau directed by Robespierre, Georges Couthon, and Louis de Saint-Just, became even more hostile to him. In the cafes he was accused of being a moderate. And Joseph Cambon, the minister of finance, detested him.

Unremitting work and frequent speeches in the Legislative Assembly and at the Jacobin Club (a total of some 450 since the beginning of the session) had undermined Robespierre's health, and he became irritable and distant. Embittered by the slanders and by the accusations of dictatorship being spread both by the royalists and by his colleagues, the Montagnards, he stayed away from the National Convention and then, after June 28, from the Committee of Public Safety, confining his denunciations of counterrevolutionary intrigues to the Jacobin Club. At the same time he began to lose the support of the people, whose hardships continued despite the recent French victories. From his partial retirement Robespierre followed, dumbfounded, the unleashing of the Great Terror in the summer of 1794 and the progress of opposition.

Attempting to regain his hold on public opinion, Robespierre reappeared at the Committee of Public Safety on 5 Therrnidor (July 23) and then, on July 26, at the National Convention, to which he turned as his judge. His last speech was at first received with applause, then with disquiet, and finally the parliamentary majority turned against him. Despite his successful reception that evening at the Jacobin Club, the next day Robespierre's adversaries succeeded in preventing him from speaking before the Legislative Assembly, which indicted him together with his brother, Augustin, and three of his associates. Robespierre was taken to the Luxembourg prison, but the warden refused to jail him. Later, he went to the City Hall (the Hôtel de Ville), where he could, doubtless, still have continued the struggle, for armed contingents from some of the ***sections*** of the city had been summoned by the Paris Commune and were awaiting his orders. But Robespierre refused to lead an insurrection and eventually his loyal contingents began to disperse. Declared an outlaw by the National Convention, Robespierre severely wounded himself by a pistol shot in the jaw at the City Hall, throwing his friends into confusion. The soldiers of the National Convention attacked the hall and easily seized the wounded Robespierre and his followers. In the evening of 10 Thermidor (July 28), the first 22 of those condemned, including Robespierre, were guillotined amidst a cheering mob on the Place de la Révolution (now the Place de la Concorde). In all, 108 people died for their adherence to Robespierre's ideas.

Assessment. Robespierre's enemies credited him with dictatorial power, both in the Jacobin Club and in the

Committee of Public Safety, a power that he did not have. Counterrevolutionaries and the rich condemned his egalitarian ideas, while popular militants accused him of lacking boldness. After his death, his memory was relentlessly attacked, and a large part of his papers was destroyed. And so history mostly portrayed him as either a bloodthirsty creature or a timid bourgeois. But, following the ascendancy of the popular movements of the 19th century, both in France and abroad, homage was paid to this "persecuted patriot," and his most famous speeches were reprinted. His social ideal consisted in reducing extreme inequalities of wealth, in increasing the number of small property owners, and in ensuring work and education for all. He was a man of his times, of the Enlightenment, a patriot, a man with a sense of duty and of sacrifice, whose influence remains considerable.

BIBLIOGRAPHY. MARC BOULOISEAU, GEORGES LEFEBVRE, and ALBERT SOBOUL (eds.), *Oeuvres complètes*, 10 vol. (1910–67), a critical edition of the complete writings of Robespierre; ERNEST HAMEL, *Histoire de Robespierre, 4* vol. (1865–67), the first general analysis and critical essay, by an author who does not conceal his sympathy for his subject; A.J. PARIS, *La Jeunesse de Robespierre et la Convocation des États généraux en Artois* (1870); ALBERT MATHIEZ, *Études robespierristes,* 2 vol. (1917–18) and *Autour de Robespierre* (1925); *Robespierre terroriste* (1921), collections of articles tending to rehabilitate Robespierre and justify his political actions; J.M. THOMPSON, *Robespierre,* 2 vol. (1935, reprinted 1969), the classic work in English; P.R. ROHDEN, *Robespierre: Die Tragödie des politischen Ideologen* (1935); LOUIS JACOB (ed.), *Robespierre vu par ses conternporains* (1938); JEAN MASSIN, *Robespierre* (1956); *Bi-Centenaire de la naissance de Robespierre (1758–1958*) (1958); MARC BOULOISEAU, *Robespierre* (1957) and *Le Comite' de salut public,* 1793–1795 (1962); GERARD WALTER, *Robespierre,* definitive edition, 2 vol. (1961); W.M. MARKOV (ed.), *Maximilien Robespierre* (1958), a collection of articles in German; *Actes du colloque Robespierre,* XII International Congress of Historical Science, Vienna 1965 (1967); GEORGE RUDE (comp.), *Robespierre* (1967), a selection of texts in English.

(M.Bo.)

Robot Devices

A robot device is an instrumented mechanism used in science or industry to take the place of a human being. It may or may not physically resemble a human or perform its tasks in a human way, and the line separating robot devices from merely automated machinery is not always easy to define. In general, the more sophisticated and individualized the machine is, the more likely it is to be classed as a robot device. Of 20th-century origin, the word "robot" was first used in the play *R.U.R. (Rossum's Universal Robots)* by the Czechoslovak dramatist Karel Capek, who had derived it from a Czech word *robota,* meaning "forced labour." An older word, automaton, has now come to be applied to a mechanical device that imitates a human or an animal, without necessarily performing useful work. The newer word android is restricted to human-like, not animal-like, mechanisms.

HISTORY

Modern robot devices descend through two distinct lines of development—the early automata, essentially mechanical toys, and the successive refinements and innovations introduced in the development of industrial machinery.

Early automata. Hero of Alexandria, a Greek savant of the 1st century, is reputed to have constructed birds that chirped, drank, and flew. The report may be exaggerated; certainly there were no known imitators for many centuries. But 18th-century Europe suddenly inspired an interest in automata among many technically gifted men. Preserved in a Vienna museum is a "writer" (1753), a mechanism capable of writing and drawing. The machinery controlling the hand is housed in a sphere that forms part of the base. Pierre Jacquet-Droz (born 1721) and his son Henri-Louis Jacquet-Droz (born 1752), French clockmakers, produced several mechanical men, which wrote, drew, or played musical instruments. Jacques de Vaucanson, a member of the French Academy of Sciences, is known from photos in the Musée des Arts et Métiers, Paris, to have made a mechanically animated duck that flapped its wings, drank water, pecked corn, and even "digested" or, at least, dissolved the corn. Another famous mechanical animal (in the Victoria and Albert Museum in London) is a tiger made by an unknown inventor for a customer in India. The tiger is in the act of mauling a prostrate European; when a handle is operated the tiger is made to growl by a worm gear and crankshaft connected to an organ inside, while at the same time the victim's arm moves. More recent examples of classic automata are modem animated toys, such as walking and talking dolls.

Industrial robot development. Possibly the earliest ancestor of industrial robot devices is the clepsydra, or water clock, which improved on the hourglass by employing a siphon principle to automatically recycle itself; Ctesibius of Alexandria is said to have made such a clock about 250 BC. Weight-driven, pendulum-controlled clocks were invented in the Middle Ages; the spring-driven clock did not come until the 18th century, which also witnessed the introduction of basic low-level automatic machinery in the textile industry. From the point of view of automation and robots, the most significant development of the Industrial Revolution was the punched-paper-tape-controlled Jacquard loom, an idea whose full exploitation had to await much further progress in technology.

Influence of the Industrial Revolution

The Industrial Revolution stimulated the invention of robotic devices to perfect the production of power itself. The steam engine inspired the governor (actuated by rotating weights), which, when it slowed under load, increased the flow of steam to the engine, and, when its load decreased, reduced it. The internal-combustion engine of the 19th century brought a recycling innovation in the form of pistons that repositioned themselves after each cycle. The later 19th and early 20th centuries saw a rapid proliferation of powered machinery in industrial operations. These at first required a man to position both the work and the machine, later, only the work, and still later, in mid-20th century, required no man at all. Antomatic cycle-repeating machines (automatic washers), self-measuring and adjusting machines (textile colour-blending equipment), and machines with a degree of self-programming (automatic elevators) followed, bringing the prospect of mechanical devices at least theoretically capable of semi-intelligent interactive operation.

The blend of automation development with the progress of automation and cybernetics began to take shape in the 1950s and 1960s. One line of development was represented by the synthetic robot, a computer simulation model of a human being. An example was SAMMIE, created at the University of Nottingham in England and displayed on a television tube. SAMMIE was designed to show how a human would perform in various environments, such as an aircraft cockpit. Another type of research involved the construction of such experimental automata as a mechanical horse built by a laboratory team at the General Electric plant, in Schenectady, N.Y., which could be ridden over rough ground at 30 miles (48 kilometres) per hour; and an android, constructed at Massachusetts Institute of Technology, which could be programmed to walk down a corridor, enter a room, position furniture so as to facilitate retrieval of a book, and return to its starting point.

MODERN ROBOT DEVICES

Only a few robot devices used in industry and elsewhere are android in character. Typically they depend on electronic components to process environmental information and permit prompt decision making and action. A familiar type of robot device that perceives a change, interprets the change in accordance with programmed instructions, and responds appropriately, is the room thermostat. As the room cools, the thermostat perceives a new environmental state, interprets it, and reacts by closing the electrical contacts to start up the heating unit. When the room has reached the programmed temperature, the thermostat again perceives, interprets, and acts.

Decision-making robots. Robots of a higher level are capable of adapting to changes in environment, but are also capable of making decisions by selecting the proper solution from among several alternatives stored in memory units. The traffic-control signal illustrates this higher level of robot activity. It may be time-switch-controlled on a fixed cycle (lowest level), or it may adjust its operation in response to information fed from sensors in the road surface about the number and timing of vehicles.

Automatic aircraft pilot and landing systems

A more complex robot device in modern transportation is the automatic aircraft pilot, which has progressed from the ability to maintain a set course to its present ability to operate as a complete control system capable of performing all routine in-flight steering and control manoeuvres. A complete flight, from takeoff to landing, is now possible without direct human intervention, as long as the aircraft can be brought within the effective range of a ground-based instrument-landing system. Such a system locates the aircraft in relation to the runway during the approach and makes automatic landings possible.

The reliability of aircraft instrument landing systems, the key to their successful application, depends on the duration of time for which the automatic device must operate. Typical figures for a system's probability of failure are 10^{-7} (or one failure in 10,000,000) for 30 seconds' operation in automatic landing, or perhaps 10^{-5} (one in 100,000) for a three-hour supersonic aircraft flight.

Advanced robotic control. Robotic control involves: (1) giving instructions (programming) to the device and storing information (memory) within it; (2) gaining knowledge of the state or position of the device and processing this information to decide a course of action (feedback); and (3) transforming this feedback data into action (response).

A high-level robot can accept and remember instructions. Basically mechanical in operation, they can also be constructed electrically to form electromechanical combinations. Programming or memory devices can also be electromechanical (electrical contacts, relays, completely electronic solid-state memorv devices), or magnetic (magnetic tape, and so forth).

An example of the application of programming and processing by robot-computer devices is seen in the navigation of spacecraft. Such vehicles require a degree of accuracy that far exceeds the capabilities of human coordination. Furthermore, in any event, complete stabilization and accurate control of rockets cannot be achieved by human hands in certain flight situations because the responses required are so rapid. Missiles and space vehicles in flight must, therefore, be controlled primarily by robot devices, although manual intervention on some levels can be interposed. Two types of robot navigators are installed in spacecraft. One, a celestial navigator, calculates position from star patterns. The other, an inertial navigator, receives responses from a gyroscope; any change of direction or velocity is sensed by the opposition of the gyroscope to such changes, and the path of the vehicle can be deduced by an appropriate summation of these reactions. Information from either navigating device is processed by a computer, making possible unmanned space flight and enormously facilitating manned space flight.

Analog and digital control systems

Control systems for advanced robots can be either analog or digital computer systems (see also COMPUTERS). Analog computers compare analogous functions with the variable over which control is required. A common analog device is the automobile speedometer, which relates electromechanical characteristics of the instrument to road speed of the vehicle. Digital control systems depend on counting. Computer digital systems work in binary form (*i.e.*, in pairs), in which information is presented as a pattern of two states or merely the presence or absence of a factor. This two-state system is used to make machine-processing of numerical information as simple as possible. An analog system is intrinsically less precise than a digital system. In the past, analog systems were common because of their low cost. Development of modern electronic components of direct digital monitoring has cut the cost of digital systems and led to their widespread adoption. Digital systems have the added attraction of being able to resolve complex equations more accurately and completely, and any changes (updating of information) can be accommodated more readily.

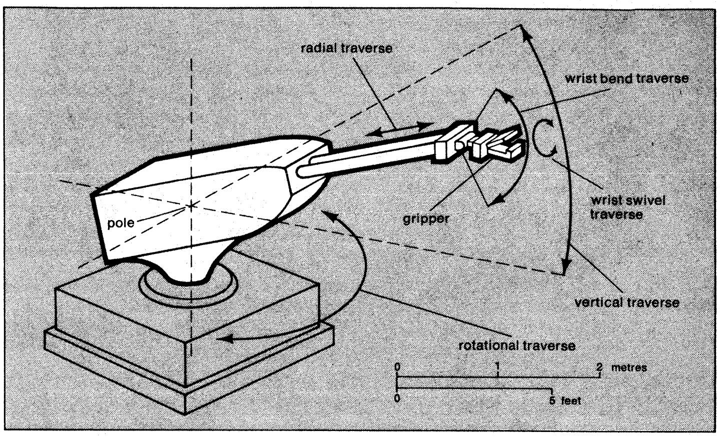


Figure 1: Polar movement of single-arm industrial robot.
By courtesy of Unimation, Inc.

Industrial robots. A human worker, however superb a craftsman he may be, has certain physical limitations. He cannot work continuously in a hostile environment. He works at a relatively slow rate and possesses little physical strength. At best, he can work continuously at peak efficiency only for relatively short periods. Most significantly, from the economic point of view, he is in short supply and often expensive to hire. Modern industrial robot devices aim to substitute a machine for a man in hostile environments; cut costs by replacing expensive hand labour with cheap, dependable machines; and provide versatile all-purpose robots or mechanical handling devices at predictable cost. The alternative is a special device for each operation, with higher costs.

An example of a primary industrial robot is illustrated in Figure 1, which shows a single-arm robot that is polar in operation; that is, the single arm is attached at the centre like the hand of a clock, but is free to move in any direction. Another popular device able to carry out similar functions has a cylindrical movement. Numerous robots, based on these concepts, have been constructed. Figure 2 illustrates the working mechanism of a single-arm robot in polar operation; when the robot is operating, signals from the memory drum are compared with feedback signals to produce a control signal that actuates a special kind of valve. That device, in turn, directs the robot's arm and hand to repeat a specific sequence of movements. The drawing illustrates only one axis of motion. Other existing types of robots involve elbow movements for greater flexibility, but these have not met with much success in industry, doubtless because of the complexity of an elbow joint. Hydraulically operated heavier robots have successfully handled products weighing hundreds of pounds. Smaller robots, operated pneumatically, can handle loads from a fraction of an ounce to several pounds. The "hand" at the end of the operating arm can be a simple mechanical gripping device, or magnetic or suction pads to handle a variety of objects.

Industrial robots have been successful in operating forging presses, loading and unloading machine tools, stack-

By courtesy of Unimation, Inc.

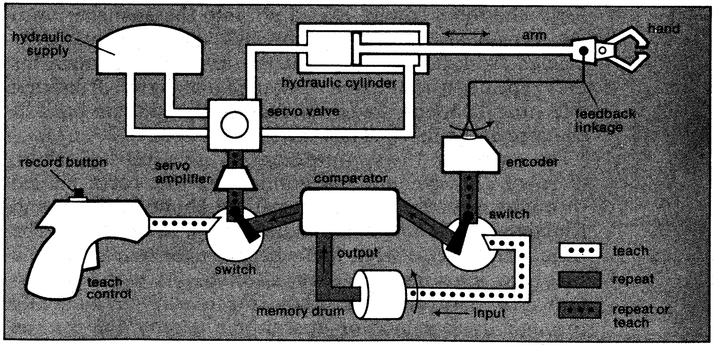


Figure 2: Working mechanism of a single-arm robot in polar operation.

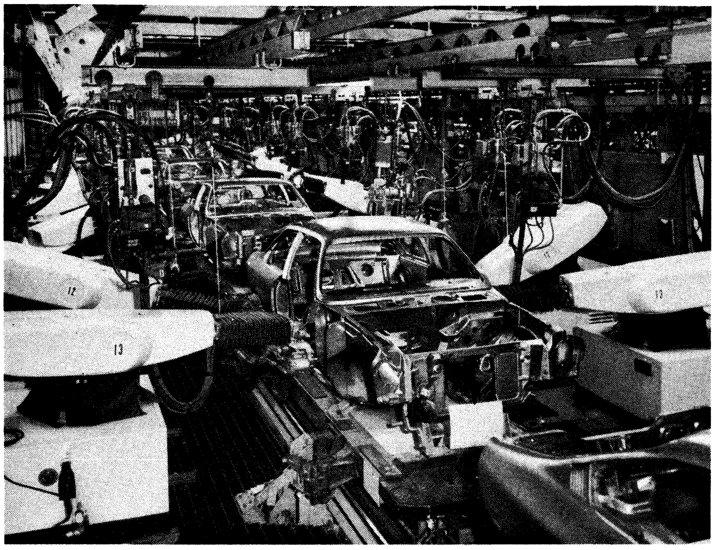


Figure 3: Programmable robots, working on both sides of an automobile assembly line, perform a series of precision welds to within 1/16 of an inch (1.5 mm).
By courtesy of Unimation, Inc.

ing parts into rectangular layers (as in loading kiln cars with bricks), spraying paint, controlling welding of automobile bodies in assembly plants (Figure 3), loading glass tubing into multilayer containers, galvanizing pails by dipping them into a molten zinc bath, unloading large injection-molding machines for making plastics, and heat-treating tractor parts by taking them through a complete furnace-to-quenching cycle. The use of robots in such jobs enabled the removal of men from toxic atmospheres, hot areas, and other hazardous or unpleasant surroundings and the easy performance of exhausting work.

Certain industrial operations that are too complex to be handled by existing robots can nonetheless effectively employ robots to simplify or modify a part of the operations. Robots can be tooled with sets of fingers designed for specific jobs. They can grasp, hold, and release light or heavy parts, and can push, pull, twist, raise, lower, and rotate parts as required. One such robot is capable of five individually controlled degrees of motion—twist, turn, rotate, up-down, and right-left. A sixth motion is used for grasping. This robot handles parts weighing up to 75 pounds (34 kilograms) and positions them within 0.05 inch (1.27 millimetre) of target position. Its memory can store as many as 180 sequential commands.

Robots in forging plants

The ability of robots to handle parts too hot for human handling is especially advantageous in forging plants. In one factory a robot plucks a connecting pin from a magazine feed (so-called because it works like the magazine of a gun), inserts it into a groove in the rotary hearth of a heating furnace, releases it, then grasps a red-hot pin, removes if from the furnace and signals the furnace door to close. The robot then places the hot pin in a quencher. The complete cycle takes about one minute. The robot is electrically prevented from loading pins into the furnace until it receives a signal that the furnace hearth has rotated to accept a pin and the furnace door is open. Similarly, the robot cannot load a pin into the quencher until the previously quenched part has been automatically ejected from it. The fact that the robot can perform its task repeatedly is due to a computer-like digital drum that can store 180 separate commands and send them back to the arm in the form of electrical impulses.

There are two ways to program, or give instructions to, a robot. In dynamic programming, the robot is led through its task while information about position and function are recorded simultaneously in its memory. The robot can then do what it has been taught by playing the recorded information back through its control system. Alternatively, the robot can be driven manually through its sequence of operations; a recording of the control responses generated is subsequently played back to operate the robot continuously.

Thus, to program the memory system described above, a plant worker takes the robot's arm in his own hand and leads it through the desired motions. At every step that calls for one of the robot's five motions, the worker pushes a "record" button to activate the memory drum. After one complete learning cycle the robot is ready to repeat the process endlessly.

Industrial robots such as these are still relatively low-level devices. They carry out their functions in a mechanical manner only after being programmed to do so. Apart from simple detection devices, such as photocells, which can be connected to them to sense the presence of objects and measure temperature, such robots are blind, deaf, dumb, and insensitive. They cannot sort out parts presented at random, and they must receive such parts one at a time in rigidly correct position for handling. Yet, in their own sphere of efficiency, they have proved so successful that applications have been suggested outside industry; for example, in household uses, such as table laying and clearing, washing, and bedmaking. **A** table-clearing robot has been successfully demonstrated by a team of researchers at Queen Mary College, London.

Experimental and scientific robots. In addition to robots with immediate practical applications in industry, scientists in the 1970s were working with a variety of robot devices, some planned for scientific-research applications, others simply with the aim of exploring the potential of the devices themselves. One of the most intriguing aspects of the robots' performance is their apparent potential for learning. A simple robot that illustrates the basic learning method is a mechanical rat that sequentially explores several paths in a maze, in a pattern of successive combinations of left and right turns. Its task is to reach a favoured position at the end of one of the channels. The rat uses a kind of trial and error process, sequentially adopting cyclic patterns of left and right turns. Eventually it ends up in the desired position, at which time a signal is injected, causing the machine to record the correct pattern. Henceforth, the machine travels quickly and accurately along the same track to accomplish any job that has been assigned.

The robot sailboat

Among a number of scientific-research applications for robots that have been proposed, an imaginative one is the robot sailboat developed by the Radio Corporation of America (RCA). An electronically controlled vessel, it is capable of carrying out a wide range of military, scientific, and commercial missions. Dubbed Skamp, for Station-

By courtesy of the Stanford Research Institute

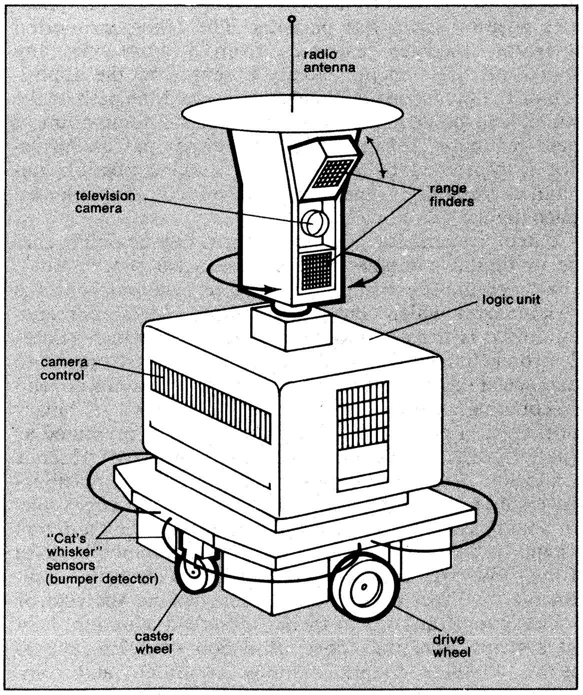


Figure 4: Shaky Robot device built by Stanford Research

Keeping and Mobile Platform, the circular, 8-foot-high plastic boat, which resembles a buoy with sails, theoretically could be sent to any point in the world's seas to radio back information. It is designed to receive orders from navigation satellites and to remain without a mooring within one-fifth of a nautical mile of its assigned position up to one year. Skamp might also serve as a navigation station for ships, submarines, or aircraft.

An example of the advanced android robots produced experimentally in the 1960s and 1970s is shown in Figure 4. Called Shaky Robot by its makers at Stanford Research Institute in California, it has perceptive capabilities and consists of a 3-wheel bogey, or base, with separate motor drives to each of the two drive wheels. Steering is achieved by driving the two main wheels at different rates; the third wheel has a caster action. The device does not carry its computational facilities within it, but communicates with a central computer over two radio channels. The swivelling head carries a television camera eye and range finder, to provide the capability for visual and positional recognition of a number of simple geometric shapes. "Cat's whisker" sensors around the robot operate microswitches, enabling it to take appropriate action if it encounters obstructions. Control circuitry in the cabinet (mounted within the device) processes commands radioed from the computer. Shaky Robot can be programmed to find simple objects, learn room boundaries as viewed by the television eye and range finder, and interpret such information numerically; problems involving object identification and position are solved by the computer through the use of projective geometry.

The humanoid robot

Another robotic device designed in the U.S. in 1969 for use in scientific fields has been called the first humanoid. A joint American design venture of the Atomic Energy Commission and the National Aeronautics and Space Administration, this mechanical man is intended to be guided by a computer for work in extreme or dangerous environments, such as retrieving highly radioactive materials and rescuing people in a contaminated area. The humanoid is equipped with artificial arms, TV camera, and sound pickup. Mounted on a flatbed vehicle, it uses a teletype machine for communication and is steered and controlled by microwave radio. Laser guidance, which would transmit more information in a tighter package, is under study for later installation. Its uses might extend to the testing of rockets during operation and the initial exploration of a planet. Case Western Reserve University in the U.S. has studied the possibility of having humanoids disassemble a nuclear reactor core, which is made up of thousands of individual fuel elements.

Such a robot is intended to improve man's senses, or add new ones, and then to do his bidding; its human features allow the operator to identify with it and thereby operate it more easily. A humanoid can be built to see via television, detect heat through thermal sensors, feel shapes with tactile devices, and transmit these sensations to the operator. Humanoids exceed humans in ultraviolet and infrared sensitivity and in magnifying images. The speed and precision of their responses is also superior to man's. The major problem facing designers of humanoids is that of communication, between the operator and the computer, and between the computer and the humanoid. In each case, human, computer, and humanoid, a different language is spoken (verbal, mathematical, and electrical), which requires translation.

BIBLIOGRAPHY. J. COHEN, ***Human Robots in Myth and Science*** (1966), a readable survey of the history of robots and automata; E. DROZ, "From Jointed Doll to Talking Robot," ***New Scientist,*** 14:37–40 (1962), an interesting paper giving factual details of past automata with information on their performance and mode of operation; A. BELLEAU, "Wiener McLuhan and the Rise of Automata," ***Architectural Design,*** 38:302–304 (1968), a philosophical survey of the principles of automata; L.T. HOGBEN, ***Science for the Citizen,*** 4th rev. ed. (1957), a general book on the development of science containing references to robot devices and automata; ***Proceedings of the 1st National Symposium on Industrial Robot~Çhicago, April 1970*** (1970), reports on the applications of modern industrial robots.

(W.B.H.)

Rock Deformation

Evidence of rock deformation today

The oldest rocks at the Earth's surface reveal abundant evidence of intense deformation, and analysis of the Earth's crust in all parts of the world demonstrates that spasmodic deformation has characterized all of geological time. Rock deformation continues unabated today. If it were not for continuing deformation, erosion would remove all the mountains within a few million years and reduce all land areas to sea level; the sediment produced would be accommodated easily in the ocean basins. Hence, the presence of mountains is strong evidence for geologically recent uplift. Earthquakes also reflect present-day rock deformation because they result from spasmodic, rapid slipping of major crustal blocks along fractures (fault planes).

Despite the importance of erosion by water, wind, and ice, rock deformation can still be considered the most important single factor controlling the environment at the Earth's surface. It results in the uplift of mountains like the Sierra Nevadas of California and the Alps and Carpathians of central Europe; in the growth of island arcs (curving chains of islands such as the Antilles of the Caribbean), and in foundering of rift valleys (fault-bounded troughs such as the East African Rift Valley). On a larger scale, although the evidence is not wholly unequivocal, many scientists believe that continental drift is continuing today. The continental drift hypothesis suggests that the American continents were contiguous with Europe and Africa some 60,000,000 years ago, and then separated slowly with the continents moving across the globe at perhaps one centimetre per year. This is yet another manifestation of rock deformation.

The development of major structural features of the Earth effects significant changes in climate and ecology. The specialized marsupial mammal fauna of Australia evolved and survived because of the deformation that effectively isolated their habitat from other continents. On a smaller scale, large mountain blocks can be raised rapidly (in a time interval of about 1,000,000 years) and these can drastically affect climatic patterns. The geologically young Sierra Nevadas in California effectively deplete the predominantly west to east air flow of moisture and create arid conditions east of the mountain mass.

The Earth's land areas can be divided into two parts: large continental shields, or cratons (areas that have been relatively stable and passive for millions--or even hundreds of millions—of years) and sinuous mobile belts (geosynclines) commonly one or two thousand kilometres long and 40 to 60 kilometres (25 to 40 miles) or more in width. These belts either occur along the margins of or transect the stable shield areas. Geologically rapid sedimentation, volcanic activity, and faulting and folding of rocks are integrated in these belts to develop intensely deformed zones. Until recently, the phases in the birth, development, and demise of mobile belts—a cycle lasting several tens of millions of years—were thought to have followed a fairly constant pattern throughout the world during the past 4,000,000,000 years of Earth history. Each cycle involves development and deformation of a geosyncline. Cratons (stable continental regions) such as the Canadian and Baltic shields comprise eroded remnants of ancient transecting mobile belts. The Canadian Shield extends southward and westward (in the United States and Canada) under a veneer of younger, mainly marine, sedimentary rocks deposited over the past 600,000,000 years, approximately. As the craton slowly oscillated up and down, periods of marine sedimentation alternated with uplift and erosion. Gentle folding, jointing (fracturing), and faulting (fracturing involving vertical or horizontal movements) accompanied these oscillations and were largely controlled by vertical movements on old joints and faults in the basement rocks. Examples are the Illinois, Michigan, and Williston basins, which are separated by broad arches and domes; each of these major structures is complicated by smaller folds and dislocations (faults). Because many of the veneer rocks (cratonic sediments) are oil and water reservoirs, their folding and faulting are of great economic importance.

The study of rock deformation

Structural geology involves the scientific study of the nature and development of deformed rocks in the Earth's crust, and also laboratory and theoretical attempts to simulate naturally deformed rocks. Interest ranges from the geometric interrelationships between crystalline grains at the microscopic or X-ray level to continental or oceanic units involved in global tectonics. Some of the oldest and unsolved geological problems concern the manner in which folds and faults form. Understanding of these problems would lead to more general concepts about the Earth's development and the nature of the most fundamental Earth processes. Initial scientific studies of fold and fault formation in the 19th century, for example, invoked compressive forces thought to result from contraction of the Earth. The essential importance of compression still pervades geologists' thinking, but during the 20th century there has come a realization that noncompressive forces are also significant, Many Soviet scientists contend that horizontal compression is a local, secondary phenomenon resulting from vertical movements of the crust and that the latter are really the fundamental movements involved. The correct balance between these extreme positions must await assembly of more factual data. Few early investigators dealt with folds and faults as prime objects of study; information accumulated incidentally during more general studies. In the last two decades, however, research aimed at understanding the mechanism of formation of folds and faults has been undertaken in several countries.

This article is concerned with the resnonse of rocks to imposed stresses, the several kinds of folding, faulting, and jointing that occur in nature, and the structural inteiference that results from the superposition of successive deformational events. For further information on the geologic and physiographic consequences of rock deformation, see MOUNTAIN BUILDING PROCESSES; MOUNTAIN RANGES AND MOUNTAIN BELTS; RIFT VALLEYS; OCEANIC RIDGES; CONTINENTAL DRIFT; SEA-FLOOR SPREADING; PHYSIOGRAPHIC EFFECTS OF TECTONISM. See also EARTH, STRUCTURE AND COMPOSITION OF; EARTHQUAKES; ISLAND ARCS.

STRESS AND STRAIN OF ROCKS

Response **to** stress. Essentially, all sedimentary, igneous, and metamorphic rocks are crystalline aggregates. When subjected to forces (stresses) in the laboratory, rock samples deform by fracture or distortion of their overall shape and internal structure or both. In nature, most rocks become deformed after their initial formation. Joints or fractures are almost ubiquitous. Shape distortion is most obvious when the initial rock contained features like bedding planes (visible surfaces on which sedimentation took place) or fossils; the present geometry of such markers is one measure of the degree of deformation. Initial horizontal bedding in sedimentary rocks may yieid simple folds that range in size from a few millimetres to several kilometres. Alternatively, rocks can be deformed and refolded to such an extent that all original structures are obliterated. Commonly, such penetrative deformation accompanies regional metamorphism (large-scale alteration of rocks under high temperatures and pressures) in mobile belts or the flow of extremely plastic rocks (*e.g.*, salt deposits) between rocks that are more rigid.

Stress components and the stress ellipsoid

Strain is the total change in volume, shape, and orientation resulting from stress. The stress (force per unit area) comprises body forces acting on every point in the rock equally (*e.g.*, gravity), and external or surface forces applied to the exterior of the rock unit. Stress commonly varies from point to point; theoretically, it can be calculated with respect to any plane through a rock. Normal stress, commonly designated *a* (sigma) is the component of total stress acting perpendicular to such a plane; tangential stress (shearing stress), commonly designated τ (tau), is the component parallel to the plane. Three mutually perpendicular (orthogonal) planes always exist such that the tangential stress τ equals zero, in which case the three normal stresses, designated $\sigma_1 > \sigma_2 > \sigma_3$, are the principal stresses. For a rock deformed deep within the Earth, the stress involves both confining pressure (an equal, normal stress across every possible plane as represented in Figure 1A) and shearing stress, which is conveniently resolved into three orthogonal normal stresses, σ'_1, σ'_2, and σ'_3, whose sum is zero. The shearing stresses can be represented by a stress ellipsoid (Figure 1C) with principal axes proportional to the magnitudes of *a*:, σ'_2, and σ'_3. Inclusion of confining pressure in Figure 1 adds a constant increment in every direction.

From (A,B) G. Wilson, "The Tectonic Significance of Small Scale Structures and Their Importance to the Geologist in the Field." Annals *of the Geological Society* of *Belgium*, vol. 84 (1961), (C) Proceedings of *the Geological Society* of *America* (1946)

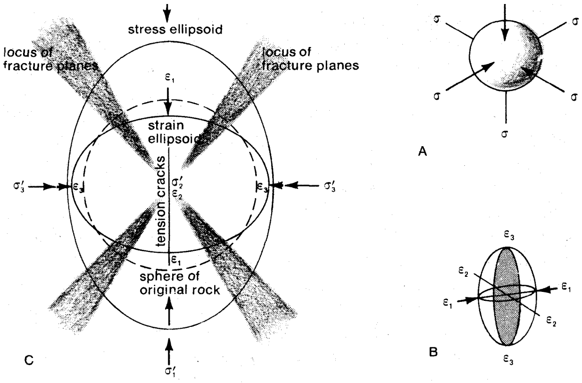


Figure 1: Relationship between stress and strain: (A) a spherical element of rock under confining pressure, with equal stress (σ) in every direction; (B) the strain ellipsoid, representing deformation of a sphere and the resulting strain axes (ε_1, ε_2, and ε_3); (C) relationship between stress and strain ellipsoids.

The response of rocks to stress is strongly affected by confining pressure, temperature, pore fluid pressure, rate at which strain is induced, length of time that stress is applied. and chemical environment. Granite and basalt, for example, are slightly ductile (can flow without fracture or faulting) at confining pressures of five kilobars at 300" C, or 570" F (average conditions about 21 kilometres [13 miles] below the Earth's surface; one kilobar [kb] equals 1,020 kilograms per square centimetre, or 14,504 pounds per square inch). Under experimental conditions of 150" C (300" F) in dry air the often used Solenhofen Limestone is brittle at 0.75 kilobar and ductile at one kilobar. Most rock deformation occurs beneath the Earth's surface where the average temperature increase with depth is approximately 1" C per 30 metres (1° F per 55 feet) and confining pressure increases at the rate of 0.23 kilogram per square centimetre per metre. At temperatures and pressures of modest depths, the strongest rocks flow and undergo permanent strain whereas, near the surface or when subjected to rapid strain rates, the same rocks may buckle or fracture.

Most theoretical stress–strain studies relate to elastic deformation (infinitesimal reversible displacements) in homogeneous substances. The actual strain of rocks, however, involves permanent deformation (finite strain) of heterogeneous media. Compositionally and structurally, the Earth's crust, whether considered on the scale of a continent or a microscope slide of a rock, is heterogeneous. When individual mineral grains are considered, the rocks containing them are always heterogeneous. In larger samples the' mineral variation becomes less important, however, and rocks approach homogeneity sufficiently to accommodate the principles that govern stress–strain relations in homogeneous rocks.

Elastic **and** plastic deformation. At the onset of stress, strain is in the elastic range, meaning that recoil to the original shape occurs with stress removal (except for minor residual strain due to pore-space collapse). Some rocks suddenly fracture (brittle failure) in this range. If brittle failure does not occur, additional stress carries the rock beyond the elastic limit (at the yield point) and plastic deformation produces permanent strain that persists after removal of the stress (*e.g.*, ε_1 or· ε_2 in Figure 2). For a given rock, yield-point stress is decreased by increased temperature, and increased by an increased

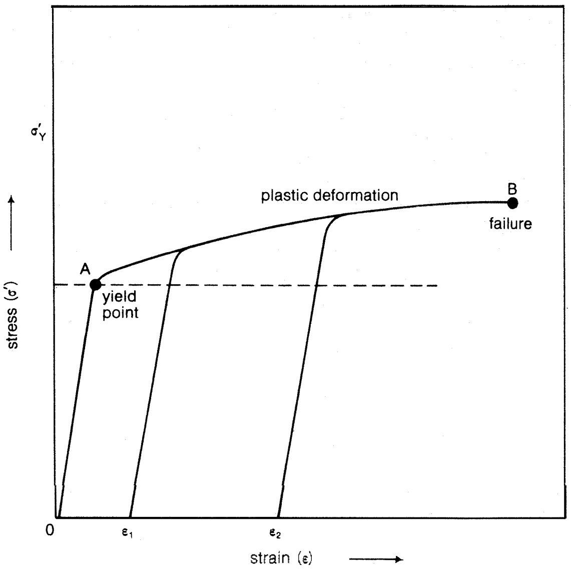


Figure 2: Generalized stress-strain relationships (see text),
From J.G. Ramsay, *Folding and Fracturing of Rocks*,Copyright 1967; used with permission of *McGraw-Hill Book Co.*

Effects of tempera- ture and time

confining pressure or strain rate. For a perfectly plastic body, the slope of AB (Figure 2) would be zero, but rocks exhibit strain hardening so that in order for plastic deformation to continue, the stress must increase beyond the yield-point stress. The mechanism of plastic flow varies: when the temperature is low, cataclastic flow by crushing, fracturing, and mechanical granulation occurs; when the temverature is moderate. distortion. bending, and crystal twinning occur, but no crushing or new grain formation take place; when the temperature is high, recrystallization and the growth of new grains with new orientations, shapes, and distributions occur—the rock behaves almost like a perfect plastic material.

Strain in rocks combines elastic, viscous, and plastic properties. Time-dependent deformation is creep. When stress is first imposed, the strain is elastic followed by fairly rapid primary creep (Figure 3). With increasing time, the rate of strain $d\varepsilon/dt$ becomes constant; during this secondary phase steady-state, or pseudoviscous, creep occurs. During the tertiary phase, creep rates increase to failure. Recoil associated with stress removal is indicated by broken lines in Figure 3. Directed stress in the Earth's crust may be maintained for many thousands of years (or even for millions of years); some authorities believe that under such conditions elastic, plastic, and other components are unimportant so that rocks essentially behave like viscous liquids. Materials in which viscous flow—as a steady-state process—is the only significant strain have been called rheids. Though this concept seems to explain many geological observations, it has been criticized by physicists. Unfortunately, it is difficult to account adequately for geologically significant time in experimental or theoretical work.

From J.G. Ramsay, *Folding and Fracturing of Rocks*, Copyright *1967;* **used** with permission of McGraw-Hill *Book Co.*

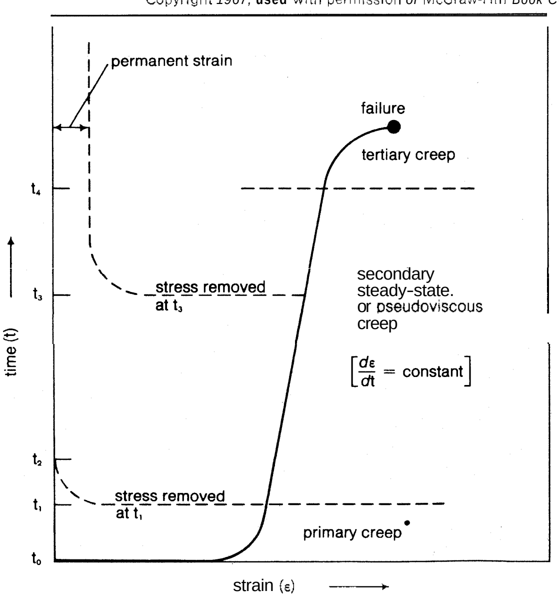


Figure 3: Generalized time-strain relationships with constant stress and with stress termination (see text).

The total strain results in changes in position (translation), orientation (rotation), volume (dilation), and shape (distortion). These are all geologically important. Translation and rotation are significant in mobile belts, for example, but observable evidence within rock samples is usually limited to dilation and distortion—*i.e.*, to pure strain. Pure strain can be considered in terms of the strain ellipsoid (Figure 1B), whose three principal axes are commonly designated by the Greek letter epsilon, $\varepsilon_1 > \varepsilon_2 > \varepsilon_3$ (which correspond to σ'_1, σ'_2, and oi); this ellipsoid is the shape assumed by an imaginary sphere of unstrained rock following the effects of stress (see Figure 1). The shapes of predeformation structures preserved in a deformed rock give the rock a geometric pattern that can be observed and described objectively and quantitatively. Though the movements of individual parts of the rock that resulted in the new geometry commonly can be reconstructed objectively, these kinematic movements cannot be directly observed. **A** unique set of forces produced the kinematic and geometric results but, being unobservable, the actual forces are commonly subject to considerable controversy. In consequence, knowledge of strain behaviour in rocks relies heavily on experimental laboratory work on rock samples and on mathematical analyses of idealized materials.

The strain ellipsoid

Numerous types of finite homogeneous strain are possible, but general strain is the most common in rocks. In this, pure strain (pure extension or compression along three orthogonal axes) is combined with pure rotation of the whole system about any axis or axes. Simpler strains are occasionally encountered in special, limited locations. Unrealistically, many geological analyses have been limited to two dimensions and based on plane strains in which strain was assumed to have occurred along only two orthogonal axes.

Wherever the strain is homogeneous, each part of the rock can be represented by the same strain ellipsoid; also a straight line, a plane, and a sphere become transformed into another line, a plane, and an ellipsoid, respectively. In most geological situations, heterogeneous, rather than homogeneous, strains occur (as in a simple fold). In a folded conglomerate (coarse-grained sedimentary rock), for example, the pebbles and the matrix are likely to have different physical properties and to be dissimilarly strained. In addition, mechanical heterogeneity causes the stress field to be uneven throughout the sample and this in turn produces additional strain dissimilarities. Commonly, smaller domains can be found in which strain is homogeneous; in many geological geometrical analyses, the first step involves identifying domains that are homogeneous with respect to structural properties of interest.

Analysis of grains in strained rocks

On the basis of extensive work in structural petrology, it is commonly believed that the symmetry of the stress factors is reflected in the symmetry of the strained rock. Each grain in a rock responds individually to a directed-stress field, so that the grains develop a preferred habit (shape) or lattice (molecular structure) orientation or both. According to classical concepts the process is believed to result from either direct componental movements (bodily rotation of existing grains) or indirect componental movements (recrystallization of existing minerals or growth of new, more stable, minerals with their molecular structure preferentially oriented with respect to the local directed-stress field). Electron microscopy, metallurgical research, and detailed consideration of the thermodynamic principles of solution and redisposition are currently shedding much light on the processes involved in lattice orientations and the development of the very common fine banding of recrystallized metamorphic rocks. Habit orientation is also a fundamental (though vexing) research topic of currency; it is discussed further in connection with foliation and cleavage development in the next section of this article.

FOLDING OF ROCKS

Types ***of tectonic folding.*** All rocks can be folded when the stress is sufficient to produce permanent strain, but folds are more obvious where original planar structures (such as bedding or schistosity) were deformed during the total strain. In the mobile belts, fold wavelengths (distance from crest to crest) vary from a few millimetres to several kilometres. Traditionally, most geologists have recognized three types of folding, but gradations between these are readily found in nature.

Flexure folding. The rock is structurally homogeneous (containing no planes of mechanical discontinuity) and no planes of discontinuity develop during flexure folding. A sheet of India rubber or modelling clay is an appropriate analogue—when folded, the concave side is compressed and the convex side is stretched without fracture development. Such strain is relatively rare in rocks.

Slip folding. The rock is structurally homogeneous but the stress induces a single set. of parallel slip planes (S') during slip folding. Folds of this type are common in slates where colour banding or very minor compositional differences define original bedding (S_0) that is folded by differential slip along innumerable new S' planes. The geometry can be illustrated with a deck of cards (Figure 4A). A heavy line can be drawn around the cards to represent bedding (S,), and it can be assumed that the cards are a solid homogeneous mass before deformation and that the planes (S') between the cards result from the stress; differential slip on these induced planes produces folds in S_0 that are variously called slip, similar, or bending folds. The genetic term bending fold implies folding caused by motion transverse to bedding that was not generated by compression along the rock layers (S,,).

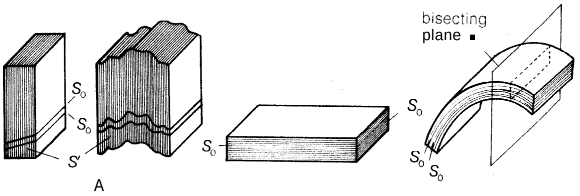


Figure 4: Simple fold models illustrated by decks of cards. (A) Slip folding of bedding S_0 along slip planes S'. (B) Flexural-slip folding of bedding S_0 (see text).

Flexural-slip folding. The initial rock has planes of mechanical discontinuity (*e.g.*, well-marked bedding planes) along which slip can occur during stress. A deck of cards can also illustrate such folding if the separation between the cards is assumed to correspond to original bedding (S,). Compression or shortening parallel to the layering (or bending of the layers) causes folding with slip (S') between the cards parallel to bedding (S,); the result is variously called flexural-slip, parallel, or buckling folds (Figure 4B).

Successive rock layers tend to be mechanically dissimilar during folding. Well-bedded sandstones, for example, may respond by flexural slip while interstratified silts and clays may develop slip-fold characteristics. Such complex relationships are common. Kock between the successive slip planes of flexural-slip folds is also strained during folding, but whether by flexure or by one of the other types of folding is not entirely clear.

From J. Ramsay, *Journal of Geology*, vol. 70 (1962); University of Chicago Press

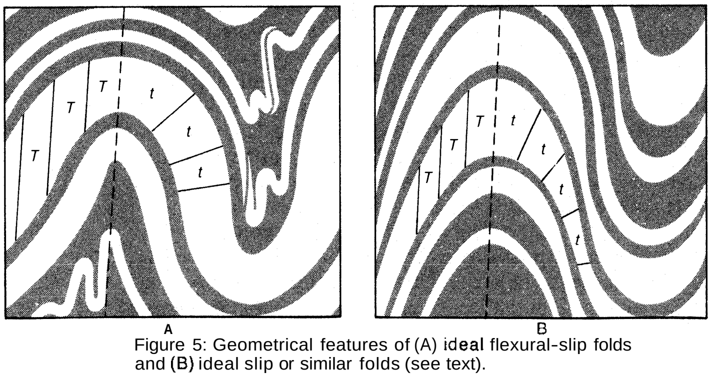


Figure 5: Geometrical features of (A) ideal flexural-slip folds and (B) ideal slip or similar folds (see text).

Theoretically, the geometries are simple and distinct. In ideal flexural-slip or parallel folds (Figure 5A), the orthogonal thickness (t) of each bed remains constant whereas "thickness" (T) measured parallel to the bisecting plane of the fold changes from point to point. By contrast, in slip folds the orthogonal thickness (t) varies while T remains constant (Figure 5B). It has generally been assumed that flexural-slip and slip folds are the dominant types in nature, with the former more common. These models have been recognized since 1896, but recently it has been noticed that few folds really have these geometries. In fact, most folds appear to depart markedly from the ideal *t* and T relationships, which calls into question the basic hypotheses about fold formation. In an attempt to explain this anomaly, it has been suggested that, after the initial fold formation, continued compression causes the structure to be flattened in the plane normal to the original maximum stress (*i.e.*, flattening approximately in the bisecting plane). Limited published *t* and T data for natural and experimentally produced folds make it difficult to evaluate the flattening hypothesis. If "flattening" is as common in numerous environments as has been suggested, most ideal fold shapes must be distorted.

Departure from ideal fold geometries

The terms flexural-slip, slip, and flexure fold clearly have genetic implications suggesting that the kinematic mode of formation is known unequivocally and that the folds conform to the *t* and T specifications. For this reason, many geologists use the older terms, similar and parallel folds, instead of slip and flexural-slip folds, respectively. The older terms refer to what now appears to be rare, idealized geometry; supposedly, in slip folds, successive folded layers have similar shape, whereas, in flexural-slip folds, boundaries between successive layers are parallel (Figure 5). Following the genetic terminology developed in strength-of-materials studies, some geologists prefer the terms bending and buckling folds.

The geological literature is burdened with dozens of terms descriptive of fold shapes. New terms accumulate rapidly and, in most cases, without the use of systematic concepts or usable operational definitions. Many terms overlap in meaning or are dissimilarly used by different geologists; some relate to purely geometrical characteristics, some to genetic concepts, and others to a mixture between geometric and genetic properties. A list of the descriptive terms for folds used in the principal English-language structural-geology textbooks (abnormal anticlinorium, abnormal synclinorium, accordion fold, allochthonous fold, ameboid fold, etc.) with brief definitions occupies 24 printed pages. Because a fold is a geometric form its description should be based on its geometry and not intertwined with the kinematic movements and dynamic forces (or both) that produced it.

Cylindrical and conical folds

When studied in units of convenient size, almost all folded rocks can be considered as cylindrical or conical geometric structures. Normals (perpendicular lines) to the sides of a cylinder are perpendicular to the cylinder axis; normals to bedding in a cylindrical fold are perpendicular to the fold axis. The fold axis is a direction (in three-dimensional space) but has no particular position; it is the imaginary line that, if moved parallel to itself, always remains within the folded surface. If the top and bottom of a sheet of ruled notebook paper are moved toward each other (so that the lines on the paper remain straight and parallel), the paper simulates a cylindrical fold with the fold axis parallel to each line on the paper. For standardization, fold shape is described in terms of the profile—the trace of the bed on a plane normal to the fold axis. An idealized cylindrical fold has identical profiles at all points along its axis. If, as commonly happens, the profile changes along the fold axis (at different rates in each fold), the geometry approaches that of a conical fold. A conical fold is the locus of a line constrained to pass through a fixed point; the fold axis is then the cone axis. A cylindrical fold is a special case of the conical fold in which the apical angle of the cone is zero. Although individual folds may be conical, when whole fold systems are considered, the folds are frequently approximately cylindrical; when seen in plan, such folds resemble a set of en echelon (parallel but offset) canoe-shaped structures. Thus, within a single system, folds of one size may be essentially cylindrical, and those of another size may be conical.

For complete description of a cylindrical fold, it is necessary to specify:

1. Fold axis orientation: usually in terms of trend and plunge. Trend is the orientation (in degrees east of north) of the projection of the axis on a horizontal surface (AB in Figure 6); plunge is the angle measured in a vertical plane between the fold axis and horizontal (angle between AB and AD in Figure 6).

From The Earth: An introduction to *Physical* Geology, by John Verhoogen F.J. Turner, L.E. Weiss, and C. Wahrhaftig. Copyright © 1970 by Holt, Rinehart and Winston, Inc. Adapted and used by permission of Holt. Rinehart and Winston, Inc.

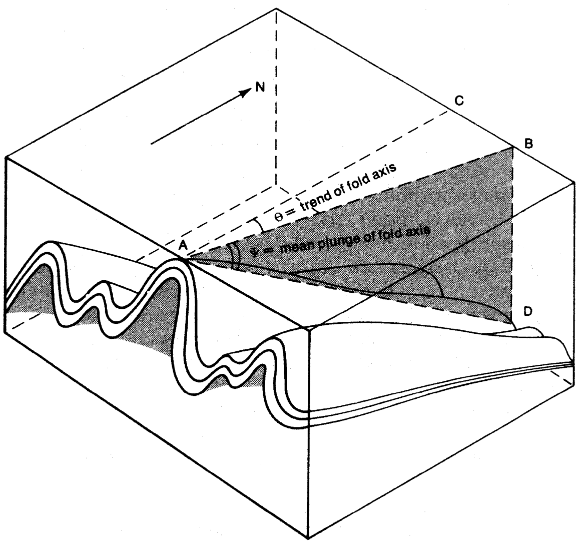


Figure 8: A group of plunging cylindrical folds, showing trend and mean plunge of one fold (see text).

2. Fold size: commonly, a wavelength measured on the profile.

3. Fold shape as seen in profile: qualitative descriptive terms are commonly used; although quantitative description techniques are being developed, wholly satisfactory methods are not yet available.

When several layers are folded together, the surface through successive fold crests is the axial surface (or plane), which may or may not coincide with the bisecting surface (or plane) defined by a dominant bedding plane. When successive beds are of different thicknesses and lithologies, folds developed in them tend to be of different sizes and of dissimilar profiles, although the axes are commonly approximately parallel (*i.e.*, homoaxial). The folds are disharmonic when profiles of successive layers are dissimilar. A fold whose limbs approach one another downward is called a synform; if the limbs diverge downward it is an antiform. Folds are upward facing when stratigraphically younger rocks are crossed in traversing upward along the bisecting plane. In intensely folded areas, the stratigraphic succession may be completely inverted so that folds face downward. If it is positively known that the folds are upward facing, they are properly called synclines and anticlines (instead of synforms and antiforms, respectively).

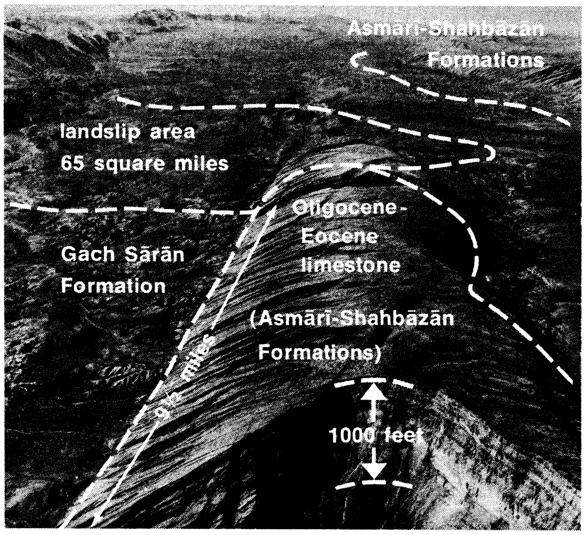

Asmārī-Shahbāzān Formations
landslip area
65 square miles
Gach Sārān Formation
Oligocene-Eocene limestone
(Asmārī-Shahbāzān Formations)
6½ miles
1000 feet

Figure 7: Large anticline in the Oligocene-Eocene Asmārī-Shahbāzān Formations, Lurestan, Iran (aerial view looking northwest). The anticline is seen disappearing under the immense landslip area.

By courtesy of the BP Petroleum Development Limited, London

Very large folds within a mobile belt can be traced for several hundred kilometres, although structures of this size can be recognized only on the basis of careful mapping. When the wave length is a few kilometres, the entire structure can sometimes be seen simultaneously if exposure conditions are particularly good; air photographs are especially useful for studying such structures (Figure 7). Folds with wave lengths ranging from several tens of metres to a few centimetres can be seen readily in the field (Figure 8). With a hand lens, details of even smaller folds that commonly occur in more incompetent rocks can be seen. Not infrequently, smaller homoaxial folds occur on the flanks and in the cores of larger folds--this applies to structures of all sizes. In indifferently exposed terrain, the small folds—parasitic folds—assist in detecting the nature of major folds that can be interpreted only by mapping and piecing together information from numerous outcrops. Individual parasitic folds may be flexural-slip, slip, or flexure folds depending on the lithologies and other factors involved; the poor genetic term drag fold is frequently used for parasitic folds.

Foliation and lineation. Commonly, in rocks strained at all but very shallow depths, foliation (rock cleavage) has developed approximately parallel to the bisecting planes of folds. Foliation can be particularly obvious in rocks containing abundant phyllosilicates (chlorites, biotite~,etc.); the approximate parallelism of such flaky

From F.A Donath and R.B. Parker, "Folds and Folding " Bulletin of *Geological* Society *of* America, vol. 75 (1964); Geological Society of America, Inc.

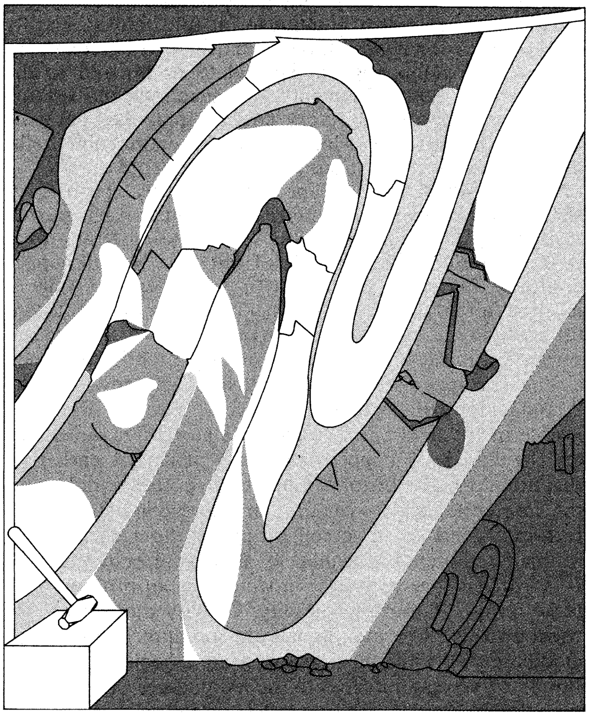


Figure 8: Flexural-flow fold in thick- and thin-bedded shales, Black Rock Mine, northwestern Queensland, Australia; hammer (lower left) provides scale.

minerals largely defines the foliation in hand specimens. If, as has commonly been supposed, such foliations develop perpendicular to either the maximum principal compressive stress that caused the strain or the direction of minimum total elongation (maximum shortening), it is difficult to explain why slip along *S′* is dominant in slip folds. This problem does not arise with the flexural-slip model, where both the bisecting plane and the foliation are approximately normal to the minimum total elongation for the whole fold. A satisfying explanation is not available. It has been suggested that in slip folds the *S′* planes are only approximately parallel, so that a small component of the greatest principal stress acts along each *S′* plane inducing differential slip. In varied clay, silt, and fine-sand formations folded at shallow depths, weak foliations have been attributed to the mechanical rotation of original clay minerals by pore water expressed during folding. Minor veinlets of sand parallel to the foliation and cutting through the silt and clay layers (and clay veinlets cutting the coarser grained units) support this hypothesis, but the theory may not adequately explain the axial-plane orientation of the veinlets and foliation.

Although most foliations form parallel to the fold axial planes, incompetent (weaker) beds located between more massive units frequently develop foliation (schistosity) parallel to bedding, due to shear during flexural-slip folding.

Various linear structures are generated in all strained rocks, except those deformed at very shallow depths. Structures within the parent rock become deformed during folding; for example, pebbles and carbonate ooliths (small, spherical particles) become elongate with their long axes approximately parallel so as to define a linear structure. Following intense deformation, individual grains may be strained sufficiently to show a distinct lineation in a hand specimen—viewed with unaided eye—especially on weathered surfaces; under the microscope, such preferred orientations can be detected and measured more easily and precisely. Such lineations tend to be parallel to the fold axes in metamorphosed rocks, whereas in less metamorphosed rocks they lie within the axial plane and normal to the fold axis. Lineations pervading the entire rock are penetrative, whereas nonpenetrative striations (slickensides) parallel to the slip direction (*i.e.*, normal to the fold axis) sometimes develop on bed surfaces subjected to flexural slip.

During flexural-slip folding of varied lithological sequences, the more competent units frequently develop a linear structure called boudinage on fold flanks; sharply defined necks or zones along which the bed is attenuated (necked) develop aligned parallel to or normal to the fold axis or both. Frequently, intersections of bedding and schistosity and the very small, ripple-like parasitic folds on the surfaces of large folds of metamorphic rocks define prominent lineations parallel to the fold axes.

Nappes and diapirs. Nappe is the general term for a sheet of rocks several kilometres in extent that has moved forward (for distances up to several kilometres) over rock formations beneath and in front of it. Not all nappes comprise individual folds, but a very large recumbent isocline (a fold with both limbs exhibiting a near-horizontal attitude) is one variety of nappe. Such structures are important in the architecture of ancient mobile belts such as the European Alps, the Scottish Highlands, and the Appalachian fold belt. In such areas, many of the folds visible in the field are second-phase structures developed on the flanks of nappes that were produced earlier in the mobile belt's history. Current hypotheses hold that the immense nappe structures are a response to gravity tectonics—that is, to creep or plastic flow of partially lithified sedimentary sequences shortly after deposition. It is implicit that very slow movements can produce very large strains if the time is long. Gravity movements toward a trough in a mobile belt are sometimes referred to as syntaphral; syntaphral tectonics are associated with the supply of creeping masses to the axial zones of a geosyncline. Such zones later become the axial zone of the tectonic orogen (mountain core) so that the early soft-sediment folds and nappes almost invariably become overprinted by younger deformation. The classical concept of folding being initiated by viselike compression is not applicable to gravity-tectonic or to syntaphral phenomena.

Gravity tectonics in mountain building

Diapirs, a special type of fold structure, appear to result from the ascent of a plastic core of less dense rock that pierces and dilates the overlying and surrounding rocks. Using a less dense, plastic medium below more dense material and simulating gravity operating over a long period with a centrifuge, a wide variety of naturally occurring diapiric structures have been simulated in the laboratory. In these experiments, a circular plug of the less dense medium typically ascends into overburden; the stalk connecting the plug to the buried parent layer attenuates so that an egg-shaped mass of the less dense material eventually separates from the parent layer completely and continues to rise upward (see further SALT DOMES).

Granitic diapirs

Aside from rock salt, the most common diapiric substance is granitic material. Numerous small (16 to 20 kilometres in diameter) granite masses (*e.g.*, Flamanville Granite on the northwest coast of France) have been emplaced by dilation of the enveloping rocks during magmatic intrusion—the ascent of hot, molten silicate material. In many Precambrian (older than 570,000,000 years) shield areas circular, dome-shaped granitic complexes abound. The so-called mantled gneiss (coarsely foliated metamorphic rock) domes of Karelia (Finland and the Soviet Union) and the Vredefort Dome (near Johannesburg, South Africa) are good examples. Although their origins are controversial, a diapiric origin, with relatively low-density granitic material rising plastically through more dense material, appears reasonable in the light of recent studies. The examples cited seem to occur at axial culminations (intersections of major perpendicular antiformal structures) that facilitated and localized diapir formation. Analogous structures have been mapped over large tracts of Rhodesia; the Canadian Shield of Saskatchewan and Manitoba has numerous comparable structures.

Nontectonic folding. The development of small, nontectonic folds in relatively unconsolidated surficial rocks and recently deposited sediments is also important. Submarine sliding and slumping is widespread and more important; the process produces small folds contemporaneously (or nearly so) with sedimentation. Gravitational slip along bedding planes prior to additional sedimentation means that the folds affect only a metre or two of sediment; the underlying strata and the overlying (as yet undeposited) strata are unaffected. Lithified examples seen in outcrop can look deceptively like tectonic folds. Following tectonic deformation, the superposed tectonic folds make it difficult to recognize the earlier soft-sediment structures; a good example of this situation was described in graywackes (a type of sandstone with a muddy matrix) from the Appalachian mobile belt of Newfoundland. Gravitational creep and flow in sedimentary rocks that are exposed in deeply dissected terrain on land causes crumpling and folding in disturbed zones up to a few metres thick, several good examples from the Italian Apennines have recently been described.

Under special conditions, numerous other soft-sediment folds can develop in largely unconsolidated sediments. For example, differential compaction, extrusion of water during compaction and dewatering, and decay of ice bodies within periglacial areas are examples of conditions that can induce minor fold structures.

FRACTURE IN ROCKS

Rock units at the Earth's surface are always intersected by fracture and rupture surfaces. As pointed out earlier, rocks respond to stress either by plastic deformation or by failure (fracture), depending on pressure, temperature, and other conditions. The fractures commonly result from stress imposed during earth movements and tectonism; however, thermal contraction of igneous rocks and dessication of sedimentary rocks also produce fractures, as do residual stresses within rocks formed at depth in the crust but brought to the surface as erosion slowly strips away the overlying rocks. Individual fractures range from minute structures visible under the microscope to some of the largest recognized geological structures in the Earth's crust. These fractures fall into two main categories—joints, which are fractures along which little or no displacement has occurred, and faults, which are fractures along which significant displacements have occurred.

Economic and geologic importance of joints

Joints, joint sets, and joint patterns. Because they facilitate easy flow within relatively impervious rocks, joints afford excellent groundwater, oil, and gas reservoirs of considerable economic importance. In addition, joints are exploited in quarrying and mining operations, and stone masons customarily utilize the many invisible planes of weakness that are parallel to the visible fractures. For water wells, some success has recently been achieved in locating the intersections of master joints on air photographs (even in country blanketed by a few metres of soil or other overburden). If the host rock is a compact limestone, for example, appreciably better water flow is obtained by sinking wells to intersecting joints. Small explosive charges are sometimes used at the bottom of a well to shatter the surrounding rock and to provide access to the natural joint system. Solution of limestone by groundwater circulating through joint systems commonly results in pot holes and extensive underground cave systems (enjoyed by speleologists) and eventually leads to development of karst topography and impressive gorges (see further CAVES AND CAVE SYSTEMS).

Joints are probably brittle fractures in which displacement (zero to several centimetres) normal to the fracture surface is characteristic; sometimes, the dilation appears to be greater, but many of the very open joints in surface outcrops were enlarged and widened by surface erosion. Commonly, two or more sets of regular, parallel, closely spaced joints—systematic joints—are clearly visible in outcrops; scattered, non-systematic joints of more variable orientation and not crossing other joint sets are not uncommon.

In laboratory compression tests of rock samples, brittle-fracture surfaces form either parallel to or oblique to the maximum compression axis and produce extension fractures or shear fractures (or faults), respectively. In simple tensile tests, the fracture surfaces form approximately perpendicular to the tensile-stress axis. High pore-water pressures reduce the rupture strength of rocks and offset the stress due to the weight of the overlying rock column. Despite these simple experimental results, the origin of natural joints is not well understood and different authorities hold quite dissimilar views. None of the numerous theories for joint development is wholly acceptable. One immediate problem is that there are few detailed studies of the nature and variability of joints, although joints are visible in essentially every outcrop and air photograph. Second, there are probably numerous dissimilar mechanisms that induce brittle fractures in rocks, so that a single genetic theory is inadequate. Analysis of natural joints is complicated by the common superposition of successive, unrelated joint sets. Slight offset of one joint set by another sometimes permits the relative ages to be established.

Relation of joints and faults to folding

It is widely believed that joints and faults have as a common origin the deformative stress that occurs during a phase of folding. Many cylindrical and conical folds are associated with well-developed cross joints (normal to fold axis) and longitudinal joints (parallel to axial surface). Less pronounced diagonal joints occur as paired (conjugate) sets symmetrical to the cross and longitudinal joints but more closely inclined to the former. Joints tend to be of two types: (1) tension fractures characterized by clean granular breaks (sometimes with plumose, or radiating, patterns reflecting the spreading fracture through the rock) oriented parallel to the maximum-stress axis and perpendicular to the minimum-stress axis; (2) shear fractures occurring as conjugate sets with slight offsets and slickensided (below) surfaces; these joints or faults develop in planes parallel to the median stress direction, and they have an acute angle between them that is bisected by the maximum-stress direction. Slickensides are grooves or striations on joint or fault surfaces produced parallel to the last direction of forceful movement of one face across the other. Close relationships between joint sets and fold geometry are widespread and, when actually established, fracture-pattern analysis can have far-reaching significance in determining the orientation of tectonic forces responsible for regional structures (*e.g.*, anticlines and mountain uplifts). Close agreement to theory is illustrated by an example from southeastern Algeria (Figure 9).

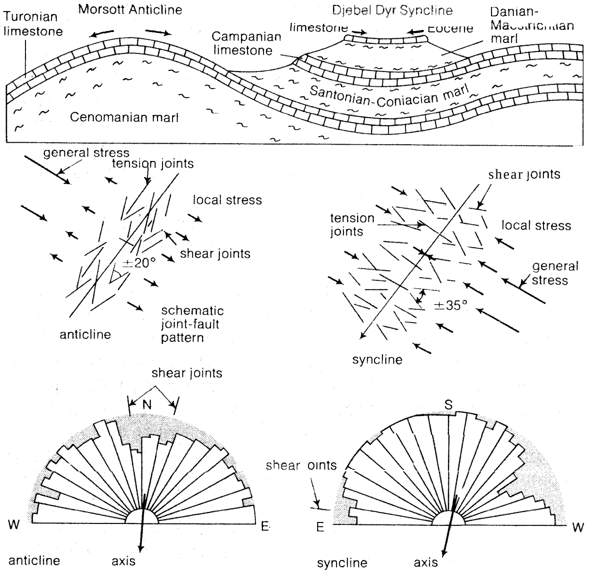

From L.U. do Sitter. *Structural* Geology, 2nd ed. (1964); McGraw-Hill Book Co.

Figure 9: Origin of joint and fault patterns in adjoining anticline and syncline, adapted from aerial photographs of adjacent structures in southeast Algeria, with frequency diagrams of the strikes of joints and faults.

Where there appears to be a clear genetic interrelationship between joints and folds, the fracture density tends to be greater in thinner rock units and in the more brittle rocks. Fracture frequency is commonly correlative with the degree of bed curvature, so that fold flanks have fewer fractures than the closures; such fractures, seen on fold profiles, have sometimes been called fracture cleavage, but they are true joints and should be differentiated from cleavage or foliation. Foliation, unlike joints, is characterized by a splitting parallel to a preferred mineral orientation (*e.g.*, schistosity or slaty cleavage). Occasionally, this type of fracture and slaty cleavage grade into each other; this is additional evidence for foliation and folding being concomitant and genetically related. In metamorphic rocks, tensional-joint formation is commonly aided by the foliation, which customarily parallels the fold axial surfaces.

Propagation of basement joints

In many other situations, joints appear to be independent of the stress field that previously effected strain of the rocks. Sedimentary rocks on stable cratonic areas, for example, are commonly almost unfolded or folded only very mildly. Such rocks are invariably jointed, however, and the joint patterns remain remarkably constant in orientation over vast areas, although varying in frequency from one rock type to another. Similarly, the highly folded and metamorphosed shield areas are characteristically transected by regional joint sets bearing no easily discerned relationship to the local structural geometry. Such joints apparently reflect major regional stress phenomena. Sometimes consistent offsets suggest shear components. More commonly, three mutually perpendicular (orthogonal) joint sets are dominant—two nearly vertical and one approximately horizontal. This suggests that basement joints—those in the oldest crystalline rocks at the bottom of the sequence—were propagated into the overlying cratonic rocks (and even into Pleistocene sedimentary rocks) by either (1) continuation of the processes that produced the original basement joints or (2) relatively small vertical movements on the old joints causing extension of those joints up into the overlying rocks. Detailed observation in Finnish mines east of the Gulf of Bothnia reveals continuing small movements on Precambrian joints; these particular movements probably reflect continuing isostatic readjustment following melting of the Pleistocene ice sheets. It has been demonstrated that movement on old basement joints is widespread and has dominated the map pattern in Finnish basement rocks (particularly in the Orijarvi area). Similar posthumous movements appear to be characteristic of the southern Canadian Shield. Such movements probably initiated jointing in the Paleozoic sedimentary rocks of the north central United States that blanket the southward extension of the Canadian Shield. This mode of joint development soon after deposition of sedimentary rocks is probably a common phenomenon. Some geologists believe that such regional joints are fatigue fractures, resulting from continued cyclic elastic strain of the rocks due to periodic waves affecting the whole Earth's crust. Seismic waves and solid earth tides have been cited, but numerous other significant periodic stresses affect the crust. It seems unrealistic at present to attribute the regional joint sets to any one of these possible causes because of the lack of knowledge that prevails.

Joints do not persist downward indefinitely because eventually the domain of plastic deformation is reached. Frequently joints are much more open at surface outcrop and more appressed (or only incipient) a few metres below the surface. It has been argued that joints are brittle fractures caused by stresses developed in rocks raised from within the Earth's crust by uplift and erosion. During such uplift a body of rock must undergo a lateral expansion simply because the Earth is spherical and this, on geometric grounds alone, might be expected to produce more open joints near the Earth's surface. It also has been suggested that uplift and erosional unloading induce formation of vertical conjugate shear joints and, at shallow depth, vertical extension fractures.

Desiccation and cooling patterns

Many joints are the product of nontectonic forces. During the desiccation of clays and silts on playa lake beds, shrinkage produces vertical tension fractures in a polygonal plan. Individual polygons are usually 0.2 to 0.6 metre (0.7 to two feet) across and, with continued desiccation, the joints extend downward. Aerial photographs show that major joints also develop on playa lake beds defining polygons 100 to 200 metres (330 to 660 feet) across. Some geologists have suggested that, during dewatering and lithification of many primary sediments, joints develop perpendicular to bedding. This is possible, though not as yet proven.

Distinctive joint patterns characterize igneous rocks and are generally attributed to tensional stresses during post-crystallization cooling and contraction. Dikes and sills, which are tabular intrusive bodies, and many lava flows tend to develop columnar joints; the latter define polygonal columns perpendicular to the cooling surfaces. Individual columns range from seven to eight centimetres (2.75 to three inches) to as much as six metres (20 feet) in "diameter." The column "diameter" seems to be controlled by the cooling rate; a zone of smaller columns occurs at the base and surface of thick lava flows, whereas much thicker columns characterize the centre. Characteristically, curving tensional joints occur perpendicular to the column length. At Giant's Causeway in northeastern Ireland, basalt flows of Early Tertiary age (about 55,000,000 years old) are exposed in sea cliffs and provide an excellent three-dimensional view of columnar jointing. The extensive basaltic lava flows of the Columbia Plateau, Washington, are equally impressive. Although very common, these simple polygonal joints are not universal. For example, basalt flows in southern Nevada are well exposed, capping mesas around the edge of which the joints appear to be columnar. In detail, the joint patterns are complex and not infrequently dominated by rectilinear vertical joints many of which are parallel to and physically continuous with the more widely spaced, but regionally developed, joints of the underlying sedimentary rocks that are significantly older.

Joint sets in large igneous bodies

In major igneous intrusions three mutually perpendicular joint sets are characteristic, with two sets roughly vertical. Many plutons (large intrusive bodies) have distinctive planar preferred orientations of the component mineral grains. In many cases, these planar structures are nearly vertical and are interpreted as products of flow during intrusion and crystallization. The vertical orthogonal fractures comprise cross and longitudinal joints that are roughly perpendicular to and parallel to the planar structure. Also, less well developed, conjugate, vertical fractures (diagonal joints) are symmetrically disposed to the flow structure. Joints of granitoid batholiths (very large masses of igneous rock) probably arise from cooling and contraction of the crystallized magma. Early formed joints sometimes become annealed prior to formation of the present fractures; for example, annealed old fracture planes, demarcated solely by iron pyrite, occur in some Front Range granites of Colorado. Joint faces (or the whole joint cavity) in many igneous rocks are encrusted with secondary minerals, formed after crystallization of the principal rock mass. Late stage veins and dikes cutting plutons commonly utilize previously formed joints. Many intrusions have additional fractures, however, some of which may be responses to postcrystallization deformation. Some batholiths even have well-developed orthogonal joints that are apparently unrelated to the internal mineral orientations; for example, the Older Granite of Donegal in northwestern Ireland is characterized by strongly developed north–south and east–west vertical joints and a subhorizontal set, which disregard well-defined flow-structure orientations.

Faults **and faulting.** Faults are fractures along which the rocks on opposite sides have moved so as to produce significant displacement. Such displacements range from a few millimetres or less up to several hundred kilometres. An initial rupture surface or plane of failure in the rock provides the plane along which fault displacement occurs. When strain in homogeneous rocks is involved, there is little problem in differentiating between joints, faults, and folds. With natural heterogeneous rock units, however, transitions between faults and folds abound. Numerous, small shear planes characterize the small internal structures of many folds. Faults large enough to be shown on a map may die out and grade into folds. Similarly, there appears to be a complete gradation from small joints to major faults, although this may be misleading because there are many different types of joints and, apparently, an even greater diversity of fault types.

The net slip or total length of displacement on a fault can range from essentially zero (as in joints) up to perhaps 600 kilometres (400 miles). The distribution of faults is very uneven; some large areas are almost unfaulted, and others are cut by innumerable faults of varying size. Movement on many faults has been or is very spasmodic with rapid movements (lasting a few seconds) of up to a few metres being separated by intervals during which stress builds up and is eventually released when frictional forces along the fault plane are overcome. Rapid release produces an earthquake. Movement on some currently active faults, however, seems to be by continuous creep rather than spasmodic jumps. Many faults were active for only a brief period, whereas some others moved spasmodically and played a major role in Earth history over hundreds of millions of years.

The three types of faults

Three types of faults can be distinguished on the basis of theoretical stress systems that can be resolved into three orthogonal principal components $\sigma_1 > \sigma_2 > \sigma_3$, based on a model in which one component is approximately vertical and the fracture planes are inclined 15° to 45" to the largest principal stress component (σ_1). The three generally accepted possibilities involve: thrust faults, wrench, tear, or transcurrent faults, and normal or tension faults (Figure 10). Reverse faults are also common in some areas but they are unexplained by this model. This system can readily be extended to joint formation, however, especially if a fourth category is added, namely tension joints (equivalent to fissures or gash fractures) developed parallel to the principal stress.

Normal faults. Normal faults appear to occur when the least principal stress (often actual tension) is horizontal and the largest principal stress (gravity) is approximately vertical. Subparallel normal faults with the down-

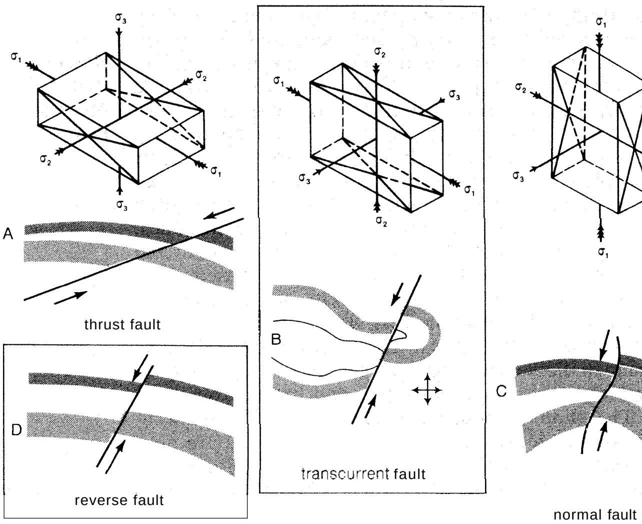


Figure 10: The three principal types of faults based on the orientations of the stress axes (σ_1, σ_2, and σ_3).

Adapted from E. H. Timothy Whitten, Structural Geology of Folded Rocks, © 1966 by Rand McNally and Company, Chicago, p. 133. Fig. 109.

thrown side facing the same direction give a steplike structure. Facing normal faults, with a downthrown block between them, produce a rift zone; rifts range from a metre or two in width to structures a few kilometres wide and several hundred kilometres long (*e.g.*, the East African Rift Valley). Horsts—upthrown blocks bounded by outward facing normal faults—rarely exceed a few kilometres. The dip of normal fault planes is usually greater than 45°, although they are sometimes steeper (about 60" to 80°) near the surface and shallower at depth; also, fault planes are usually less steep in soft, incompetent strata. Hanging wall and footwall (old mining terms) are names used to designate the blocks on either side of a normal fault.

Transcurrent faults. Transcurrent or tear faults produce structures of major significance in the Earth's crust. Their planes are characteristically almost vertical, so that the fault traces on a map are essentially straight. Horizontal compression apparently induces two intersecting fractures at an acute angle to the maximum compressive stress; friction between the rock masses can produce extensive slickensided surfaces, fault breccias (coarse fragmental rocks), mylonites (rocks produced by intensive crushing and shearing), or even ultramylonites along the fault planes. In many cases, one transcurrent fault direction is dominant, and the conjugate set is not obviously developed. Examples are the San Andreas Fault Zone, California (Figure 11), and a widespread set of northeast–southwest faults that cut the Scottish Grampian Highlands. In the latter area, a major fault and an associated breccia zone of great width resulted in erosion of the Great Glen (Inverness to Fort William); the net slip (movement) was some 96 kilometres (60 miles), but to the south numerous subparallel faults with smaller slips ranging up to seven or eight kilometres (four to five miles) have been mapped (*e.g.*, along mid-Strathspey). These Scottish faults are left lateral (or sinistral) because, looking across the trace of the fault, the opposite side moved to the left. The opposite movement sense is termed right lateral or dextral.

Scottish and midocean examples

John S. Shelton and Robert C. Frampton

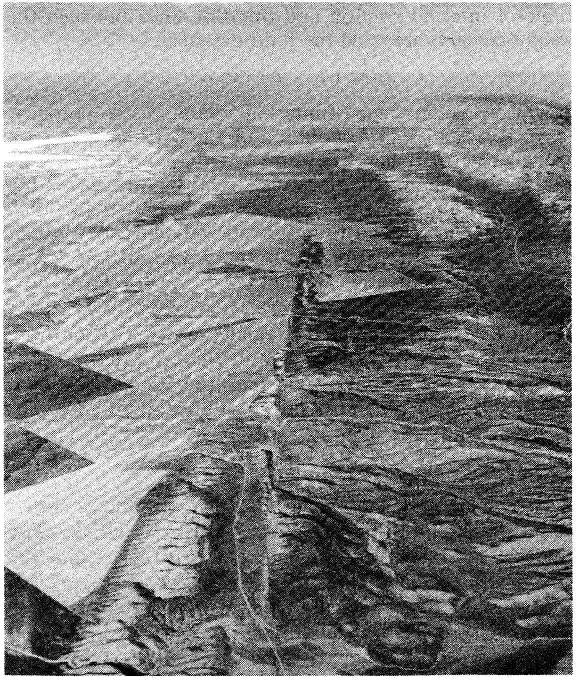


Figure 11: San Andreas Fault looking northwest from Elkhorn Hills. California.

Transcurrent faults in ocean basins are widespread and are responsible for repeated offset of midoceanic ridges (*e.g.*, Mid-Atlantic Ridge, which has the dimensions of a major mountain range). Many of these are thought to be transform faults; that is, the transcurrent movement terminates sharply where it is transformed into a structure of another type. The hypothesis involves unusual geometry because these transcurrent faults transect midoceanic ridges interpreted to be zones of active oceanic growth and outward spreading (see further OCEANIC RIDGES; SEA-FLOOR SPREADING).

Thrust faults. According to the model referred to above, thrust faults involve horizontal maximum compression and near-vertical least compression. The model involves a fault plane inclined less than 45°. This widely quoted model is too simple to explain fully a large proportion of actual thrusts, or the locally important although relatively small (less than a few hundred metres) reverse faults with dips greater than 45°. Some horsts are upthrust blocks between two reverse faults. Most thrusts tend to be steep near the surface but flatten to a nearly horizontal attitude at depth and over most of their extent. Many authorities distinguish overthrusts as thrust faults with an initial dip of less than 10° and a net slip measured in kilometres. The first overthrusts were identified in the 19th century in Switzerland and near Dresden, Germany.

In zones of apparently strong compression within the Appalachian fold belt (eastern United States and Canada) many geologists still support the classical concept that folding occurred where competent rock units overlie weaker shales and other incompetent rocks, whereas thrusting occurred where incompetent members were underlain by competent units (see Figure 12). Thrusts are

By courtesy of the BP Petroleum Development Limited. London

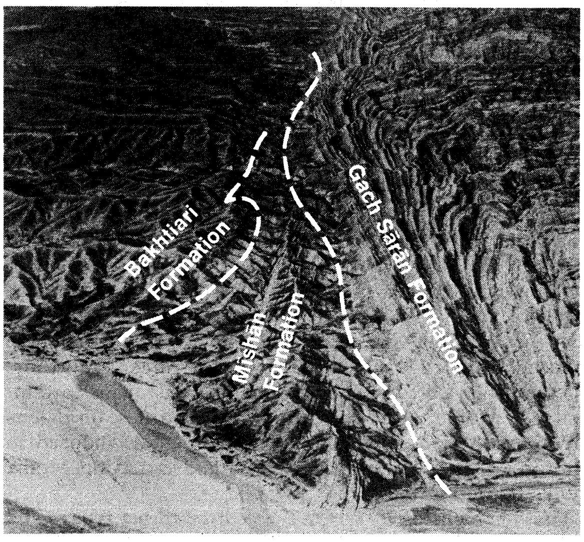


Figure 12: The major Ram Hormuz thrust fault, with steep dip at outcrop, near Māmātīn, Khuzestan, Iran (aerial view looking northwest). The gypsiferous Gach Sārān (Lower Fars) Formation has been thrust from the right. Gently dipping Bakhtiari conglomerates unconformably overlie the folded and thrust formations.

extremely important structural elements of major geosynclinal or mobile belts. In Alberta and Kentucky, for example, oil wells have frequently penetrated the same horizon three or four times where several thrust slices overlie each other. Thrust faults have frequently been considered to be convincing evidence of crustal shortening in tectonic zones of compression; many authorities now reject this hypothesis, which, in some mobile belts, would involve incredible distances (several hundred kilometres) of crustal shortening. In addition, many of these thrust slices are only a few tens of metres thick but several kilometres long (parallel to the movement direction), and it is difficult to conceive that brittle fracture, followed by transmission of stress through such a lithic unit, could move it up a thrust plane--especially if the rocks involved are relatively incompetent. There is considerable difference of opinion about whether thrusts continue into the local basement or whether they are restricted to the "thin skin" above the basement. Several dissimilar types of major thrusts exist in the Earth's crust, although currently textbooks tend to emphasize one type to the virtual exclusion of others. For illustration, three tectonically dissimilar types are discussed here, with examples of each type:

Types of major thrust faults

(1) Marginal to many deformed mobile belts (foothill belts), oil wells commonly penetrate numerous thrusts and then one major sole fault (main fault surface) before entering essentially undeformed strata. The basement is apparently not involved; the rocks are essentially unmetamorphosed and contain few, if any, small folds. Examples were exhaustively studied during petroleum search in the Rocky Mountain foothills of Alberta. In the European Alps and Carpathians, thin thrust-bounded slices are termed nappes (or decken); many nappes comprise gigantic recumbent isoclines (folds with near-horizontal limbs and bisecting planes). Such structures involve décollment, meaning that the sedimentary rocks become detached from and thrust across underlying, undeformed basement. The surfaces between nappes resemble the "slides" in the Scottish Grampian Highlands; subsequent to development of these slides, the whole sequence was regionally metamorphosed and concomitantly subjected to several superposed phases of folding.

(2) Some thrusts cut the basement, causing introduction of crystalline rock slices into overlying sedimentary sequences. The Scottish Northern Highlands provide an excellent example. Here, Lewisian gneisses crystallized about 1,600,000,000 years ago; following profound erosion, these basement rocks were overlain by some six to seven kilometres of Moine shales and sandstones about 900,000,000 years ago. The earliest recognized folding of the Moine rocks involved intimate isoclinal interfolding with the Lewisian basement (as at Glenelg, east of Skye). Current interpretation of the region farther northeast (*e.g.,* Scardroy) is that several thrusts brought basement slices up into the Moine sequence during the initial deformation (prior to intense folding and metamorphism of the Moine rocks). Such thrusts possibly characterize a more central zone of mobile belts than does the preceding type.

(3) The Scottish Northern Highlands yield a classic example of another thrust type. Several folding and metamorphic events (more than 570,000,000 years ago) affected the Moine rocks before deposition of the Cambrian and Ordovician quartzitic and dolomitic sediments. The Moine Thrust Zone (excellently exposed in the Assynt area of northwestern Scotland) shows unmetamorphosed Cambro-Ordovician rocks overridden by Moine metamorphosed rocks. Several major thrust planes are involved. Moine rocks were thrust westward several tens of kilometres. Local but intense folding was induced in the underlying Cambro-Ordovician and overlying Moine rocks bordering the thrust surfaces.

The mechanisms by which these three thrust types (and other possible tectonically distinct types) were generated are not fully understood, although two hypotheses that help to avoid the need for extreme crustal shortening in mobile belts have gained much support in recent years.

Possible thrust mechanisms

(1) The voids between sediment grains are full of water at deposition; continued sedimentation causes loading and compaction of underlying sediment and, not infrequently, the pore-water pressure builds up more rapidly than water can be expelled. Water pressure above the hydrostatic pressure forces grains apart, reducing internal friction and viscosity of the sediment and facilitating gravitational flow down gently inclined slopes. Local, or general, vertical tectonism during development of geosynclinal mobile belts apparently causes gravity gliding of recently deposited sediments out onto adjacent zones. Continuing rise of central areas (more rapidly than the flanks) would promote lateral spreading off raised welts toward the bounding craton. This activity produces nappes and recumbent folds of semiconsolidated sedimentary materials, together with thrusts at the junction of the deeper-water and shallow-water zones of sediment accumulation, (respectively, the eugeosynclinal and miogeosynclinal belts). The gliding would laterally compress the flanking portions of the geosynclinal belt. Such concepts are now widespread, and Russian authors have championed the idea that vertical movements, rather than crustal shortening and crustal compression, are the dominant cause of folding and faulting.

A few geologists place emphasis on syntaphral tectonics, which involve basinward gravity sliding. Slumping of unconsolidated wet sediments and penecontemporaneous folding are examples, although much larger examples appear to develop on slopes of only a degree or two. It is argued that troughs in developing geosynclines receive creeping masses, compressive foldings, nappelike sheets, and turbidity flows separated by "thrust" surfaces (that could be preserved and misidentified as early orogenic phenomena). A sheet one kilometre *(0.6* mile) thick and 10,000 square kilometres (4,000 square miles) in area travelled many kilometres on the Grand Banks off eastern North America; study of this sheet suggested that a significant part of the deformation of Alpine-type geosynclines occurred while sedimentation was continuing and before tectonic forces began to uplift the geosyncline. Evidence of extensive troughward gravity sliding in the central Appalachian fold belt also is germane; the sliding occurred in the form of widespread wedges—small thrusts (one to four metres) on which one thinly bedded sedimentary member rode up and over itself.

(2) Another mechanism invoked is that of high water pressures to reduce frictional resistance and to facilitate movement on major thrusts. The weight per unit area, S, of a block is supported by a solid stress σ plus the interstitial-fluid pressure p. The critical shear stress, τ, required to slide a block is then the product of the tangent of the angle of internal friction and the difference between the weight per unit area and the fluid pressure:

$$\tau = (S - p) \tan \phi$$

in which ϕ is the angle of internal friction; as p approaches S, corresponding to flotation of the overburden, τ approaches zero. It has been claimed that, under tectonic compression, if maximum stress is horizontal, p readily equals S. Fluid pressures approximately equal to 0.9 S and greater have been found in oil wells in geosynclinal basins and in areas probably now tectonically compressed. Such pressures permit formation of large, low angle overthrusts. For the western Wyoming overthmst belt, it has been calculated that, if the geosyncline limb sloped at 2.5" to 3°, a fluid pressure–overburden ratio of 0.91 to 0.92 would have been required for gravitational sliding, which there must have totalled about 80 kilometres (50 miles).

STRUCTURAL INTERFERENCE

Events of several ages generally have produced superposed strains in all but the most recent rock bodies. The kinds of geological phenomena that may occur and the time spans associated with them are shown in Figure 13. Superposition of tectonically induced strains makes analysis of the individual events difficult.

Superposed-fold geometry within metamorphosed mobile belts has been studied intensively for 15 years. Folding and metamorphic events within a belt tend to be

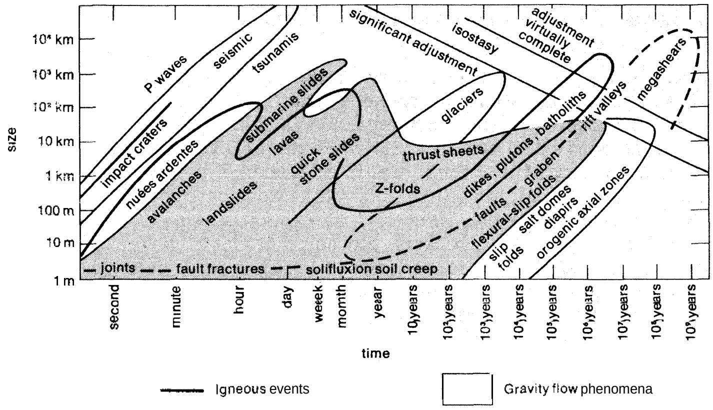


Figure 13: A size-time model for geotectonic phenomena, showing the kinds of events that may occur and their associated sizes and time spans.

From W.S. Carey, "Scale of Geotectonic Phenomena," vol. 3 (1962); *Journal of the Geological Society of India*

Superposed folding events

spasmodic in space and time. Remarkably similar but not synchronous events may occur at locations parallel to a geosynclinal axis. In the Caledonian mobile belt of Scandinavia, Scotland, and Ireland or in the Appalachians of eastern North America, major nappes apparently formed during initial deformation prior to major metamorphism. Without detailed lithologic mapping over hundreds of square kilometres, such nappes are difficult to detect. Two or three subsequent phases of slip, flexural-slip, or flexure folding (interspersed with metamorphic episodes) produced the folds that are readily seen in the field. Dating methods used in the Caledonides have shown that despite the similarity of the folding events in the Dalradian rocks of western Ireland and of the eastern Grampian Highlands (Scotland), they were not synchronous. During the waning deformation, mobile belt rocks commonly become less plastic and small, angular, kink folds are commonly superposed on earlier folds.

Compression of planar beds commonly produces cylindrical or conical folds with subparallel axes, although the axial orientation depends on the original bedding orientation with respect to the compressive stress. Superposed refolding does not produce a single new fold axis, but a differently oriented fold axis for each differently oriented portion of the original fold. Succeeding fold systems thus tend to produce smaller folds on the flanks of earlier structures; exceptions occur if all old structures are obliterated by transposition and a new schistosity is produced and folded, or if the new folding is of a much larger wavelength (*e.g.*, gentle regional warping of a range). The theoretical geometry of superposed slip and flexural-slip folding has been described, but unravelling superposed folds in the field is very complicated. Several hundred papers describe field relationships in different parts of the world, but few give unequivocal geometrical solutions. Current methods require analysis of thousands of field measurements. Complications stem from: (1) inadequate three-dimensional exposure, (2) the fact that although superposed slip and flexural-slip folding produce dissimilar geometries it is often uncertain which was involved or whether a hybrid of both mechanisms occurred, and (3) factors like flattening that may have significantly distorted the theoretical strain models. It has proved effective for deciphering complex superposed fold geometry (when enough data are available) to subdivide the region into smaller, homogeneous domains in which specified linear structures are approximately parallel. This identifies domains in which bedding (or schistosity) was planar after the first folding. Extrapolation to adjacent areas would be important for prospecting (*e.g.*, Precambrian iron ores, Quebec), but currently it is extremely difficult. Where two simple and open fold phases are superposed, structural culminations (domes) at antiformal intersections and depressions (basins) at synformal intersections commonly occur. Because the domes and basins tend to be arranged on a checkerboard pattern some predictions are possible. This model was used successfully to predict another cratonic basin of gold-bearing Witwatersrand rocks east of the known subsurface goldfields near Johannesburg, South Africa; drilling proved the existence of the predicted basins.

Superposed joints and faults

Relatively little work has been done on fracture sets in areas of superposed folding, although joints specifically related to superposed folds are usual. Superposed joint and fault sets are common even where folding has not been induced, but the regional significance of superposed fracture systems has been evaluated in only a few cases. Renewed stress may induce a new joint set or accentuate or cause movement along old joints. Analysis of the offsets, mineralization, or dike-swarm intrusion along certain joints can sometimes produce a chronology. Central Scotland provides an excellent example. During the waning phases of the Caledonian orogeny (episode of mountain building), a dike swarm intruded northeast–southwest trending fissures; in late Paleozoic time (about 275,000,000 years ago), transcurrent faults (*e.g.*, Great Glen Fault) cut the Central Highlands. The latter probably reflect north–south compression (Hercynian orogeny?), and release of this stress may have accounted for intrusion of the east–west dike swarm. Later, the Tertiary basaltic dike swarms occupied new northwest–southeast fractures, although across parts of the Donegal Granite (northwestern Ireland) Tertiary dikes alternately occupy short dilated sections of the old north–south and east–west joints to maintain the regional northwest–southeast trend.

Currently, great interest centres on global and plate tectonics (the concept that the outer part of the Earth consists of a small number of great plates that shift about and abut each other), subjects that received impetus from the implications of extensive continental drift and concomitant growth of the ocean basins by upward and outward flow along the midoceanic ridges. These hypotheses are not universally accepted, although their advocates marshal1 impressive arrays of supporting evidence and have begun to develop an integrated picture of crustal development and evolution. On the continents, it is increasingly clear that many old structures continue to control structural development of younger features. Major basement fractures may remain active and significantly affect younger sedimentation. Evolution of cratonic basins and domes of southern Africa and North America supports the hypothesis that basement fold structures may control the pattern of cratonic basins and sedimentation during all succeeding periods, unless they become involved in mobile belt geosynclinal activity.

BIBLIOGRAPHY. Recent structural geology textbooks that broadly cover the topics treated in this article include: P.C. BADGLEY, *Structural and Tectonic Principles* (1965); M.P. BILLINGS, *Structural Geology*, 2nd ed. (1954); N.J. PRICE, *Fault and Joint Development in Brittle and Semi-Brittle Rock* (1966); H. RAMBERG, *Gravity, Deformation and the Earth's Crust; As Studied by Centrifuged Models* (1967); J.G. RAMSAY, *Folding and Fracturing of Rocks* (1967); L.U. DE SITTER, *Structural Geology*, 2nd ed. (1964); F.J. TURNER and L.E. WEISS, *Structural Geology of Metamorphic Tectonites* (1963); arid E.H.T. WHITTEN, *Structural Geology of Folded Rocks* (1966).

Translations of Russian and French works of importance include: V.V. BELOUSSOV, *Basic Problems in Geotectonics* (1962; based on the Russian ed. of 1954); and edited with A.A. SORSKII, *Folded Deformations in the Earth's Crust: Their Types and Origin* (1965; orig. pub, in Russian, 1962); and J. GOGUEL, *Traité de tectonique* (1952; Eng. trans., *Tectonics*, 1962).

Books containing modern approaches to regional structural studies are: M.R.W. JOHNSON and F.H. STEWART (eds.), *The British Caledonides* (1963); M. KAY (ed.), *North Atlantic: Geology and Continental Drift, a Symposium* (1969); E-AN ZEN *et al.* (eds.), *Studies of Appalachian Geology: Northern and Maritime* (1968); J. RODGERS, *The Tectonics of the Appalachians* (1970); and G.W. FISHER *et al.* (eds.), *Studies of Appalachian Geology: Central and Southern* (1970).

For a summary of computer applications to structural problems, see E.H.T. WHITTEN, "Trends in Computer Applications in Structural Geology," in D.F. MERRIAM (ed.), *Computer Applications in the Earth Sciences* (1969); and for a glossary of terminology, see J.G. DENNIS (ed.), *International Tectonic Dictionary, English Terminology* (1967).

(E.H.T.W.)

Rockets and Missile Systems

Rocket is the generic term used broadly to describe a variety of jet-propelled missiles, research vehicles, thrust devices, fireworks, and space-launch vehicles. Forward motion results from reaction to the rearward ejection of matter, usually hot gases, at high velocity. The propulsive jet of gases usually consists of the combustion products of solid or liquid propellants. Stored, high-pressure cold gas, heated hydrogen gas, or ions can also be used.

Rocket propulsion is a unique member of the family of jet-propulsion engines that includes turbojet, pulse-jet, and ramjet systems. The rocket engine is different, however, in that the elements of the propulsive jet (fuel and oxidizer) are self-contained within the vehicle. The thrust produced is independent of the medium through which the vehicle travels. Other kinds of jet-propulsion engines carry only their fuel and depend on the oxygen content of the air for burning. Thus these varieties of jet engines are called air breathing and are limited to operation within the Earth's atmosphere. The upper limit of travel for air-breathing jet engines is about 90,000 feet (27 kilometres). A rocket engine is necessary for flight beyond the atmosphere into the immense reaches of space.

A guided missile is broadly any military missile that is capable of being guided or directed to a target after having been launched. Modern guided missiles are powered by some type of jet propulsion, usually rocket propulsion. There are numerous kinds of guidance systems as well as range and functions of such missiles. Accuracy is of prime importance. Explosive warheads may be high explosive or nuclear.

Rocket power has been used in crude form for hundreds of years for military purposes, signalling, and fireworks displays. During the first half of the 19th century, several European armies had rocket brigades, but interest waned with improvement in accuracy and range of artillery. Guided missiles had their origin during World War I but were not then developed to operational use. During World War II, however, rocket power and guided missiles were extensively developed, primarily by Germany. Since then, most technologically capable nations have developed rocket-propelled guided missiles for military purposes and for space research.

Nearly a century ago it was recognized that rocket power was the key to exploration of space beyond the Earth's atmosphere. Technology, however, did not reach the point at which such investigations were possible until mid-20th century. In 1957 man opened a new era of exploration by applying the laws of celestial mechanics through the use of rocket power.

This article is outlined as follows:

I. Development of rockets and guided missiles

EARLY HISTORY

There is no reliable early history of the "invention" of rockets. Most historians of rocketry trace the development of rockets to China, a land noted in ancient times for its fireworks displays. In 1232, when the Mongols laid siege to the city of K'ai-feng, capital of Honan Province, the Chinese defenders used weapons that were described as "arrows of flying fire." There is no explicit statement that these arrows were rockets, but some students have concluded that they were because the record does not mention bows or other means of shooting the arrows. In the same battle, it is reported, the defenders dropped from the walls of the city a kind of bomb described as "heaven-shaking thunder." From these meagre references some students have concluded that by 1232 the Chinese had discovered black powder (gunpowder) and had learned to use it to make explosive bombs as well as propulsive charges for rockets. Drawings made in military documents much later show powder rockets tied to arrows and spears. The propulsive jet evidently added to the range of these weapons and acted as an incendiary agent against targets.

In the same century rockets appeared in Europe. There is indication that their first use was by the Mongols in the Battle of Legnica in 1241. The Arabs are reported to have used rockets on the Iberian Peninsula in 1249; and in 1288 Valencia was attacked by rockets. In Italy, rockets are said to have been used by the Paduans (1379) and by the Venetians (1380).

Nature of early rockets

There are no details of construction of these rockets, but it is presumed that they were quite crude. The tubular rocket cases were probably many layers of tightly wrapped paper, coated with shellac. The propulsive charge was basically a mixture of finely ground carbon (charcoal), saltpetre (potassium nitrate), and sulfur. The English scientist Roger Bacon wrote formulas for black powder about 1248 in his Epistola. In Germany, a contemporary of Bacon, Albertus Magnus, described powder charge formulas for rockets in his book *De mirabilibus mundi.* The first firearms appeared about 1325; they utilized a closed tube and black powder (now referred to as gunpowder) to propel a ball, somewhat erratically, over varying distances. Military engineers then began to invent and refine designs for guns and rockets on a parallel basis.

The French historian Jean Froissart suggested that firing rockets from tubes would give them better direction. An Italian writer conceived a number of novel weapons based on rocket propulsion, including a rocket-driven car, designed to breach walls or gates, and a naval torpedo, designed to skim across water and ram its spiked nose into ships. Many other ideas were suggested in print, such as rockets with parachutes and underwater explosive rockets. The extent to which many of these designs were reduced to working models or weapons is not known. By this time rockets were used also for signalling and, especially by pirates, for setting fire to the tarred rigging of sailing ships. They had many names, such as flying fire or wild fire. There is reason to believe that some of these devices were hurled or catapulted incendiaries rather than actual rockets, particularly in night attacks.

Of considerable interest are the historical origins of rocket designs that were to be employed many years later: the staged (or step) rocket, the clustered rocket, and the winged rocket. These designs evolved both from applications and requirements of fireworks and their adoption by ordnance masters. The earliest records of these concepts are contained in a manuscript by Conrad Haas, an artillery officer and chief of the arsenal at Sibiu, Romania, in the 16th century. In 1590 a German fireworks expert published a small illustrated book reproducing Haas's designs. A Polish artillery expert published essentially the same designs (see Figure 1) in a book translated into French, German, English, and Dutch, but practically no use was made of these concepts for clustered, winged, and step rockets for several hundreds of years, except perhaps in fireworks displays.

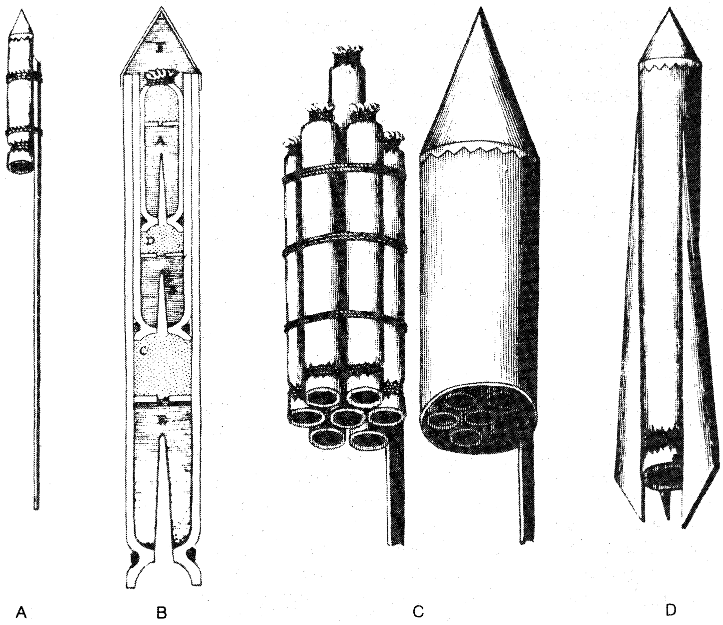


Figure 1: Historical examples of rockets.
(A) Conventional stick-guided war rocket or fireworks skyrocket; some early design concepts of (B) a step rocket, (C) a clustered rocket, and (D) a fin-stabilized glide rocket. Drawings by Casirnirus Siemienowicz from *Artis* magnae *artillerae*, pars prima, 1650.
By courtesy of M. Subotowicz (Poland)

By 1668 military rockets had increased in size and performance. In that year, a German colonel designed a rocket weighing 132 pounds (60 kilograms); it was constructed of wood and wrapped in glue-soaked sailcloth. It carried a gunpowder charge weighing 16 pounds (seven kilograms). Nevertheless, the use of rockets seems to have waned, and for the next 100 years their employment in military campaigns appears to have been sporadic.

Hyder Ali's metal-clad rockets

A revival commenced late in the 18th century in India. There Hyder Ali, prince of Mysore, developed war rockets with an important change: the use of metal cylinders to contain the combustion powder. Although the hammered soft iron he used was crude, the bursting strength of the container of black powder was much higher. Thus a greater internal pressure was possible, with a resultant greater thrust of the propulsive jet. The rocket body was lashed with leather thongs to a long bamboo stick. Range was perhaps up to three-quarters of a mile (more than a kilometre). Although individually these rockets were not accurate, dispersion error became less important when large numbers were fired rapidly in mass attacks. They were particularly effective against cavalry and were hurled into the air, after lighting, or skimmed along the hard dry ground. Hyder Ali's son, Tippu Sultan, continued to develop and expand the use of rocket weapons, reportedly increasing the number of rocket troops from 1,200 to a corps of 5,000. In battles at Seringapatam in 1792 and 1799 these rockets were used with considerable effect against the British.

19TH-CENTURY DEVELOPMENTS

The news of the successful use of rockets spread through Europe. In England, William Congreve began to experiment privately. The fact that his father, of the same name, was comptroller of the Royal Arsenal at Woolwich undoubtedly facilitated his efforts.

Several important improvements in rockets were made by Congreve. First, he experimented with a number of black-powder formulas and set down standard specifications of composition. He also standardized construction details and used improved production techniques. Also, his designs made it possible to choose either an explosive (ball charge) or incendiary warhead. The explosive warhead was separately ignited and could be timed by trimming the fuse length before launching. Thus, air bursts of the warheads were feasible at different ranges.

Congreve's metal rocket bodies were equipped on one side with two or three thin metal loops into which a long guided stick was inserted and crimped firm. Weights of eight different sizes of these rockets ranged up to 60 pounds (27 kilograms). Launching was from collapsible A-frame ladders. In addition to aerial bombardment, Congreve's rockets were often fired horizontally along the ground.

These side-stick-mounted rockets were employed in a successful naval bombardment of the French coastal city of Boulogne in 1806. The next year a massed attack, using hundreds of rockets, burned most of Copenhagen to the ground. In the Battle of Leipzig and the siege of Danzig, which led to the surrender of that city, rockets played a significant role.

Rockets in the War of 1812

During the War of 1812 between the United States and the British, rockets were employed on numerous occasions. The two best known engagements occurred in 1814. At the Battle of Bladensburg (August 24) the use of rockets assisted British forces to turn the flank of the American troops defending Washington, D.C. As a result, the British were able to capture the city. In September, the British forces attempted to capture Ft. McHenry, which guarded Baltimore harbour. Rockets were fired from a specially designed ship, "Erebus," and from small boats. The British were unsuccessful in their bombardment, but on that occasion Francis Scott Key, inspired by the sight of the night engagement, wrote "The Star Spangled Banner," later adopted as the United States national anthem. "The rockets' red glare" has continued to memorialize Congreve's rockets ever since.

In 1815 Congreve further improved his designs by mounting his guide stick along the central axis. The rocket's propulsive jet issued through five equally spaced holes rather than a single orifice. The forward portion of the guide stick, which screwed into the rocket, was sheathed with brass to prevent burning. The centre-stick-mounted rockets were significantly more accurate. Also, their design permitted launching from thin copper tubes.

Maximum ranges of Congreve rockets were from one-half mile to two miles, depending upon size. They were competitive in performance and cost with the ponderous ten-inch mortar and were vastly more mobile.

Congreve wrote a brief but classic work, ***The Details** of the Rocket System* (1814), detailing the deployment and use of his rockets. Recognizing the importance of massed fire for maximum effect, he recommended the simultaneous firing of 50 rockets at a time, and never less than 20. Volleys could be fired as rapidly as every 30 seconds.

The use of Congreve rockets spread rapidly through Europe. Rockets based on Congreve designs were developed in France, Denmark, Spain, Italy, Switzerland, Sweden, Austria, and Russia. One interesting rocket design of the Swedish Rocket Corps was a stickless, delta-winged glide missile that was fired from a hand-held launcher.

The next significant development in rocketry occurred about the middle of the 19th century. William Hale, a British engineer, invented a method of successfully eliminating the deadweight of the flight-stabilizing guide stick. By designing jet vents at an angle, he was able to spin the rocket. He developed various designs, including curved vanes that were acted upon by the rocket jet. These rockets, stabilized by means of spin, represented a major improvement in performance and ease of handling.

Manufacturing techniques and materials had improved, too. The use of improved steel and rivetted rocket bodies permitted pressures as high as 23,000 pounds per square inch.

Decline in use of rockets

Even the new rockets, however, could not compete with the greatly improved artillery with rifled bores. The rocket corps of most European armies were dissolved, though rockets were still used in swampy or mountainous areas that were difficult for the much heavier mortars and guns. The Austrian Rocket Corps, using Hale rockets, won a number of engagements in mountainous terrain in Hungary and Italy. Other successful uses were by the Dutch colonial services in Celebes and by Russia in a number of engagements in the Turkistan War.

Hale sold his patent rights to the United States in time for some 2,000 rockets to be made for the Mexican War. 1846–48. Although some were fired, they were not particularly successful. The U.S. Ordnance Manual of 1862 lists 16-pound Hale rockets with a range of 1.25 miles (two kilometres).

Rockets were used in a limited way in the American Civil War (1861–65), but reports are fragmentary, and apparently they were not decisive.

During the 19th century two important peacetime applications of the war rocket developed: the lifesaving rocket and the whaling rocket.

In the era of sailing ships many lives were lost in the grounding and breakup of ships a short distance offshore, particularly during storms. The idea of using a rocket to carry from ship to shore or shore to ship a light line, which then could haul a heavier lifesaving line to aid passengers and crew to reach shore, was developed fiist by Henry Trengrouse of Cornwall, in England. In 1807, after watching the loss of 100 crewmen in a shipwreck near the coast, he built a series of line-carrying rockets. The first lives were saved by a line-carrying rocket in 1832. John Dennett modified a small Congreve rocket and in that year was able to save 19 persons from the "Bainbridge" aground on Atherfield Rocks. By mid-century numerous other designs had appeared (see Figure 2).

By courtesy of Mitchell R Sharp

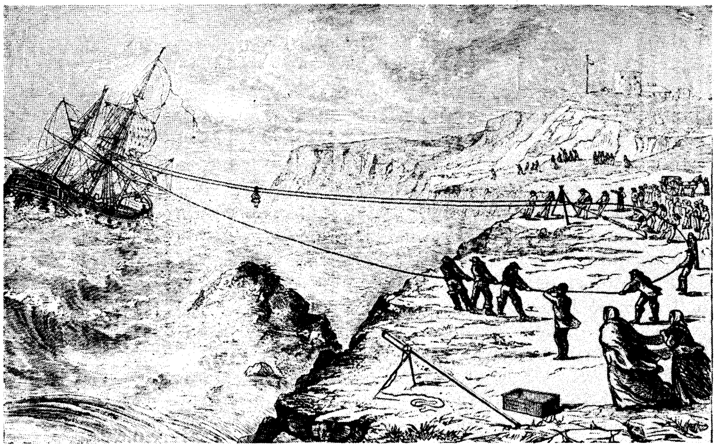


Figure 2: Rescue operation using lifesaving rocket, from a drawing of the 1870s. The light line carried to the ship by the rocket (centre foreground) has been replaced by a heavier line, and a breeches-buoy transfer is in operation.

Lifesaving rockets spread to maritime nations and to the United States. Records kept of their use in Great Britain alone list at least 15,000 lives saved between 1871 and 1962.

As steam and diesel power replaced sails, the need for lifesaving rockets declined. In the late 1960s, however, The Netherlands still maintained coastal stations from which special trucks could rush line-throwing rockets to the nearest point on shore to aid a grounded ship.

Rocket-propelled harpoons

Rocket propulsion was first applied to whaling harpoons in 1821. The most successful concept was that of Capt. Thomas W. Roys, of the United States, whose design of a rocket harpoon carried an explosive warhead. The harpoon was fired through a tube launcher from a small boat. If the aim, burning time of the rocket, speed of flight, and distance travelled were correct, the explosive bomb killed the whale quickly and affixed a toggle and rope in the carcass. These devices enjoyed only a brief lifetime, however.

Both the United States Army and the British Navy conducted experiments with rocket-propelled torpedoes in the period 1860–80, and several ingenious devices were designed. Nevertheless, none of this work led to a real weapon. One of the fundamental problems of the rocket torpedo, as pointed out by William Hale, was that as the powder in the rocket motor was consumed, the torpedo grew lighter, making it difficult to hold the torpedo at constant level below the surface.

Meanwhile, there were individual enthusiasts and inventors in nearly every country. Largely unknown, and sometimes considered a dangerous nuisance, they achieved varied success and recognition. In the early 19th century Claude Ruggieri, a prominent Italian fireworks maker, staged a number of shots in Paris in which rats and mice were sent aloft by rockets and returned by parachute. It is reported that Ruggieri even planned to send up a small boy, using a rocket cluster, but the police intervened.

In 1881 a Russian explosives maker, Nikolay Ivanovich Kibalchich, imprisoned for his part in an assassination attempt on Tsar Alexander II, conceived a rocket airplane that functioned by successive explosions of compressed powder candles. Kibalchich was executed and his writings remained in prison archives until the Revolution in 1917.

A few years later a German inventor, Herman Ganswindt, conceived of an intermittent-firing propulsion system similar to that of Kibalchich but employing steel cartridges loaded with dynamite. Ganswindt went further. He wanted to give his vehicle sufficient speed to attain escape velocity; *i.e.*, to leave the Earth. Apparently, Ganswindt was the first to connect rocket propulsion with space flight. But in Russia in 1895 a young mathematics teacher published his first article on space travel. Konstantin Eduardovich Tsiolkovsky was among the first to grasp the importance of exhaust velocity and the reason that rockets had been limited by black-powder formulas. He saw that by using liquid propellants (*e.g.*, liquefied hydrogen and oxygen) much greater efficiencies would result. Tsiolkovsky made important contributions to the theory of space vehicle design, including the concept of a closed biological cycle, utilizing plant life to produce oxygen on long voyages. The fact that he wrote in Russian, combined with his retiring nature, left him scarcely known abroad for many years.

Importance of exhaust velocity

20TH CENTURY TO WORLD WAR II

In Sweden, about the turn of the century, Wilhelm Unge invented a device described as an "aerial torpedo." Based upon the stickless Hale rocket, it incorporated a number of design improvements. One of these was a rocket motor nozzle that caused the gas flow to converge and then diverge. Another was the use of smokeless powder based on nitroglycerin. Unge believed that his aerial torpedoes would be valuable as surface-to-air weapons against dirigibles. Velocity and range were increased, and about 1909 the Krupp armament firm of Germany purchased the patents and a number of rockets for further experimentation.

In the United States, meanwhile, Robert Hutchings Goddard was conducting theoretical and experimental research on rocket motors at Worcester, Massachusetts. Utilizing a steel motor with a tapered nozzle, he achieved greatly improved thrust and efficiency. Another of Goddard's concepts was a high-altitude research rocket whose motor was fired in pulses. Somewhat similar to the pulsed-momentum principle suggested by Kibalchich and Ganswindt, the impulses were to have derived from charges of solid fuel injected into the combustion chamber in rapid succession. In 1916 Goddard approached the Smithsonian Institution for financial support. It was forthcoming in a few months.

During World War I Goddard developed a number of designs of small military rockets to he launched from a lightweight hand launcher. By switching from black powder to double-base powder (40 percent nitroglycerin, 60 percent nitrocellulose), a far more potent propulsion charge was obtained. These rockets were proving successful under tests by the United States Army when the Armistice was signed; they became the forerunners of the bazooka of World War II. Goddard's main interest, however, was in utilizing the new potential of rockets to reach high altitudes. His notebooks ultimately revealed such imaginative concepts as a circumlunar rocket to carry a camera to photograph the far side of the Moon, ion and nuclear rocket propulsion, and manned and unmanned interplanetary exploration.

Backed by the Smithsonian Institution, Goddard switched from solid to liquid propellants and launched

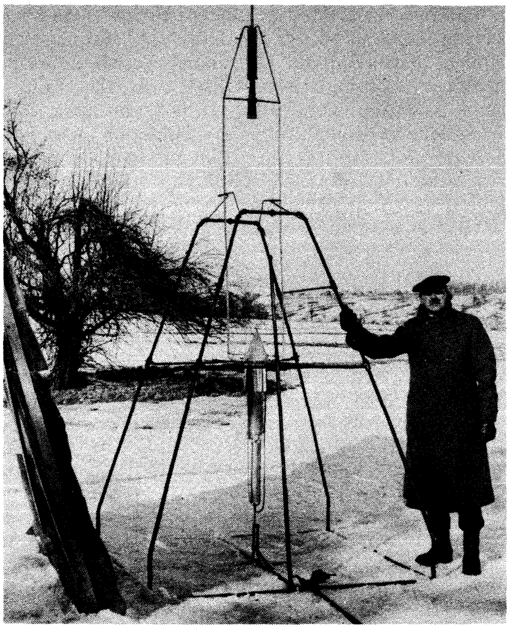


Figure 3: R.H. Goddard, U.S. rocket pioneer, with the liquid-propellant rocket he designed and launched in 1926.
By courtesy of Mrs. Robert H. Goddard

the first liquid-propellant rocket (liquid oxygen and gasoline) on March 16, 1926 (see Figure 3). It reached an altitude of only 41 feet (12 metres) and landed only 184 feet (56 metres) away, but the significance was as great as the few feet flown by the Wright brothers at Kitty Hawk.

Picking up the pieces, Goddard modified the motor nozzle length, increased the throat diameter, strengthened the launching stand, and flew the apparatus again on April 3. No public report of this success was made.

Later, supported by Clark University and Guggenheim Foundation funds, Goddard developed further research rockets in Roswell, New Mexico. Although he never achieved the results he knew were possible, Goddard was a brilliant inventor. His later designs utilized both turbine-driven pumps and gyroscopic stabilizers. Vanes were used to deflect the rocket exhaust to correct deviations from the planned flight path.

In his paper "A Method of Reaching Extreme Altitudes" Goddard discussed the possibility of a rocket reaching the Moon with a payload of flash powder to signal its arrival to astronomers. The significance of this work, like that of Tsiolkovsky's, was that the suggestions were made not by an exuberant inventor or enthusiast but by a thoughtful scientist with a sound academic background.

World War I actually saw little use of rocket weapons, despite successful French incendiary antiballoon rockets and a German trench-war technique by which a grappling hook was thrown over enemy barbed wire by a rocket with a line attached. Many researchers besides Goddard used the wartime interest in rockets to push experimentation, the most noteworthy being Elmer Sperry and his son, Lawrence, in the United States. The Sperrys worked on a concept of an "aerial torpedo," a pilotless airplane, carrying an explosive charge, that would utilize gyroscopic, automatic control to fly to a preselected target. Numerous flight attempts were made in 1917, some successful. Because of interest in early military use, the U.S. Army Signal Corps organized a separate program under C.F. Kettering in Ohio late in 1918. Seventy-five aircraft were ordered built just before World War I ended.

Features of the Kettering design

The Kettering design used a gyroscope for lateral control to a preset direction and an aneroid barometer for pitch (fore and aft) control to maintain a preset altitude. A high angle of dihedral (upward tilt) in the biplane wings provided stability about the roll axis. The aircraft was rail-launched. Distance to target was determined by the number of revolutions of a propeller. When the predetermined number of revolutions had occurred, the wings of the airplane were dropped off and the aircraft carrying the bomb load dropped on the target.

The limited time available to attack the formidable design problems of these systems doomed the programs, and they never became operational.

In Britain, even earlier, A.M. Low conceived of a similar project, but two tests made in 1917 had little success. Following World War I the British continued a small development program of radio-controlled seaplanes, with known flights being made in 1927 and 1930. In the United States the Sperry automatic control system was utilized in work from 1920 to 1926.

In 1923 an obscure German mathematics teacher, Hermann Oberth, published ***Die Rakete zu den Planetenräumen*** ("The Rocket into Interplanetary Space"). In this thin pamphlet was set forth, with a grasp remarkable for his time, the potentialities of rockets to achieve great velocity and to provide the means for manned space exploration. In ***Wege zur Raumschiffahrt*** ("Way to Space Travel"), in 1929, Oberth not only set forth design concepts for immense interplanetary space vehicles utilizing clustered liquid-propellant motors but also included a chapter on electric propulsion and the ion rocket (see below ***Electrical propulsion systems)*** predating actual development work on electrostatic propulsion by 30 years.

The decade that followed was an exciting one. Germans such as Walter Hohmann and the Austrians Baron Guido von Pirquet and Hermann Noordung published technical studies on rocket power and space vehicles. In France, the famous aviator and test pilot Robert Esnault-Pelterie lectured and wrote on high-altitude rockets and interplanetary flight. It was Esnault-Pelterie who first used the term astronautics. In 1929 Esnault-Pelterie and the French banker André Hirsch established an annual astronautics award for the experimenter who had done most to further space flight. In the Soviet Union, researchers were interpreting and expanding Tsiolkovsky's work. In the period 1927–33 rocket and space flight societies were formed in Germany, Austria, the Soviet Union, the United States, and Great Britain. These groups provided a meeting place for discussion and experimentation, and their journals became a means of disseminating information.

Formation of rocket and space flight societies

As World War II approached, minor and varied experimental and research activities on rockets and guided missiles were underway in a number of countries. But in Germany, under great secrecy, there was concentrated effort. In the Soviet Union, experimental sounding rockets were built by enthusiasts in Moscow and Leningrad. Minor financial support was given by the government. Liquid-propellant motors were designed and tested, and in 1936 a sounding rocket altitude of more than three miles was reported. Other research was conducted on an engine for a rocket-powered airplane designed by S.P. Korolev, who later became the leading Soviet space vehicle designer.

In the United States, Goddard was at Roswell, working with a small crew of technicians, making many improvements in his sounding rocket design. There he constructed and launched 31 rockets, reaching 2,000 feet (600 metres) altitude in 1930 and 7,500 feet (2,300 metres) in 1935. Goddard was a brilliant innovator. He worked under self-imposed secrecy, and few details of his work were known until after World War II.

Elsewhere in the United States, the American Rocket Society conducted a number of rocket engine and flight tests in the vicinity of New York City. At the California Institute of Technology, under Theodore von Kármán, a number of graduate students started a program of rocket research and development in 1936. Backed by Guggenheim Foundation funds, detailed scientific studies of solid- and liquid-propellant rocket technology were made, and with military support solid-propellant formulations were developed. Rockets also were studied for the purpose of assisting the takeoff of heavily loaded aircraft.

In France, Esnault-Pelterie and others were receiving some military support on small liquid- and solid-propellant rocket motors. Across the Channel in Great Britain, amateur rocket enthusiasts were prevented from experimentation by stiff laws relating to explosives. Military rocket research was confined to utilizing smokeless cordite powder.

Rocket development in Germany

But in Germany, successful flights as high as one mile with gasoline–oxygen-powered rockets were made in 1931–32 by the German Rocket Society. Funds for such amateur activities were scarce, and the society sought support from the German Army. The work of Wernher von Braun, a member of the society, attracted the attention of Capt. Walter R. Dornberger. Von Braun became the technical leader of a small group developing liquid-propellant rockets for the German Army. By 1937 the Dornberger–Braun team, expanded to hundreds of scientists, engineers, and technicians, moved its operations from Kummersdorf to Peenemiinde, a deserted area on the Baltic coast. Here the technology for a long-range ballistic missile was developed and tested (see below Surface-to-surface: guided missiles).

Two other prewar German activities deserve mention. One was the rocket engine design and development work of the Austrian Eugen Sanger. A brilliant engineer, Sänger had built and tested rocket motors at the University of Vienna. The German Air Force invited Sanger to build a rocket research establishment at Trauen near Hannover.

The other development was the work of Hellmuth Walter, who developed hydrogen peroxide rocket motors. An experimental Heinkel aircraft powered by such an engine made a successful flight in February 1937.

II. Military Systems—World War II to present

World War II saw the expenditure of immense resources and talent for the development of rocket-propelled weapons. Except in the case of Germany, the greater part of the effort was in development of unguided rockets. As German resistance collapsed in 1945, the Allies avidly sought details of rocket and guided missile development. Particularly sought were the key personnel, who had years of experience in the design, development, and testing of missile system components: propulsion, airframe, guidance, and warheads. More than 100 specialists, headed by the foremost expert, Wernher von Braun, surrendered to the U.S. Army and moved to the United States in 1945

MILITARY ROCKETS AND GUIDED MISSILES

The four major categories of military rockets and guided missiles are: surface-to-surface, surface-to-air, air-to-surface and air-to-air. Free flight, or unguided military rockets are included in appropriate surface- or air-launch categories. Submarine- (underwater-) launch missiles are included in the surface-to-surface launch category. In the following summation, the developments relating to each of the four categories are taken up in turn.

Surface-to-surface. Free-flight rockets. Germany used widely a battlefield rocket, the Nebelwerfer, of 15-centimetre (about six-inch) and 21-centimetre calibre, from six-barrelled launchers. The 15-centimetre rocket was about 40 inches (100 centimetres) long, weighed about 80 pounds (35 kilograms), could be fired at the rate of one per second, and carried a high-explosive warhead. Maximum range was more than 6,000 yards (six kilometres).

The Rheinbote

Other German developments included the Panzerfaust and Panzerschreck, hand-held tube-launched rockets. One other noteworthy rocket was the Rheinbote, a four-stage, solid-propellant rocket. It weighed nearly two tons, was about 40 feet (12 metres) long, and had an impressive range—135 miles (220 kilometres). The payload of 88 pounds (40 kilograms) of high explosive, however, could not compete with other weapons such as aircraft bombardment and the V-2. The Rheinbote was significant, nevertheless, in that practical recognition was made of the value of the step principle suggested by Haas almost 400 years before.

In Great Britain, a five-inch rocket with a 30-pound explosive warhead was developed. Its range was two to three miles. These rockets, fired from specially equipped naval vessels, were used in heavy coastal bombardment prior to landings in the Mediterranean. Firing rates were 800–1,000 in less than 45 seconds from each ship.

The United States did not commence active development of rockets until mid-1940. At that time, the National Defense Research Committee (NDRC) authorized a program under the direction of C.N. Hickman, who had worked with Goddard on the tests of hand-launched rockets at Aberdeen in 1918. Hickman supervised the development of a refined design, known as the bazooka. About 20 inches long and weighing 3.5 pounds, the rocket was fired from a shoulder launcher that weighed 14.5 pounds. The bazooka was used extensively against tanks. Its maximum range was short (600 yards) and it travelled slowly, but it carried a potent shaped-charge warhead.

Another United States Army development was the Calliope, a 60-tube launching projector for 4.5-inch rockets mounted on a Sherman tank. The launcher was mounted on the tank's gun turret, and both azimuth (horizontal direction) and elevation were controllable. Rockets were fired in rapid succession (ripple-fired) to keep the rockets from interfering with one another as they would in salvo firing. Other launchers were developed for trucks and jeeps.

Close liaison was established between the U.S. committee and similar research groups in Great Britain. Advantage was taken of earlier rocket developments by the British. In this cooperative effort the United States developed an antisubmarine rocket-propelled weapon, known as Mousetrap, based upon a similar British design, Hedgehog.

Other conventional rockets developed in the United States included a 4.5-inch barrage rocket with a range of 1,100 yards and a five-inch rocket of longer range. The latter was used extensively in the Pacific theatre of war from launching barges against shore installations, particularly just before landing operations. The firing rate of these flat-bottom boats was 500 per minute. Other rockets were used for smoke laying and demolition. Because the United States had great industrial capacity, unhampered by air raids, and urgent needs for its own forces as well as for its Allies, its production of solid-propellant rockets increased to a high rate by the end of the war. The United States produced more than 4,000,000 of the 4.5-inch rockets and 15,000,000 of the smaller bazooka rockets during the war.

Soviet barrage rockets

As far as is known, Soviet rocket development activity during World War II was limited. Extensive use was made of barrage, ripple-fired rockets. Both A-frame and truck-mounted launchers were used. The Soviets mass-produced a 5.1-inch rocket known as Katyusha. From 16 to 48 Katyushas were fired from a boxlike launcher known as the Stalin Organ, mounted on a gun carriage.

After World War II, the impetus of advanced rocket designs of free-flight rockets continued at a much slower pace. By 1950 the United States Army had begun development of the Honest John rocket; it became operational in 1955. The Honest John is 24.8 feet (7.6 metres) long, about 30 inches (75 centimetres) in diameter, and weighs 4,700 pounds (2,100 kilograms). It has a range of about 23 miles (37 kilometres) and can be equipped with either a high-explosive or nuclear warhead. The rocket is both spin-stabilized and fin stabilized. After it leaves the launcher, small rocket motors located forward of the head of the sustainer engine (but arranged tangentially to it) fire and impart a slow spin to the rocket. The four large stabilizing fins at the aft end of the rocket are set at a slight angle to maintain the spin.

A smaller version, Little John, was developed especially for use with airborne infantry divisions. Unlike the larger rocket, Little John is spin stabilized by the launching rail as it is fired; the spin is maintained by slightly offset fins.

Following World War II the Soviet Union produced its Frog series of large, solid-propellant, ballistic free-flight rockets. (In the absence of public designation of Soviet missiles, the code names [*e.g.*, Frog] used by the North Atlantic Treaty Organization are used in this article.) The Frog 1 missile, in service since about 1957, is a single-stage spin-stabilized rocket 31 feet (nine metres) long with a body diameter of about two feet (0.5 metre), a bulbous nose, and six fins with a span of three feet three inches (one metre). The gases exhaust through a cluster of seven nozzles. The missile is carried on a tracked armoured vehicle and is believed to be capable of delivering either a conventional or a nuclear warhead a distance of about 15 miles (25 kilometres). The latest version is the Frog 7, displayed in 1965. This appears to be a much cleaner design, cylindrical in shape, with four small fins and a main nozzle ringed by 12 small nozzles. It is rail-launched and carried by a wheeled transporter erector.

Between the end of World War II and 1960 the development of smaller solid-propellant rockets spread to other countries, primarily because of the effectiveness of such weapons. Thus by the 1960s all major powers were developing and producing antitank and battlefield-support rockets.

Guided missiles. In Gerniany, as mentioned above, the army was developing a long-range ballistic missile at Peenemiinde. The A-4 (popularly known as the V-2, for Vergeltungswaffen Zwei, "vengeance Weapon Two") was a remarkable engineering achievement (see Figure 4).

Camera Press—Publix

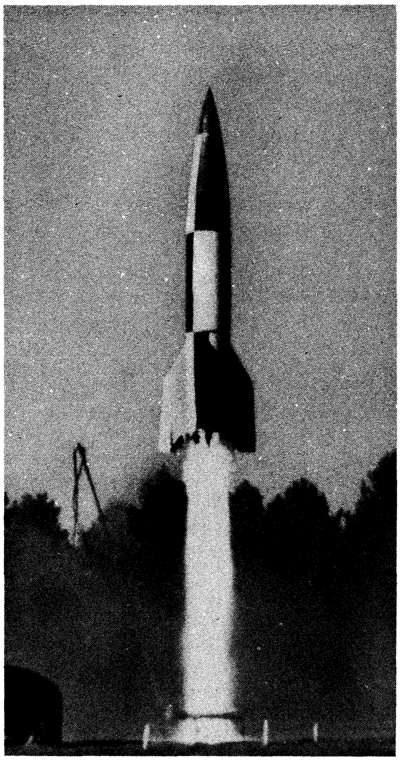


Figure 4: German V-2, nearly 47 feet long, a 13.5-ton missile with a 200-mile range. First fired in 1944, it carried a conventional warhead weighing one ton.

The V-2 rocket was nearly 47 feet (14 metres) long, had a cylindrical diameter of 5.5 feet (1.5 metres), and weighed about 13.5 tons at takeoff, including the one-ton warhead. Propellants were liquid oxygen and a 75 percent ethyl alcohol–water mixture. Approximately 8,400 pounds of alcohol and 10,800 pounds of oxygen were burned at a rate of 300 pounds per second. The rocket motor produced a 55,000-pound thrust for about one minute, after which the engine was shut off. Propellant supply was by turbopump (see below *Propellant supply system*). Yaw and pitch control was accomplished by two pairs of graphite carbon vanes in the rocket exhaust. Ignition of the alcohol–liquid oxygen propellants was effected by a pyrotechnic pinwheel inserted in the motor. Maximum range was about 200 miles (320 kilometres).

The first operational V-2 was fired against Paris on September 6, 1944. Two days later, the first of more than 1,000 missiles was fired against London. The missile travelled on a ballistic arc trajectory, reaching a maximum speed of more than one mile per second and an altitude of 60 to 70 miles. Since it approached the target faster than the speed of sound, there was no warning of its approach. By the end of the war about 4,000 of these missiles had been launched from mobile bases against Allied targets. During February and March 1945, only weeks before the war in Europe ended, an average of 60 missiles was launched weekly.

V-2 fired against London

Although the V-2 did not become a decisive weapon, it was a landmark achievement in rocketry. Large rocket motor developments after World War II, in both the United States and the Soviet Union, drew heavily upon V-2 engine design.

The only other relatively long range, surface-to-surface guided missile to see action in World War II was the V-1, or "buzz-bomb," a development of the German Air Force. The V-1 has been called, from an engineering point of view, an aerial counterpart of the naval torpedo. The analogy is striking in several ways. Once the V-1 was launched, its course could not be corrected, being determined by a preset guidance system. An accelerometer, a sensitive instrument for detecting accelerations, sensed deviation to the left and right of the programmed flight path and sent correcting signals to the control surface actuators. The missile's altitude was sensed and maintained in a similar manner by a barometer, and range was determined by the number of revolutions turned by a small propeller in the nose of the missile. When the requisite number of revolutions, corresponding to a linear range, had been made, a signal was sent to the aerodynamic control surface actuators to cause the V-1 to dive vertically onto the target. The average range was 150 miles. Launching ramps and a hydrogen peroxide-powered catapult boosted the V-1 to flying speed.

Pulse-jet engine in the V-1

The V-1 used a unique pulse-jet engine that found later employment in early post-World War II Soviet and U.S. missiles. Air enters a diffuser through a series of spring-loaded flapper valves that open as the air passes through them and then close when the air is in the combustion chamber. At this point in the firing cycle, kerosene from a gas-pressurized tank is sprayed into the chamber and ignited by a spark plug. As the chamber pressure of the burning gas rises, the flapper valves close and the exhaust gases are forced through the nozzle of the motor. The V-1 warhead was a 2,200-pound charge of high explosive set off by an impact fuze.

A final but significant weapon under development by the Germans during World War II, and one that spurred investigation in Great Britain, France, the Soviet Union, the United States, and Switzerland, was the X-7 antitank missile. Although this missile was not developed in time for combat, its technology set the pattern for practically all antitank guided missiles developed during the next 20 years. The X-7 was a solid-propellant winged rocket with a shaped-charge warhead and a range of about half a mile. This missile was to have been guided in flight by means of signals sent over two wires connected to control surfaces located on the wings, similar to the X-4 air-to-air missile (see below). The operator visually followed the missile in flight and sent guidance correction signals by manual control.

When World War II came to a close in August 1945, development and production of rockets was a massive enterprise in the United States and Great Britain. Despite a few ingenious uses of rocket power, however, the major effort was along conventional lines, adapting technical improvements to existing weapons. Nowhere else was there the depth of appreciation and scope of plans for tactical guided missiles or long-range ballistic rockets that there was in Germany. Recognition of the potential of rocket power by the Allied powers came too late in the war to enable them to catch up with German developments.

Appreciation of the potential importance of German technical efforts was indicated by the speed with which technical intelligence teams followed close behind front-line troops as they moved across Germany. These teams obtained masses of technical data, design drawings, and missiles. They also interrogated key scientists and missile engineers. The top planning and technical staff of Peenemiinde, headed by Dornberger and von Braun, fled south in the last few days of the war in order to surrender to U.S. troops. The Soviet troops that captured Peenemiinde found mostly wreckage and had orders to destroy what was left. Soviet forces, however, took many German technicians to the Soviet Union. Thus, the U.S., France, Great Britain, and the Soviet Union all had the benefit of information on rockets and guided missiles captured in Germany. In general, early postwar development of rockets followed most of the lines suggested by German work. Rockets were used to propel all types of guided missiles, competing successfully in these missions with air-breathing jet engines.

The United States obtained not only the services of top German rocket experts but also, from the underground V-2 factory at Niedersachswerfen, enough V-2 components for about 100 complete vehicles. About 70 V-2s were fired during the period 1946–51 at White Sands (New Mexico) Proving Ground. These firings provided much experience in the handling and launching of large rockets.

V-2 firings in the U.S.

The long-range ballistic missile was not a major development in the United States until 1954. Until that time air-breathing subsonic missile developments such as the Snark and Navaho were counted upon by military planners to supplement and supersede long-range, strategic-bomber aircraft. These vehicles were essentially unmanned, turbojet- or ramjet-engine-powered, high-speed aircraft equipped with warheads. The Corporal (75-mile range), with a mobile launcher, developed by the U.S. Army, was operational by this time as a battlefield support weapon.

By the 1960s the Pershing missile (400-mile range) replaced the Corporal. Capable of carrying either a conventional or nuclear warhead, Pershing is deployed in Europe.

Numerous wire-guided antitank rockets presently operational have been built by the United States, Great Britain, France, and many other countries.

In 1954 two developments put the intercontinental ballistic missile (ICBM) in a new light. One of these developments was the thermonuclear bomb, with destructive power measured in megatons (millions of tons of TNT equivalent). The other was the miniaturization and refinement of inertial guidance systems (see GYROSCOPE), so that they became sufficiently accurate to place the warhead of the ICBM close to a target 5,000 miles or more away.

The first ICBM authorized in the United States was the Atlas, followed later by the Titan. The Atlas was a one and one-half-stage vehicle. Whereas a two-stage vehicle drops both first-stage rocket engine and tankage, the one and one-half-stage vehicle drops only the rocket engine, and the second stage continues to use the original tankage. The first Atlas was fired full range from the Cape Canaveral (now Cape Kennedy, Florida) missile range in November 1958. The Titan I ICBM was a conventional two-stage vehicle. First successful tests of the Titan were made in 1959. Both the Atlas and Titan burned liquid oxygen and hydrocarbon fuel (similar to kerosene).

A major logistic problem in ballistic missiles using cryogenic (extreme low-temperature) propellants such as liquid oxygen is the necessity for fuelling just prior to launching. This weakness, which means delay in defensive, retaliatory fire, can be eliminated by the use of solid propellants or storable noncryogenic propellants. A later version of Titan, Titan II, used a higher performance propellant combination of nitrogen tetroxide and hydrazine-based fuel. This propellant combination permits the missile to stand ready, fully loaded for firing.

Development of solid propellants to a level of performance near that of liquids achieved such progress by 1958 that a solid-fuelled ICBM, the Minuteman, was authorized.

For greater defense against attack, the Minuteman was designed to be launched from a 90-foot underground concrete silo. Minuteman I was first launched in 1961. Consisting of three stages, the missile was 54 feet (16 metres) long. In September 1964, 650 Minuteman I missiles were in place and an improved Minuteman II was in development. By 1971 a further improved Minuteman III was test launched. A recent version is nearly 60 feet long and has the capability of carrying MIRV warheads (see below Warheads). Range is over 7,000 miles.

In 1947 the United States Navy demonstrated that a V-2 could be launched from the deck of a ship at sea. The problems associated with the production and storage of liquid propellants aboard ships soon turned the navy toward a quest for a solid-propellant IRBM; the result was the two-stage Polaris, first tested in 1958 and fired underwater in 1960.

The Polaris missile

Launching of Polaris missiles is accomplished by compressed air. The first-stage rocket motor fires only after the missile is in the air, above the surface. Sixteen Polaris missiles are carried in two parallel rows of eight on each of the nuclear-powered United States fleet ballistic missile submarines. In 1971 a program to replace Polaris missiles with the longer, wider diameter Poseidon missile (see Figure 5) was begun. The Poseidon, like Minuteman III, is capable of carrying MIRV warheads.

By courtesy of the U.S. Navy

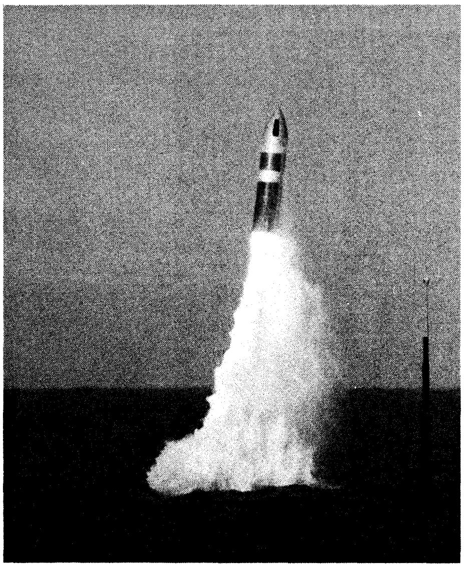


Figure 5: U.S. Poseidon missile being launched from a nuclear- ower red submarine.

Another submarine-launched missile is the antisubmarine Subroc. Launched submerged, the missile breaks the surface and is rocket-propelled on a ballistic trajectory toward an enemy submarine. Re-entering the water, the missile assumes the role of a homing torpedo.

Ten years after World War II, the United States Air Force and Army each undertook the development of an IRBM. The air force Thor and the army Jupiter used the same liquid-propellant engine, and each had a range of 1,600 miles (2,600 kilometres).

Great Britain commenced development of the Blue Streak IRBM but cancelled it in 1960. France, after a late start, began development of a three-stage, solid-propellant, 2,000-mile-range ballistic missile, but progress has been slow. The technology has been used also for the French space launch vehicle, Diamant.

Developments in IRBM and ICBM design in the Soviet Union paralleled United States efforts in the years immediately following World War II. The Soviets fired several captured V-2 missiles from the rocket proving ground at Kapustin Yar in 1946 and 1947.

Early Soviet surface-to-surface guided-missile developments borrowed heavily from German technology, but later Soviet weapons were original designs. The Scud A is a liquid-fuelled, single-stage missile that appears to derive from the German Wasserfall surface-to-air missile. This missile and its successor, the Scud B, are both believed to have a range of about 50 to 100 miles and probably are guided by radio commands.

The first Soviet IRBM's were the Shyster (700-mile range), Sandal (1,200-mile range), and Skean (2,000-mile range), all single-stage vehicles. The Shyster is basically a stretched and uprated version of the German V-2 rocket and is fuelled by liquid oxygen and kerosene. The Sandal and Skean are powered by storable propellants and appear to be essentially original Soviet designs. The Sandal was the missile deployed to Cuba in 1962 and withdrawn after crisis negoiiaiions.

The Soviet Scud missiles

Soviet tactical and medium-range missiles are highly mobile, mounted on cross-country vehicles that allow them to advance with combat troops. They can be concealed in wooded country, making detection by aircraft and satellites more difficult. Scud is an example. Two variants of this storable-liquid-propellant missile are known. Scud A travels on a tracked vehicle of the type used for the unguided missile Frog, while the larger Scud B is carried on a fully enclosed cross-country vehicle of the type first shown in November 1965. Each is placed on a simple launch platform at the rear of the vehicle by a tubular cradle that elevates into a vertical position. It is believed that guidance is achieved by a preset inertial system acting upon steerable tail fins.

Scapegoat is carried by a tracked transporter-erector, the full weapon system being designated Scamp in NATO nomenclature. The 35-foot missile appears to be the top two stages of Savage, the Soviet silo-based solid-propellant ICBM.

Scrooge, also mobile on a heavy tracked transporter, is concealed within a cylinder 62 feet long. This launch tube is raised vertically behind the vehicle, and the missile is fired directly from it.

Weapons of this type can be moved effectively along the entire frontier of the Soviet Union, providing coverage of western Europe, the Middle East, and Southeast Asia. Scrooge has been reported deployed as far east as the Chinese border near Buir Nuur in Mongolia. Estimated range is 3,000 to 3,500 miles.

The first Soviet ICBM is believed to be a dual-purpose vehicle used both as an ICBM and as the basic booster for much of the Soviet space program. It was displayed publicly for the first time at the 1967 Paris Air Show, where it was placarded as the booster for the Vostok spacecraft, but it has probably been in use since the late 1950s. The vehicle is parallel staged, with a central sustainer core to which are attached four large, tapered boosters. The sustainer and boosters are powered by clusters of thrust chambers, all of which ignite upon launching; midway in the powered flight the four-booster tank and engine combinations are detached while the sustainer continues burning. The propellants are liquid oxygen and a hydrocarbon, and the engines develop a total sea-level thrust of about 1,000,000 pounds.

In 1964 and 1965 the Soviets displayed the Sasin and Scrag, both tandem, liquid-propellant, long-range rockets. Scrag, which has three stages separated by interstage trusses, was described as being from the same family as the vehicles that placed Soviet cosmonauts into orbit. It does not appear, however, to have become operational as an ICBM.

The premier Soviet missiles in this class are the SS 11, SS 13 Savage, and SS 9 Scarp. The SS 11 is reported to employ storable liquid propellants, but by 1971 it had not been publicly displayed by the Soviet authorities. Savage, solid-fuelled, follows the format of the U.S. Minuteman ICBM but is somewhat larger. The liquid-fuelled Scarp is much larger, on the lines of the U.S. Titan ICBM. The first stage has six fixed-thrust chambers and four swivel-mounted vernier rocket motors, small auxiliaries used for trajectory correction. It can be used both as a conventional ICBM launched across the Northern Hemisphere or as a fractional orbit bombardment system (FOBS) weapon (see below *Military* space vehicles).

In the FOBS application the Scarp could be launched almost to orbital velocity to attack the United States over the Southern Hemisphere, thus avoiding the Ballistic Missile Early Warning System (BMEWS) radar stations in Britain, Greenland, and Alaska.

By the spring of 1970 the number of SS 11 and SS 13 ICBM's deployed were stated by the U.S. secretary of defense to be around 800. Also deployed were about 220 SS 9s, each capable of carrying a warhead of some 25 megatons yield.

The Soviet Sawfly

In 1962 the Soviets displayed a 48-foot-long missile, Sark, describing it as a submarine-launched missile, although its very heavy construction and large size made it seem unsuitable for operational use in submarines. Two years later the smaller Serb was displayed and also credited with a submarine application. The 33-foot-long missile is similar in configuration to the U.S. Polaris. It has two solid-fuelled stages with a range of less than 1,000 miles. A submarine-launched ballistic missile that has recently become operational is Sawfly. First displayed in 1967, the two-stage missile is 42 feet (13 metres) long and about five feet nine inches (1.75 metres) in diameter. The four first-stage nozzles are gimbal-mounted, permitting the stages to be moved freely so that their direction (thrust vector) can be controlled. Range is reported to be around 1,300 miles (2,100 kilometres).

The most advanced Soviet nuclear-powered submarine in the early 1970s was the Y-class, which has 16 launch tubes for missiles of the Sawfly family.

Regardless of the country of origin, all IRBM's and ICBM's are equipped with either nuclear or thermonuclear warheads, because (1) it is economically impractical to use such an expensive missile to deliver a relatively small amount of high explosive; and (2) the unavoidable inaccuracies in range and direction make essential the far-reaching effects of a nuclear explosion.

Table 1 gives information on representative surface-to-surface missiles of various nations.

Surface-to-air. The major purpose of the class of surface-to-air weapons is to intercept and destroy enemy aircraft, particularly at altitudes beyond the effective capability of conventional anti-aircraft artillery. During World War II high-altitude bombing above this range necessitated the development of rocket-powered weapons.

In Great Britain, initial effort was aimed at achieving the equivalent destructive power of the three-inch and later the 3.7-inch anti-aircraft gun. Single, double, and then multiple launchers were produced.

Two important innovations were developed by the British in connection with the three-inch rocket. Both devices were directed against German dive bombers. One was a rocket-propelled aerial-defense system. A parachute and wire device was rocketed aloft, trailing a wire that unwound at high speed from a bobbin on the ground. Altitudes as high as 20,000 feet were attained. Several versions of this were used, quite successfully, from ships. The other device was a type of proximity fuze utilizing a photoelectric cell and thermionic amplifier. A change in light intensity on the photocell caused by light reflected from a nearby airplane (projected on the cell by means of a lens) triggered the explosive shell.

The only significant anti-aircraft rocket development by the Germans was the Taifun. A slender, six-foot liquid-propellant rocket of simple concept, the Taifun was intended for altitudes of 50,000 feet. The design embodied coaxial tankage of nitric acid and a mixture of organic fuels. This weapon, though planned for mass production, never became operational.

Wasserfall supersonic missile

During World War II Germany made efforts to produce effective surface-to-air missiles. Three of those under development were subsonic (less than the speed of sound): the Schmetterling and Enzian resembled stubby midwing aircraft, and the Rheintochter had cruciform wings. All used solid-propellant boosters, or takeoff rockets, and sustainer motors, except that there was a liquid-propellant version of the Rheintochter. Altitude capability was about 50,000 feet. The Wasserfall was a superior weapon designed to operate at supersonic (greater than the speed of sound) speeds. A single-stage missile weighing 7,800 pounds at takeoff, this weapon was powered by a 17,000-pound-thrust motor and reached a velocity of 2,500 feet per second. Maximum altitude was 60,000 feet, and the weight of the warhead was 330 pounds. Propellants were nitric acid and an organic liquid, vinyl isobutyl ether. All of these weapons would have been effective against enemy aircraft flying at 15,000 to 30,000 feet. Although none

Table 1: Representative Surface-to-Surface Missiles*

country and service	name	length (ft)	takeoff weight (lb)	range (mi)	propulsion	characteristics
Australia, navy	Ikara	11	—	—	rocket, solid	antisubmarine, homing torpedo
France, army, navy	RAP-14	6.6	115	∼9	rocket, solid	multiple launcher
France, army	SS11	3.9	66	∼2	rocket, solid	wire-guided, antitank
France, army, navy	SS12	6	167	>3	rocket, solid	wire-guided
France, air force	SSBS	48.5	70,000	>2,000	rocket, solid; 2 stages	MRBM, silo-launched, development
France, navy	MSBS	∼36	∼40,000	>1,200	rocket, solid; 2 stages	submarine-launched, development
West Gennany, army	Cobra	3.1	22.5	1	rocket, solid	wire-guided, antitank
Great Britain, navy	Seacat	5	130	∼10	rocket, solid	radio-commanded
Great Britain, army	Vigilant	3.0	31	∼1	rocket, solid	wire-guided, antitank
Israel	Gabriel	11	880	—	—	—
Italy, navy	Sea Killer	11.5	375	>5	rocket, solid	radio-commanded
Japan, army	KAM-3D	3.1	34.6	∼1	rocket, solid	wire-guided, antitank
Norway, navy	Terne	6.4	298	∼5	rocket, solid	antiship
Sweden, army	Bantam	2.8	16.5	>1	rocket, solid	wire-guided, antitank
Sweden, navy	RB08A	18.8	2,650	∼150	turbojet	antiship
Italy, army	Mosquito	3.7	31	∼1.5	rocket, solid	wire-guided, antitank
U.S.S.R.	Swatter	3.7	—	—2	rocket, solid	wire-guided, antitank
U.S.S.R.	Skean	75	—	—2,000	rocket, liquid	IRBM
U.S.S.R.	Sasin	80	—	—6,500	rocket, liquid	ICBM
U.S.S.R.	Scarp	113.5	—	unlimited	rocket, liquid	fractional orbital bombardment system (FOBS)
U.S.S.R.	Sawfly	42	—	∼1,500	rocket, solid	submarine-launched IRBM
U.S.S.R.	Scrooge	∼55	—	∼3,000	rocket	mobile IRBM
U.S.S.R.	Savage	64	—	∼5,000	rocket, solid	silo-lnunchcd ICBM
U.S., navy	Asroc	15	1,000	6	rocket, solid	rocket-boosted, acoustic-homing torpedo
U.S., army	Honest John	24.8	4,700	23	rocket, solid	battlefield support, spin-stabilized
U.S., army	Lance	20	3,200	30	rocket, liquid	battlefield support
U.S.. air force	Minuteman III	59.8	76,000	>7,000	rocket, solid; 3 stages	ICBM, inertial guidance. silo-launched
U.S., army	Pershing	34.5	10,000	400	rocket, solid; 2 stages	battlefield support
U.S., navy	Polaris A-3	31	∼30,000	2,875	rocket, solid; 2 stages	submarine-launched
U.S., navy	Poseidon	34	65,000	∼3,000	rocket, solid; 2 stages	submarine-launched
U.S., army	Sergeant	34.5	10,000	85	rocket, solid	battlefield support
U.S., army	TOW	—	∼75	>2	rocket, solid	wire-guided, antitank

*∼ Approximate; > more than.

became operational, their development reached a point of experimentation that proved that missiles could be guided by radar beams.

Two anti-aircraft missiles were designed in the United States in the latter days of the war: the Lark and the Little Joe. The former was a 14-foot-long liquid-propellant weapon with semi-active radar homing. Weighing 1,200 pounds, it was accelerated to operational speed by two solid-propellant jato (jet-assisted-takeoff) units. It could deliver a 100-pound warhead to a range of 38 miles. The Little Joe was an 11-foot-long solid-propellant missile with a range of 2.5 miles. Its payload was a 100-pound warhead designed to counter Japanese kamikaze attacks on U.S. Navy ships. Neither weapon became operational.

Following World War II the major nations continued the development and refinement of anti-aircraft guided missiles. In the United States the navy initiated a research program the purpose of which was to develop a ramjet engine suitable for use in a high-performance anti-aircraft missile. Ultimately the Talos missile resulted. Thirty-three feet long, this ship-launched missile has a slant range of about 75 miles. Other U.S. Navy surface-to-air missiles are the Terrier, Tartar, and Sea Sparrow. The Standard missile is replacing Terrier and also has a surface-to-surface capability.

The United States Army concentrated its efforts in the field of anti-aircraft weapons on the Nike group of missiles. The first of these was the Nike Ajax, which was first test-fired in 1951 and which remained deployed with the army until 1961. The missile was 21 feet (six metres) long and weighed 2,455 pounds (1,115 kilograms). With a range of 25 miles (40 kilometres), it had a liquid-propellant engine and a velocity of 1,500 miles (2,400 kilometres) per hour.

An advanced version of the Ajax was conceived in 1953, known first as Nike B, later as Nike Hercules. The Hercules was 41 feet long and weighed 10,000 pounds. It had a solid-propellant motor, could carry both nuclear and high-explosive warheads, and had a range of 85 miles. Both the Nike Ajax and the Nike Hercules employed radar-command guidance.

Safeguard anti-ballistic missile system

Out of the early experience in the Nike program also grew the possibility for an antimissile missile. The U.S. version was the Safeguard antiballistic missile (ABM) system. Safeguard consists of two missiles, the 55-foot Spartan for high-altitude missile interceptions and the 27-foot Sprint for low-altitude operations. Both are designed to carry nuclear warheads. This system was under final development in the early 1970s.

The U.S. Air Force Bomarc is ramjet powered and capable of interception at ranges as great as 400 miles. Solid-propellant boosters are used on ramjet-powered missiles to bring the engine up to operational speed of flight.

Other significant current U.S. surface-to-air missiles include the Hawk, capable of intercepting low-flying aircraft, as is the shoulder-launched Redeye. Both have homing devices.

The Soviet Union has shown numerous surface-to-air missiles: Guild, Guideline, Goa, Griffon, Gainful, Ganef, and Galosh. Of these, Guideline has been widely deployed in the Soviet Union, in Warsaw Pact countries, and in certain Middle East countries, as well as in Cuba, Indonesia, and North Vietnam. It has been used in Vietnam against U.S. aircraft with limited success.

In Egypt, during the Six-Day War of 1967, an advanced version of Guideline was used against the Israeli Air Force. The weapon, similar in many respects to the U.S. Nike Ajax, has a solid-fuel booster and a liquid-propellant (nitric acid–kerosene) sustainer. Later versions are reported to have a slant range of about 28 miles, an altitude ceiling above 60,000 feet, and a maximum speed of mach 3.5 (3.5 times the speed of sound).

Goa, first displayed in 1964, is a low-to-medium-altitude weapon mounted on a twin launcher. Used on ships of the Soviet Navy, it has also been developed as a field weapon and appeared in Egypt in 1970.

Guild was an early surface-to-air missile. Solid-fuelled, it was about 39 feet long and appeared in Moscow parades mounted on an articulated trailer behind a truck.

Another low-level weapon is Gainful, first displayed in 1967. Three of the 19.5-foot-long solid-fuel missiles are carried, on a tracked vehicle for the defense of Soviet units in the field.

As a surface-to-air weapon, Ganef is unique among Soviet missiles in its use of a ramjet for main propulsion. Four small, strap-on, solid-fuel boosters provide initial velocity. Two missiles are carried on a tracked vehicle, and the system depends on command guidance.

Novosti Press Agency

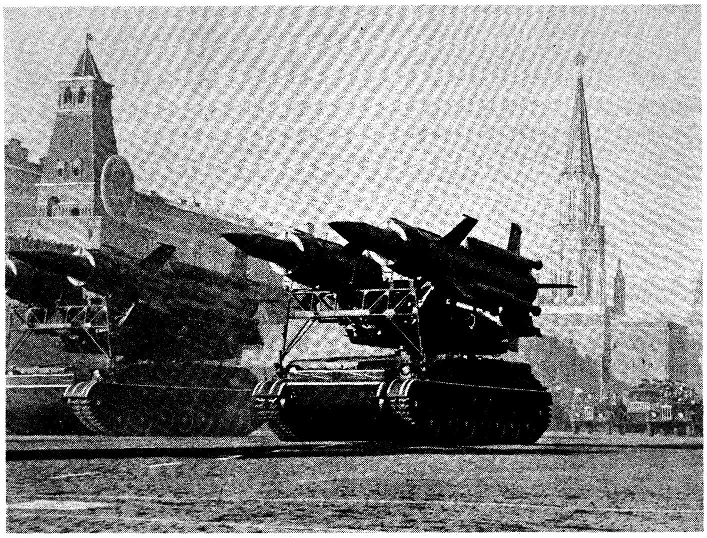


Figure 6: The Ganef, a Soviet mobile air defense weapon.

Griffon, a long-range anti-aircraft missile, is also stated to have an antimissile capability.

The more definitive antimissile missile is Galosh, a cone-shaped weapon first shown in 1964 within a tubular container measuring 67 by nine feet. The first stage has four rocket nozzles. Launching is from fixed emplacements under radar command. First sites were being established on the perimeter of Moscow in 1967.

Table 2 provides additional information on representative surface-to-air missiles.

Table 2: Representative Surface-to-Air Missiles

country and service	name	length (ft)	weight (lb)	range (mi)	propulsion
France, navy	Masurca	30	4,380	20	rocket, solid
France	Crotale	—	165	—5	rocket, solid
France, West Germany	Roland	7.3	140	—4	rocket, solid
Great Britain, air force	Bloodhound	27.8	—4,000	>50	rocket, solid; ramjet
Great Britain, navy	Seacat	58	—130	—5	rocket, solid
Great Britain, navy	Seaslug	20	—3,500	—20	rocket, solid
Great Britain, army	Thunderbird	20.8	—3,000	—15	rocket, solid
Italy	Indigo	10.5	215	—6	rocket. solid
Switzerland	Micon	18	1,760	20	rocket, solid
U.S.S.R.	Galosh (anti-ICBM)	—64	—	—	rocket, solid
U.S.S.R.	Gainful	19	—	—20	—
U.S.S.R.	Guideline	35	3,000	28	rocket, solid; liquid
U.S., army	Chaparral	9.5	—200	—8	rocket, solid
U.S., army	Hawk	16.5	1,300	22	rocket, solid
U.S., army	Spartan (anti-ICBM)	55	—	>300	rocket, solid
U.S., army	Sprint (anti-ICBM)	27	—	25	rocket, solid
U.S., army	Redeye	4	20	—2	rocket, solid
U.S., navy	Standard	27	3,000	>10	rocket, solid
U.S., navy	Talos	33	7,000	—75	rocket, solid; ramjet

Air-to-surface. *Free-Bight rockets.* In World War II, Great Britain, Germany, the Soviet Union, Japan, and the United States all developed airborne rockets for use against surface as well as aerial targets. These were almost invariably fin stabilized because of the effective aerodynamic forces when launched at speeds of 250 miles per hour and more. Tube launchers were used at first, but later straight-rail or zero-length launchers, located under the wings of the airplane, were employed.

One of the most successful of the German rockets was the two-inch-diameter R4M. The tail fins remained folded until launch, facilitating close loading arrangements.

The U.S. achieved great success with a 4.5-inch rocket, three or four of which were carried under each wing of Allied fighter planes. These rockets were highly effective against motor columns, tanks, troop and supply trains, fuel and ammunition depots, airfields, and barges.

Whereas British and German aircraft rockets were fin stabilized, the U.S. designs were spin stabilized, resulting in greater accuracy. The largest such rocket was the Tiny Tim. Slightly over 10 feet long, it carried 150 pounds of TNT. The solid-propellant motor developed 30,000 pounds of thrust for one second. More than 7,000,000 aircraft rockets were produced in the U.S. during 1940–45.

Glide bombs

A variation on the airborne rocket was the addition of a rocket motor and fins to conventional bombs. This had the effect of flattening the ballistic trajectory, extending the range, and increasing the velocity at impact, useful against concrete bunkers and strengthened (hardened) targets. These weapons were called glide bombs, and the Japanese had 224-pound and 815-pound versions. The Soviet Union employed 56-pound and 220-pound versions, launched from the Stormovik fighter aircraft.

***Guided** missiles.* Among the first air-to-surface guided missiles developed by the Germans was the radio-controlled, armour-piercing Fritz-X, which sank the Italian battleship "Roma" after its surrender to the Allies in 1944. Another was the Hs 293 winged bomb. More than 11 feet (three metres) long and weighing about 2,300 pounds (800 kilograms), this radio-controlled weapon destroyed a number of merchant ships in Allied convoys. Although these early versions of glide bombs required optical tracking (and, therefore, clear weather), plans were under way for television viewing of the target by the weapon and radar spotting for nighttime and foul-weather use. The U.S. developed a controlled glide bomb called the Bat, similar to the Hs 293.

The Hs 293 was a radio-guided bomb with a rocket engine using hydrogen peroxide as a propellant. This chemical was decomposed into steam by the use of aqueous potassium permanganate as a catalyst. A crew of three, a pilot, an observer, and a bombardier, were necessary to launch and direct the Hs 293 to its target. The observer's duty was to set the gyroscope that established a reference plane for the missile, and the bombardier piloted the bomb by remote control after its release from the aircraft. It was especially adaptable to targets at sea and found its greatest use there, sinking many merchant ships.

The United States Bat had a much greater range—up to 20 miles. In guidance it was in advance of the Hs 293, since it employed an active radar homing device. The Bat was a glide bomb, however, having no propulsion system. It weighed 1,000 pounds and was 12 feet long, with a wing span of ten feet. Developed primarily as a weapon for use against ships, the Bat was to be released from an airplane at an altitude of three to five miles. The bomb's velocity of 300 miles per hour was considerably slower than the 470-mile-per-hour Hs 293. Built in considerable numbers before the end of the war, the Bat was not actually used until April 1945, when it was credited with sinking a Japanese destroyer at the maximum range of 20 miles.

After 1945 the U.S. Air Force explored more complex versions of the air-to-surface missile. One of these was the Rascal, a liquid-propellant missile with inertial guidance and a range of 100 miles. By the mid-1950s development turned to achieving longer ranges. This effort was aided by the advent of lighter weight nuclear warheads. Two of the weapons developed, the United States Hound Dog and the British Blue Steel, had ranges of about 500 miles. The U.S. Skybolt missile, cancelled in 1963, was to have had a range of 1,000 miles. The cause of the cancellation was essentially the difficulty in guidance from an aerial launch; development was stopped in favour of alternative weapons, such as the fixed-base ICBM.

By the early 1970s the United States had seven operational air-to-surface missiles. Systems utilizing a basic missile, with modification for either air or surface launch, such as the Standard missile, were gaining favour.

The Soviet Union has displayed a variety of air-to-surface missiles on bomber aircraft that have been given the NATO designations Kangaroo, Kennel, Kipper, Kitchen, and Kelt. The first three have turbojet propulsion; the last two appear to have liquid-plopellant rocket engines.

The largest of these missiles is the 50-foot-long, swept-wing Kangaroo carried by the Tu-20 aircraft. Its range probably exceeds 300 miles.

Kennel, another aircraft-launched missile, is carried by the Tu-16 Badger aircraft. Powered by a turbojet, it is about 28 feet long with a wing span of about 16 feet. Its role is antishipping. A variant, Kelt, is rocket-powered with an enlarged radome.

Kipper, another antishipping missile, is also launched by the Tu-16. Turbojet-powered, it has a range of about 120 miles.

Kitchen, a more advanced Soviet air-to-surface cruise missile, is powered by a liquid-fuelled rocket engine and carried by the Tu-22. Range is probably about 200 miles.

The complexity and expense of air-to-surface missiles is great. Particularly difficult is the guidance problem. At extreme ranges, programmed and command systems generally are not feasible, and it is necessary to rely on such delicate and exacting systems as inertial guidance.

Table 3 lists representative air-to-surface missiles.

Table 3: Representative Air-to-Surface Missiles

country and service	name	length (ft)	weight (lb)	range (mi)	propulsion
France	AS 12	6.2	165	~5	rocket, solid
France, air force, navy	AS 20	8.5	315	4	rocket, solid
France, air force	AS 30	12.8	1,150	6.5	rocket, solid
France, Great Britain	Martel	13.1	—	~40	rocket, solid
Great Britain, air force	Blue Steel	35	—	~500	rocket, liquid
Sweden, air force	Robot RB 04	14.6	1,000	~6	rocket, solid
Sweden, air force	305 A	11.5	660	~5	rocket, solid
U.S.S.R., air force	Kelt	—	—	>100	rocket, liquid
U.S.S.R., navy	Kennel	28	—	50	turbojet
U.S.S.R., air force	Kangaroo	50	—	>300	turbojet
U.S.S.R., air force	Kipper	—31	—	120	turbojet
U.S., navy, air force	Bullpup A	10.5	570	7	rocket, liquid
U.S., navy, air force	Bullpup B	13.5	1,785	10	rocket, liquid
U.S., navy	Condor	—	—	~40	rocket, solid
U.S., air force	Standard ARM	14	1,300	~35	—
U.S., air force, navy	Shrike	10	390	10	rocket, solid

Air-to-air. ***Free-flight rockets.*** As might be suspected, aircraft-launched rockets can be used against aerial as well as ground targets, provided that the aerial targets are propeller-driven aircraft with a speed of 400 knots or less. Near the end of the war the German Me 262 jet fighter carried 48 such rockets. On one sortie, six Me 262s shot down 14 B-17 bombers during a daylight raid, without the loss of one fighter aircraft.

Rockets launched by Me 262 aircraft

After World War II the United States fitted many aircraft with 2.75-inch Mighty Mouse and Aeromite folding-fin rockets. Firing could he singly or in clusters. Another postwar development by the U.S. was a version of the army's five-inch, spin-stabilized rocket loaded by belt and fired from within the wing of the navy Skyraider aircraft.

Other countries fitted airborne rockets to aircraft for air-to-air weapons, notably Sweden (Gerda, three inches, 15.5 pounds), Italy (2.4 inches, 7.9 pounds), and Switzerland (3.15 inches, 22 pounds).

Guided ***missiles.*** As in the case of other guided missiles, Germany led in the development of the air-to-air category, although, similarly, the weapon, called the X-4, never reached operational use. An ingenious design, the X-4 was 6.5 feet (two metres) long, 8.6 inches (21.8 centimetres) in diameter, and weighed 132 pounds (60 kilograms). It had two sets of cruciform wings and fins. Two opposing wings carried pyrotechnic guide flares, and the other two carried streamlined bobbins, which trailed as much as four miles of fine copper wire. Along these wires travelled corrective directional signals that automatically imparted drag to the rear fins while the vehicle, travelling almost 600 miles (950 kilometres) per hour, spun slowly. The X-4 was powered by a rocket motor using nitric acid–hydrocarbon fuel and carried a 44-pound warhead.

After World War II the first U.S. air-to-air guided missile was the Firebird. It was ten feet long and had a solid-propellant booster and a liquid-propellant sustainer motor. It was aimed initially by radar from the launching aircraft, and it homed on the target, guided by an on-hoard radar system. First test-launched in 1947, the Firebird was considered obsolete three years later. It was replaced by more sophisticated supersonic missiles, such as the Sparrow, Falcon, and Sidewinder.

Increased range of air-to-air missiles

By the 1970s air-to-air missiles had become increasingly sophisticated. Although short-ranged (two miles) missiles such as Sidewinder 1A were still in operational use, supersonic aircraft required longer range missiles. Ranges have grown to ten to 12 miles, and the Phoenix and Falcon (AIM-47A) have ranges as great as 100 miles.

The Soviets have displayed five air-to-air missiles, NATO-coded as Alkali, Anab, Ash, Atoll, and Awl. Atoll, which resembles the U.S. Sidewinder, is widely used by the Soviet Union and by Warsaw Pact countries. It has also been exported to Afghanistan, Egypt, Cuba, Finland, India, Indonesia, Iraq, North Vietnam, and Syria.

Anab, in both infrared and semi-active radar homing versions, arms a number of all-weather interceptors. The much larger Ash, again with alternative infrared and radar homing versions, is carried by the Tupolev Tu-28 Fiddler long-range interceptor.

The earlier Alkali missile arms the all-weather MiG-17 and MiG-19 aircraft.

Table 4 gives information on representative air-to-air missiles.

Table 4: Representative Air-to-Air Missiles

country and service	name	length (ft)	weight (lb)	range (mi)	propulsion
France, air force	R 530	10.8	430	11	rocket, solid
Great Britain, navy	Firestreak	10.5	300	0.75–5	rocket, solid
Great Britain, navy	Red Top	11.5	—350	7	rocket, solid
U.S.S.R., air force	Atoll	—10	—	—4	rocket, solid
U.S.S.R.	Ash	—18	—	—	rocket, solid
U.S., air force	Falcon	6.5	110–120	> 5	rocket, solid
U.S., air force	Genie	9.5	825	6	rocket, solid
U.S., navy	Phoenix	—13	840	> 10	rocket, solid
U.S., navy, air force	Sidewinder 1*C*	9.5	185	10	rocket, solid
U.S., navy	Sparrow 3	12	450	12	rocket, solid
U.S., air force	Super Falcon	7	140	5	rocket, solid

MILITARY SPACE VEHICLES

Military applications of space flight are discussed in this section. For nonmilitary applications see SPACE EXPLORATION.

Following the launch of Sputnik 1 in 1957 by the Soviet Union, there was public concern that nuclear bombs might be placed in orbit for an indefinite period to be called down at will. Such concern is uninformed, however, since a bomb in orbit follows a fixed path. Hitting a planned target may require a waiting period of hours or days. On short notice such a bomb might not be able to strike a useful target. Further, a bomb system in orbit is subject to malfunction, and space maintenance is still a matter for the future. A bomb in orbit may be precisely located and therefore is subject to interception by another space power. Thus ground-based or submarine-based missile systems are capable of greater command and control, and continual testing and maintenance of silo-launched ICBM's is feasible. For these reasons the U.S.

and Soviet Union experienced no great difficulty in agreeing to a treaty to ban weapons of mass destruction from Earth orbit and from other celestial bodies.

The general uses of Earth orbital flight for military purposes are reconnaissance and surveillance; satellite inspection and interception; for securing communications; and as part of the fractional orbit bombardment system (FOBS).

The United States Air Force and United States Navy military satellites are launched by Scout, Thrust-Augmented Delta (TAD), Thor-Agena, Atlas-Agena, and Titan IIIC launch vehicles.

Soviet space launch vehicles

The Soviet military space program, as in the case of the U.S., has relied primarily on modified military ballistic missiles for space launch vehicles. Thus the Sapwood, Sandal, Skean, and Scarp have been used for more than 400 Earth orbital launches. All of the Soviet military launches are included in their Cosmos series program.

In 1968 and again in 1970 the Soviet Union apparently launched "inspector–destructor" satellite vehicles. In each case, a spacecraft was manoeuvred in orbit to the vicinity of a previously launched Cosmos satellite, then moved away, after which an explosion into many bits of debris occurred. From these observations it is generally conceded that the Soviets possess at least a rudimentary capability of satellite inspection and destruction.

FOBS is an alternative system for conventional ICBM trajectory. Instead of a trajectory on the shortest great circle path, the Soviet Union has launched test vehicles into orbit and caused the re-entry vehicle to land in the Soviet Union just before the end of the first orbit. Such a strategic ballistic missile system would have the advantage of nullifying the United States Ballistic Missile Early Warning System (BMEWS) radar watch, which "looks" north from Alaska, Greenland, and England. The Soviet SS 9, Scarp, has flown several flights that would seem to indicate interest in such a system. Inherent in FOBS would be a certain loss in accuracy because of the greater distance flown as well as a smaller warhead. The U.S. does not have an FOBS program.

DRONES, DECOYS, AND TEST VEHICLES

This category contains a variety of special-purpose vehicles closely related to guided-missile systems.

Drones. A drone is basically a pilotless aircraft, usually under radio-command guidance and widely employed as a target for anti-aircraft weapons. Powered by propeller or by turbojet, ramjet, or rocket engines, a large number of these manoeuvrable vehicles have been developed, with operating velocities as high as four times the speed of sound. Recovery, if the drone is not destroyed, is effected by controlled landing or by parachute.

Drones in reconnaissance

Another application for drones is military reconnaissance. Equipped with photographic or television cameras, the drone may be shifted from radio control to programmed inertial guidance and flown over targets at a relatively low speed and low altitude under operationally hazardous conditions.

Other applications of drones are electronic communications, surveillance, antisubmarine warfare (detection equipment and depth charges are carried by a drone helicopter), and even the rescue of downed airmen at sea.

An example of a drone target vehicle is the United States Army Roadrunner. This 25-foot, swept-wing, unmanned aircraft is powered by a ramjet engine. Boosted to flight velocity by a solid-propellant rocket, the Roadrunner can fly 30–40 minutes at speeds as great as mach 1.5. It flies at low altitude and is used to train personnel to operate Hawk surface-to-air missiles and to evaluate performance of other defense missiles.

Decoys. The decoy is a device designed to simulate by electromagnetic radiation an aircraft or ballistic-missile warhead. The purpose is to divert defensive anti-aircraft or antimissile fire. Thus a relatively small decoy can be made to appear, by electromagnetic techniques, as large as a bomber on radar screens. Or it can emit infrared and other radiation to simulate a re-entering ICBM warhead. For military reasons, details of such devices are not released to the general public.

Test vehicles. The term test vehicle covers a variety of special-purpose rockets and missiles designed to produce data on components or design features that will ultimately be used in much more expensive missiles and rockets. They must be scaled-down versions of the eventual rocket or missile. After World War II, for example, many aerodynamic shapes were tested at subsonic and supersonic velocities under dynamic flight conditions by affixing them to the front of solid-propellant rockets. Data were obtained by optical tracking and telemetry. In the development of a satisfactory design for ballistic missile re-entry vehicles, a five-stage rocket vehicle carried a scale model of the test shape to high altitudes; the final stage was then fired toward the Earth, achieving the re-entry velocity of an ICBM. In Project Fire, the National Aeronautics and Space Administration (NASA) in a similar manner achieved velocities equivalent to those of a lunar vehicle returning to Earth. Another special test vehicle is the Little Joe solid-propellant booster. It was developed to launch the Mercury and (a larger version) the Apollo spacecraft to an altitude at which the parachute recovery system could be tested.

The ingenious design and use of such special-purpose vehicles built from available components can reduce costs significantly, confirm design concepts, and save time.

GUIDANCE AND CONTROL

Most free-flight rockets require stabilization of some sort to minimize flight-path deflection caused by wind, nonuniformity of rocket structure, rocket-jet misalignment, and other factors.

Methods of stabilization

In the case of some barrage-type rockets launched by the thousands, dispersion errors may be accepted because of the overlapping explosive effect in the target area. Aircraft rockets are often stabilized by fixed fins located at the rear of the rocket. Folding fins that open by inertia after firing are also used. Short-range bazooka rockets and ballistic rockets of five- to ten-mile range are usually fin stabilized. Spin stabilization is used on some rockets. In these designs, the rocket nozzle is replaced by a series of smaller nozzles, canted at an angle to impart torque as well as thrust.

A unique method of wire control, based upon the German air-to-air X-4 rocket, has been used. This technique is utilized in some antitank rockets of about one-mile range. Fine wire is trailed from a pair of bobbins on the fins of the rocket, and command signals are given by the operator observing its flight through a telescope.

If a rocket is designed to operate at high altitudes, as in the case of sounding rockets, satellite launchers, and ballistic missiles of ranges greater than 100 miles, aerodynamic forces are no longer available for control because of the low density of the air. One technique, employed first by Goddard, is to place carbon or molybdenum deflection vanes within the rocket exhaust. Usually two pairs are employed for corrections about all three axes; one pair for pitch, the other for yaw and, in opposition, for roll. In flight, deviations from flight path are sensed by gyros within an automatic control (autopilot) system, and corrective signals are sent to motors that operate the vanes. This deflection system may be used also to program the tilt from vertical launch to ballistic flight path. Radio command also may be used to initiate the tilt program.

A modification of the jet vane method is the "jetevator," a ring-shaped deflector mounted at the nozzle periphery that rotates slightly into the edge of the rocket jet as required. Gimbal mounting of the engine, permitting the motor to swivel a few degrees in any direction, is widely used for flight-path control. This system was developed for the U.S. Viking and has been used successfully on IRBM'S and ICBM's.

Injection of high-pressure gas in the rocket nozzle to create a local shock wave and differential pressure is a further technique of jet deflection in use. This technique is known as thrust vector control (TVC). Magnetohydrodynamic (MHD) deflection of a rocket jet is another possible flight-control concept; since a rocket jet may be highly ionized, it is conceivable that a magnetic field may induce deflection of the exhaust gases without physical contact.

Roll control (*i.e.*, rotation about the long axis) is sometimes required, particularly on larger ballistic missiles. Small jets mounted transversely on the side of the missile are commonly used for this purpose.

Details of guidance systems in modern missiles are, of course, secret, particularly frequencies of operation, since electronic "jamming" or confusion by deliberate transmission of spurious signals by the enemy may be possible. Pulse-cdded directional signals (see TELEMETRY) have been used to avoid jamming. Some missiles carry their own small radar systems and are launched when the missile is electrically "locked-on" to the target. A ship may utiiize its large radar to track the target aircraft or vessel and continuously compute the range and bearing of the target while sending flight-correction signals to the missile in flight. Some air-to-surface missiles use television guidance, maintaining the image on the scanning screen in the same relative position after launching.

Sometimes the target may emit a signal that a missile can sense and track. Some missiles can lock on to the exhaust jet of an enemy aircraft, others on enemy radar transmitting systems.

The Snark was an early United States postwar intercontinental guided missile that utilized stellar navigation, comparing the position of stars and correcting an on-board inertial guidance platform. Powered by turbojet engines, this missile became operational in 1958 but was superseded by the Atlas ICBM. Long-range ballistic missiles have a flight time of about 30 minutes. The guidance system for these missiles is invariably inertial, using gyros and accelerometers to sense variations in velocity in each of the three axes: roll, pitch, and yaw. Having sensed a variation, the system gives correction to the flight controls to place the missile back on the programmed trajectory. Travelling along a preplanned launch path, the final-stage rocket engine is shut down at the precise moment when required velocity (about 15,000 miles per hour) and direction are achieved. The warhead and final stage continue to travel upward and follow a ballistic trajectory to the target.

Inertial guidance in ICBMS

WARHEADS

Rocket and missile warheads are conveniently divided into three categories: (1) high-explosive, (2) nuclear, or atomic, and (3) special-purpose. The first of these is generally employed on short-range tactical weapons. While it is possible to design such warheads to produce damage by either concussion (blast) or fragmentation, most depend upon some form of fragmentation. The efficiency of high-explosive warheads is a function of the size, number, weight, and velocity of the fragments produced, and the reliability of the warhead is largely a function of its fuzing system. The spray pattern of the fragments is also important and is established by the geometry of the high-explosive charge and the arrangement of the fragments upon it. Nonfragmenting high-explosive warheads for antitank weapons usually incorporate the shaped charge principle to achieve maximum penetration by concentrating the blast at one point. Nuclear warheads are used primarily, but not exclusively, on the IRBM'S and ICBM's, since it would be uneconomical and inefficient to use high-explosive heads on these long-range missiles. Typical of such missiles are the U.S. Titan, Minuteman, and Polaris. Very long range air-to-surface missiles (particularly the so-called standoff bombs), some air-to-air missiles, as well as certain short-range tactical or battlefield rockets and missiles also may employ nuclear warheads. The nuclear warheads may be either fission or fusion types; in the latter case they are called, popularly, H-bombs. The power of nuclear warheads is indicated by comparing the release of energy of a certain weapon to an equivalent weight of TNT. Thus a ten-kiloton warhead or bomb has the same explosive force as 10,000 tons of TNT, a five-megaton device has the force of 5,000,000 tons of TNT (see also NUCLEAR WEAPONS).

Because of the development of antiballistic missiles and tracking radar that can locate and identify an incoming warhead, several further refinements of warhead systems should be noted. One concept is a rocket-powered re-entry vehicle that would cause the warhead to change course to a target. By the mid-1960s, as a result of the development of antiballistic missile (ABM) sites around Moscow and other cities, the United States pressed the development of MIRV—multiple independently targeted re-entry vehicles. This system embodies a multiheaded ICBM, one that has three separately targeted nuclear warheads sent on their independent ways after the main propulsion stages of the ICBM have shut down. Installation of MIRV warheads on Minuteman III ICBM's is reported to have commenced in May 1971. U.S. Navy Poseidon fleet ballistic missiles are also scheduled to carry MIRV warheads.

Development of MIRV

III. Rocket-propulsion systems

PRINCIPLES OF ROCKET PROPULSION

Jet propulsion—of which the rocket is one type—is based upon reaction of a body to the rearward thrust of a jet of gas. The physical principle involved was set forth by Sir Isaac Newton in 1687. Newton's third law of motion states, in its simplest form, that for every action there is an equal and opposite reaction.

A simple example of reaction propulsion is the brief flight of a toy balloon after it has been filled with air and released. The forward motion of the balloon results from the rearward expulsion of air from the balloon. The thrust does not result, as sometimes erroneously presumed, from the jet pushing against the surrounding air. The force imparted by the jet is equivalent to the product of the mass flow of the air and the velocity of the jet expanded to atmospheric pressure; see equation (2), below. In the case of a chemical-propellant rocket, the jet is composed of the gaseous combustion products of the propellant mixture, which is burned inside a thrust (combustion) chamber and ejected at supersonic velocity through a nozzle. The gas velocity in a properly designed converging–diverging nozzle is sonic at the throat (most narrow) portion of the nozzle and becomes supersonic as it travels through the diverging (exhaust) end of the nozzle.

The concept of propulsive force

The concept of propulsive force (thrust) and specific impulse is best explained by the aid of simple mathematics. The propulsive force, or thrust, is basically equal to the momentum of the gas, which in turn is equal to the product of the weight of the exhaust gas and its velocity. In mathematical terms, this propulsive force is expressed as

$$F = \frac{\dot{W}}{g} \times V_e \qquad (1)$$

in which F represents thrust in pounds, W the propellant flow rate in pounds per second, g the acceleration of gravity in feet per second per second, and V_e the nozzle exhaust velocity in feet per second. For greater precision, another term must be added; the difference between the exit and ambient pressures multiplied by the nozzle exit area. The complete equation, then, becomes

$$F = \frac{W}{g} \times V_e + (P_e - P_o)A_e, \qquad (2)$$

in which P_e is the exit pressure, P_o ambient pressure, and A, the nozzle exit area. Since P_o decreases with altitude and is equal to zero in a perfect vacuum or in outer space, it may be seen that rocket engine performance increases with altitude. Thrust ratings of upper-stage rocket engines are often given both at sea level and at operating altitudes; the latter rating is always higher. Effective exhaust velocity c may be expressed

$$c = \frac{Fg}{\dot{W}} = V_e + \frac{P_e - P_o}{\dot{W}} A_e g. \qquad (3)$$

Inspection of equation (1) shows that thrust can be increased by increasing the jet velocity or the mass flow of propellant gases, or both.

Limitations on mass flow of propellant gases arise from such design considerations as maximum temperatures and pressures feasible within permissible engine weights.

Since the kinetic energy in a rocket jet is derived from conversion of the chemical energy of combustion to directed kinetic energy, high thermochemical energy per unit weight of propellant is desired. In practice, the kinetic energy of the rocket exhaust jet may be only 40 to 70 percent of the theoretical heat energy from combustion of the propellant. Minor losses arise from incomplete combustion and heat lost to the thrust chamber walls. Larger efficiency losses come from unavailable thermal energy that leaves the exhaust nozzle.

An important rocket term is specific impulse, I,,, which is the thrust per pound per second of propellent burned:

$$I_{sp} = \frac{F}{\dot{W}} \frac{\text{lb}}{\text{lb/sec}}. \qquad (4)$$

Similarly, from equation (3),

$$I_{,,} = \frac{c}{g} \frac{\text{ft/sec}}{\text{ft/sec}^2}. \qquad (5)$$

Since the pound and foot units cancel out, I,, is given in units of seconds.

Further calculations show that specific impulse is increased by increased chamber pressure and combustion chamber temperature and lowering of molecular weight of exhaust gases.

As a rocket-powered vehicle moves, the weight (mass) of the vehicle continuously decreases as the propellant is consumed and disappears to the rear as exhaust jet. In the design of efficient long-range rockets, every effort is made to reduce structural weight to a minimum and increase to a maximum the percentage weight of the propellant. In the case of the V-2, 69 percent of the takeoff weight was propellants. Modem lower-stage boosters have values as high as 94 percent. Takeoff acceleration may vary from several g's to a few tenths of a *g* in large rockets taking off vertically.

The technique of staging

The final velocity of a rocket-powered vehicle can be increased, theoretically, to any desired value by the technique of staging; *i.e.*, setting a series of rockets one on top of the other and firing them successively. Thus each successive stage is smaller and commences firing at a higher initial velocity. In practice, however, the complications of duplicating mechanical items, cost, and reduction in reliability have set a practical limit to the number of stages that are employed.

The Bumper-WACprogram conducted in the U.S. in 1948–50 is an example of the benefit of staging. A WAC/Corporal rocket placed on top of a V-2 was fired after burnout of the V-2. Whereas a V-2 alone achieves a maximum altitude of about 100 miles and a maximum velocity of 3,500 miles per hour, the Bumper-WACsent the WAC/Corporal to a velocity of 5,000 miles per hour and a peak altitude of 250 miles. All long-range ballistic missiles and space launch vehicles are staged. In the case of the Saturn 5, the S-1C (first stage) reaches a cutoff velocity at 5,350 miles per hour and separates; the S-2 (second stage) brings the vehicle to 14,750 miles per hour and separates; the S-4B (third stage) delivers thrust until an orbital velocity (at 103 nautical miles) of about 16,600 miles per hour is reached and shuts off (see also SPACE EXPLORATION).

A rocket-propelled vehicle is highly inefficient at low speeds. Efficiency increases from zero at rest to a maximum when the velocity of the vehicle is the same as the jet exhaust velocity.

A rocket motor or thrust chamber assembly is composed of a combustion chamber and nozzle (see Figure 7). The injector head is considered a part of the combustion chamber in the case of a liquid-propellant motor. The term rocket engine (liquid-propellant only) refers to the motor plus the associated propellant and feed system, propellant lines, valves, regulators, mounting lugs, igniter, etc. A rocket power plant refers to the complete propulsion system, including propellant tankage, pressurizing system. gimbal mounting (or jetevators, vanes, etc.), tankage-level sensing devices, and associated computers.

Conventional rocket motors burn chemical propellants, either solid or liquid. Combustion of the propellants provides the hot gas, which is exhausted in a jet through a convergent–divergent nozzle to the rear. Other types of rocket motors employ hybrid systems (liquid propellant burned in a solid-propellant core) and air turborockets, which utilize some ram air for part of the combustion process. Two nonchemical types of rocket propulsion systems are nuclear and electrical. These systems have received serious attention only since World War II and have particular application for space flight missions.

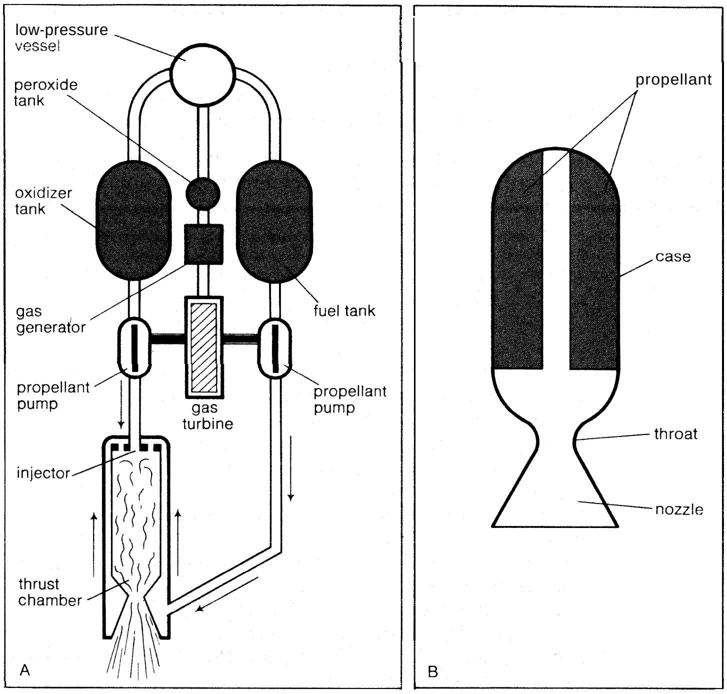


Figure 7: Representative design of (A) liquid-propellant, regeneratively cooled rocket power plant and (B) solid-propellant rocket motor.

SOLID-PROPELLANT ROCKET MOTORS

Characteristics. The outstanding feature of solid-propellant motors is their relative simplicity of configuration and state of readiness for use. All of the propellant is contained in the combustion (or thrust) chamber, to which is attached an exhaust nozzle. An electrical or pyrotechnic igniter fires the propellant charge.

Disadvantages arise from a somewhat lower specific impulse range at sea level compared with liquid-propellant motors, performance variation caused by storage temperature, and heavier rocket casing to contain combustion gases at a typical pressure of 1,000 pounds per square inch.

The design configuration and chemical composition of solid-propellant charges vary widely. The charge (or grain) may burn from one end only, as does a cigarette. Or the grain may be a hollow cylinder or have a star-shaped interior burning surface. Important elements in solid-propellant charge design are surface area and the burning rate of particular propellant mixtures.

Generally speaking, a solid-propellant rocket motor may be designed to wide performance specifications, with individual motors operating from a fraction of a second to 30 seconds or more. Current motor designs range from a few pounds to more than 1,000,000 pounds. Although modern motors are generally made of high-strength alloy steel of tubular construction, glass-fibre-wound motor cases are also used.

The largest solid-propellant rocket motors in use in the United States are the two strap-on booster motors ten feet in diameter and 86 feet long, on the air force Titan IIIC space launch vehicle. The thrust of each of these two rocket motors is 1,200,000 pounds. Each motor weighs 500,000 pounds.

Solid propellants. Various formulations of black powder (gunpowder) were the only source of propellant for rockets until nitroglycerin was discovered late in the 19th century. Saltpetre (KNO,) was the oxidizer in this mixture, while the sulfur (S) and charcoal (C) served as fuel. This exothermic (heat-releasing) reaction may be written approximately

$$2KNO_3 + S + 3C \rightarrow K_2S + 3CO_2 + N_2$$
$$+ \text{ 550 calories per gram of mixture reacting.}$$

Specific impulse from black powder

In a rocket motor, black powder yields a specific impulse (I,,) of 50–70 seconds, depending upon chamber pressure and formulation (typical, by weight: 75 percent KNO_3, 15 percent C, and 10 percent S). The flame temperature of gunpowder is about 1,500"–3,000" F (800"–1,600" C), and the mixture burns smoothly even below combustion chamber pressures of 100 pounds per square inch. The volume of gas produced is about 400 times the original volume of the charge.

With the advent of nitro explosives there arose new possibilities of obtaining high-temperature gases smoothly and reproducibly. Motor design changes were required to accommodate the higher combustion temperatures and pressures. Rocket propellants based upon nitrocellulose (guncotton) are known as single-base propellants. Mixtures of nitrocellulose and nitroglycerin are known as double-base propellants. These double-base mixtures are similar to smokeless powders for firearms. To improve the physical and chemical properties of double-base propellants, small amounts of additives are usually present as stabilizers. These additives prevent decomposition in storage or improve combustion characteristics and bond the propellant to the motor casing.

Double-base propellants require a minimum operating chamber pressure of about 500 pounds per square inch for smooth burning. Lower pressure results in irregular burning and oscillatory combustion. Flame temperature is of the order of 5,000" F (3,000" C). Heat yield is of the order of 850–1,200 calories per gram. Specific impulse is about 180–210 seconds. The volume of gases produced from double-base propellants is about 1,500 times initial propellant volume.

Solid propellants are considered to be divided into two main classes: "homogeneous" is applied to solid-propellant mixtures in which the propellant or mixture of propellants is intimately associated. Single- and double-base propellants are examples of homogeneous propellants. In composite (or heterogeneous) propellants, the substances, although finely ground, are in distinctly separate phases. An example of composite propellant is gunpowder.

Nitrocellulose or nitroglycerin is different from gunpowder with respect to the process of chemical combustion. The molecules of saltpetre, sulfur, and carbon in gunpowder must be intimately mixed in order to react. Nitrocellulose contains within each molecule sufficient fuel and oxidizer for complete reaction without the addition of other substances.

Advances in composite propellants

During World War II, a number of important new composite propellants were produced. One of these was the GALCIT series, which was a mixture of 75 percent potassium perchlorate and 25 percent asphalt oil. Another series consisted of the NDRC mixtures; a typical example of these was about 46 percent each of ammonium picrate and sodium nitrate mixed with 8 percent plastic resin binder.

Other composite mixtures use ammonium nitrate as oxidizer, and still others utilize synthetic rubber as fuel. An important advantage of composite propellants is that they can be cast directly in the case.

In the early 1970s both double-base and composite propellants were in operational use in missiles and space launch vehicles. In the composite mixtures, ammonium perchlorate was the standard oxidizer. The fuels (and binding agents) most used were polyvinyl chloride, polyurethane, and synthetic rubber, sometimes with finely divided aluminum as an additive.

LIQUID-PROPELLANTROCKET MOTORS

In this class of rocket motors, the liquid combustibles (*i.e.*, the propellants) are contained in tanks and are fed into the thrust chamber through an injector head by a propellant supply system. Most liquid-propellant rockets use two combustibles (bipropellant system), such as liquid oxygen and liquid hydrogen. Monopropellant systems that depend upon the exothermic (heat-r-leasing) decomposition of a substance, such as high-strength (90–95 percent) hydrogen peroxide, are in use. Such systems usually have lower performance ratings but are simpler to design and may be used for auxiliary propulsion systems.

The chief advantages of liquid-propellant rocket engines are the preciseness of control and restart capability that can be achieved. Also, liquid engines can be checked out, fired, and calibrated precisely before use. Further, the range of specific impulse is higher than for solid-propellant systems. Finally, only liquid-propellant motors are capable of burning for several minutes.

Elements of design of a liquid-propellant rocket engine include the combustion chamber, cooling techniques, exhaust nozzle, injector head, propellant supply system, propellant tankage, and the liquid propellants.

Combustion chambers. As in the case of all components of a rocket vehicle, a premium is placed on reduction of weight to the minimum adequate to perform the necessary function reliably. Combustion chambers are usually made of high-strength steel alloy and are usually cylindrical, although sometimes bell shaped. Solid-propellant motors are uncooled because the walls of the motor are protected from the combustion flame, since burning is from the core outward. Further, the duration of burning is relatively short. The length and volume of a liquid-propellant combustion chamber is related to the length of time for the propellants to mix and burn before passing through the nozzle.

Cooling techniques. Motors that operate for a fraction of a second to a few seconds only (such as in attitude-control systems) may be uncooled, radiating absorbed heat. Liquid rockets may be required to run five or more minutes, however. The conventional method of cooling is known as regenerative cooling.

Regenerative cooling

In this technique, one of the propellants flows along the outside of the nozzle and thrust chamber wall before it is injected into the motor. The heat that the propellant thus absorbs is not lost but is added to the heat of combustion in the thrust chamber. Design of the cooling flow passages is critical, since it is necessary to prevent boiling of the propellant and yet obtain a lightweight structure and low pressure drop. A high pressure drop results in greater design weight and higher pressure flow systems. Because the heat transfer rates to the walls are greatest in the narrow neck of the nozzle, the coolant flow rate is usually highest at this point.

In the early 1960s a new technique, ablative cooling, was developed, using materials similar to those employed as heat shields for re-entering space vehicles. Rocket chamber liners and entire chambers were made of these materials, which absorb and dissipate large amounts of heat as they vaporize and are lighter than comparable regeneratively cooled systems. They are also useful when propellant flow rate may be insufficient for regenerative cooling (*e.g.*, at low thrust levels) or when neither of the propellants has inherent sufficient cooling capacity to protect the rocket engine from heat damage.

Ablative-cooled motors have been used on both the descent and ascent stages of the Apollo Lunar Module.

Other cooling techniques that were tried in early developments were ceramic liners and film cooling, wherein a portion of one of the propellants was admitted through holes in the chamber providing a cooling film on the wall.

Exhaust *nozzle.* Exhaust nozzles are designed to expand the combustion gases to ambient pressure. In the case of a sounding or ballistic rocket, the ambient pressure decreases with altitude. For operation at high altitudes or in the vacuum of space, the diameter of the nozzle exit becomes very large and the construction weight prohibitively high. Thus nozzle design is usually subject to compromise, and certain reductions in efficiency (nozzle losses) are accepted.

Injector head. The injector head meters the propellants at a predetermined rate and mixture ratio and atomizes and combines the mixture within the combustion chamber so that burning takes place smoothly and completely. If the pressure drop across the injector is too great, an excessive burden is placed upon the propellant supply system and the weight of construction is increased unnecessarily. On the other hand, if the pressure drop is too low, oscillatory combustion will result. Good injector design reduces to a minimum the volume and length of a combustion chamber. A number of types of injectors have evolved. In the impinging spray type, propellant streams in pairs or clusters are injected to intersect at high velocity so that they break up into small droplets, evaporate, and burn. In the showerhead type, concentric rows of holes spray the propellant into the chamber. The rows of sprays may or may not impinge. Sometimes concentric slots are used to produce intersecting conical sheets of propellant sprays.

Impinging spray injectors

A World War II German design utilized a splash plate within the thrust chamber on which liquid-propellant streams played. Good mixing is obtained in this design, but hot spots may occur on the splash plate. Premix injectors have been tried wherein mixing is accomplished just prior to injection into the thrust chamber. Since rocket propellants by their nature are high-energy substances, this type of design is subject to explosion. Swirl-type sprays are another approach to injection configuration. Shop producibility, critical tolerances, and cost enter into injector head considerations, and some compromises in performances are usually accepted.

Often related to injector head dynamics is the phenomenon of oscillatory combustion instability in the combustion chamber. This is usually accompanied by audible effects of chugging and screaming. High-speed pressure differentials can destroy a rocket motor within seconds.

Because the mixture ratios of different propellants vary widely, a rocket motor is tailored for specific propellants, and it is not generally possible to operate an engine efficiently on propellants different from those for which it was designed.

Gas pressurization

Propellant supply *system.* The simplest method of forcing liquid propellants into the combustion chamber is by gas pressure. An inert gas is used, such as helium or nitrogen. Because the combustion chamber pressure is usually 300 pounds per square inch or more, the pressurizing gas must be higher to overcome frictional losses in propellant lines, valves, the thrust chamber cooling jacket, and the injector head. This high pressure necessitates heavy tanks. Gas pressurization is a simple and reliable system, however. For small vehicles helium gas at several thousand pounds pressure may be used, with operating pressure being reduced through a regulating valve. Another source of pressurizing gas is the reaction of small quantities of the propellants themselves in a special gas generator or even within the propellant tanks. In turbopump systems a source of high-velocity gas drives a turbine wheel, which in turn drives centrifugal pumps between the propellant tanks and the rocket motor. High-strength hydrogen peroxide (95–98 percent) is commonly decomposed by a catalyst bed within a gas generator. The decomposition products, oxygen and steam, are led through a nozzle and impinge on the turbine blades. The main propellants themselves may be used to provide turbopump power, simplifying the system.

The propellant pumps are usually centrifugal. They may be located on opposite ends of the turbine shaft or, in cases of large pumps, may be driven by a gear train. A certain pressure head is necessary to prevent cavitation (the formation of a partial vacuum at the blades). This is supplied from a lightweight high-pressure storage bottle through a pressure regulator to the propellant tanks. Other turbopump-drive systems utilize a high-pressure stored gas start cycle and then bleed a portion of one of the pumped propellants to an expansion jet to drive the turbine. This turbine gas exhaust may then be burned in the combustion chamber, utilizing its thermochemical heat content.

Valves for rocket engines offer design problems from the nature of the propellants and requirements for high reliability and precise operation. Many rocket flights have failed because of sticking valves. Valves are usually electrically, pneumatically, or hydraulically actuated.

Ignition of the propellants upon entry into the combustion chamber must be rapid to prevent a buildup of combustibles. The ignition of some propellant combinations such as Aerozine 50–nitrogen tetroxide is hypergolic; that is, it ignites spontaneously when the components are brought together. Otherwise, as in the case of hydrogen–oxygen systems, a source of flame or spark; or a hot wire or pyrotechnic igniter is required.

Propellant tankage. Tankage for propellants has evolved in recent years to integral systems, wherein the thin wall of the tank is the skin of the rocket vehicle itself. In addition to the weight of the propellant, rocket tankage must be able to withstand a certain amount of gas pressure for turbopump systems and several hundred pounds of pressure for gas-pressurized propellant supply systems.

Liquid propellants. The early rocket pioneers all recognized the advantages of liquid hydrogen and liquid oxygen as rocket propellants; however, such substances were in short supply and expensive. Goddard used liquid oxygen and commercial gasoline in his historic rocket test flight in 1926. The Germans used liquid oxygen and diluted alcohol in their V-2. Other German missiles used concentrated nitric acid and hydrocarbons such as vinyl isobutyl ether or a mixture of xylidine and triethylamine.

Common bipropellants

By the early 1970s the most common bipropellants in use in the United States were: liquid hydrogen and liquid oxygen (upper stages of Saturn 5); liquid oxygen and RP-1 (Atlas and first stage of Saturn 5); Aerozine 50 and nitrogen tetroxide (Titan II and Apollo Command, Service, and Lunar modules). RP-1 is a hydrocarbon similar to kerosene. Aerozine 50 is a 50–50 mixture of hydrazine and dimethylhydrazine.

Monopropellants in use in auxiliary rocket systems are high-strength hydrogen peroxide and hydrazine. Decomposition occurs by forcing the propellant through a catalyst of silver or platinum.

Considerable research and development effort has been expended on liquid propellants of higher energy, such as liquid fluorine and borohydrides.

Although experiments have been performed on thousands of fuels and a lesser number of oxidizers, no completely ideal propellants have emerged. Each propellant has some disadvantages that must be weighed against the particular design application and mission; for example, the problems of corrosivity and toxicity of fluorine or the low density of hydrogen may be accepted to obtain their high performance in space applications. Or the lower specific impulse of solid propellants may be accepted to obtain the advantages of readiness in antimissile defense systems.

In evaluation of liquid propellants the following properties are of engineering importance to the rocket designer: heat of reaction, average molecular weight of combustion products, stability (*e.g.*, to heat, shock), speed of reaction, ignition characteristics, density, viscosity, vapour pressure, specific heat, and corrosivity. For regenerative cooling—and all large, long-burning engines are so cooled—at least one of the propellants must have sufficient stability, specific heat capacity, thermal conductivity, and high saturation temperature to serve as a coolant. Despite its low temperature of —297" F (—183" C), for example, liquid oxygen is an unsatisfactory regenerative coolant. To improve cooling ability in the V-2, the ethyl alcohol fuel was diluted to 75 percent with water.

The logistic and handling qualities of liquid propellants must be taken into consideration. For bulk storage and transfer of propellants, corrosivity, stability, and vapour pressure are important, as are the freezing point and inflammability. Toxicity is important to personnel. Cost and bulk availability are also important considerations to program planners.

OTHER PROPULSION SYSTEMS

Nuclear propulsion systems. The development of reliable nuclear fission reactors has led to consideration of nuclear energy as a source of power for rockets. In this case the energy does not derive from the heat of combustion of a chemical reaction but from fission of nuclear

Design problems of nuclear systems

particles. Although the amount of energy potentially available is very much greater, the conversion to kinetic energy in a rocket exhaust jet is more complicated. Several nuclear rocket systems have been studied, including the rather audacious notion, suggested shortly after World War II, of obtaining impulse by a series of small nuclear explosions, an idea that was soon abandoned. The more conventional approach to a nuclear rocket is to use the heat of a fission reactor to heat a working fluid and expel the hot gas through a nozzle. Since the nuclear products are in a closed cycle, no radioactive particles are in the exhaust. The most obvious working fluid is low-molecular-weight hydrogen. The working fluid, or propellant, in this system would not be burned but simply heated and ejected. Some of the major problems associated with the design of a nuclear reactor are related to the design of an efficient heat exchange to transfer heat energy from the reactor to the propellant; cooling the thrust chamber walls; shutdown and restart; and nuclear radiation. The range of specific impulse (*i.e.*, pounds of thrust per pound per second of propellant flow) achievable with such nuclear rockets is estimated in the order of 700–1,000 seconds. Chemical propellants have a limitation of about 430 seconds at sea level. In the early 1970s work was proceeding on a flight-weight engine that would develop 75,000 pounds thrust and weigh 16,000 pounds.

Electrical propulsion systems. Whereas chemical-propellant rockets are characterized by high thrust for short durations, a number of electrical systems have been proposed that would yield a low thrust (in the order of a thousandth to a millionth of a pound) over a long period of time. Specific impulses of such systems range from about 1,000 to 10,000 seconds and more. Several approaches to electrical thrust systems are under development. All such systems, because of low thrust, must be carried to a low-altitude orbit by conventional rocket-powered systems. From this orbit these devices will operate in pulsed fashion or continuously for weeks, months, or even years in the case of distant interplanetary flight. Electric power to operate these thrust devices would be obtained from direct conversion of solar energy (solar batteries) or from nuclear electrical generators, as in a plasma thermocouple. Since specific impulses of 1,500 to 5,000 seconds are optimum for space missions from low-altitude Earth-satellite orbit (22,300 miles), additional flight time (of extra days or weeks) is not necessarily important. The advantages of using electrical propulsion are, however, truly significant. Earth takeoff weights may be reduced as much as one-half over the weight using chemical rocket propulsion throughout. In the particular case of a communications satellite payload, which could utilize the electrical power supply of the propulsion system, Earth takeoff weight might be reduced to one-third.

Among the electrical thrust devices under development is the electrothermal arc jet. This utilizes an electric arc to heat the propellant, or working fluid, which is expelled through a rocket nozzle to the rear. Another device studied accelerates the plasma, produced by the electric arc, by means of a magnetic field. Still other electric thrust devices operate in a pulsed fashion and utilize magnetohydrodynamics (MHD) to accelerate high-temperature gases.

Ion propulsion systems

Best known of the electrical thrust devices is the ion rocket. In this system electrostatic fields rather than electromagnetic fields are used to accelerate dust particles or positively charged ions such as metallic cesium. Specific impulse ranges of this device are estimated at 5,000 to 100,000 seconds. In 1971 two systems reported being investigated were a mercury bombardment engine (0.001–0.025 pound thrust) and a colloid accelerated system (8×10^6 pound thrust). Although in the early 1970s a few ballistic and orbital test flights had been made of rudimentary ion thrust motors, their operational use appeared to be some years in the future.

Interestingly enough, Robert H. Goddard conceived the use of accelerated charged particles for space propulsion and conducted related laboratory experiments as early as 1916–17, and Hermann Oberth independently proposed ion propulsion in 1929.

Future space propulsion systems. There are two other rocket systems proposed for space propulsion. One is the so-called hydrogen heater. In this concept, solar energy is concentrated by hemispherical reflectors onto a heat exchanger, which in turn heats hyrdogen to a high temperature; the hydrogen is then accelerated through a rocket nozzle to the rear. Finally, mention should be made of the photon rocket, wherein energy is converted to light and expelled as such. Such a system is wholly theoretical at this time, since the extremely high temperatures (nuclear fission range) required are not compatible with currently available materials.

TESTING

Cost of testing facilities

Many tests are required in the development of a rocket motor design. The term static test is used to describe operations wherein the motor is bolted to a test stand and fired. Static test stands provide for measurements of thrust, propellant flows, and a variety of significant temperatures and pressures. Control is maintained from within a reinforced concrete bunker or blockhouse. The rocket firing is viewed by means of mirrors or periscopes or at a distance through bulletproof glass. Closed-circuit television may also be used in the case of large motors. In liquid-motor development, so-called battleship (heavyweight) propellant tankage is employed, and gas pressure often is used to force the propellants into the thrust chamber. In initial chamber-design tests, water cooling provides a large safety factor over regenerative cooling. Flow measurement tests of injector heads and gas generator and turbine pump development usually proceed separately. Eventually all components are combined for system tests. When missiles are in the final stages of development, the entire vehicle is mounted in other test stands and fired. These are known as captive tests. The cost of rocket test facilities, with associated plumbing, tankage, propellant storage, and the necessary safety provisions of reinforced concrete and large safety distances, is naturally great. Because of the initial large cost of such facilities, particularly for large missiles and space launch vehicles, they represent a major investment and national asset of spacefaring nations.

AUXILIARY ROCKET SYSTEMS

In addition to the main propulsion systems of sounding rockets, guided missiles, and space launch vehicles, there are a wide variety of other rocket motors used to provide thrust for specific purposes during flight. These are called auxiliary, or secondary, propulsion systems.

At high altitudes and in space, small rocket motors (vernier motors) are used to provide attitude control about the three axes of flight: pitch, yaw, and roll. Two motors mounted on opposite sides of the missile can provide forces required to maintain correct flight path as directed by the inertial guidance system. Four quadrants of four motors each are installed on the Apollo Lunar Module for attitude control and manoeuvring. In the case of the Apollo Command Module a requirement exists for reliable, precise control of attitude to position the craft for atmospheric re-entry. Two sets of six rocket motors, flush-mounted on the command module, are provided. The thrust of each motor is 93 pounds. The second set of motors is redundant to provide for malfunction of the first system.

Ullage rockets

Under conditions of weightlessness between stage firings on launch or while in orbit, the liquid propellants float freely in their tankage. It is necessary before firing a rocket engine to have the outlet of the propellant tanks covered with fuel and oxidizer, respectively, so that continuous flow of propellant may occur. Thus an auxiliary rocket motor, known as an ullage rocket, is fired to impart a small amount of thrust to the space vehicle, causing the liquids to move to the lower end of the propellant tanks. Rocket motors that are used to provide spin (for flight stabilization) for missiles are usually of the solid-propellant type. Other small spin (or despin) motors are used on some satellites. Tiny rocket motors are sometimes used to control attitude of satellites operating automatically or upon command. Stored high-pressure inert gas is used occasionally for this purpose rather than more common monopropellants; *e.g.*, hydrogen peroxide or hydrazine. A small rocket motor may be used to separate a satellite from the final stage of the launch vehicle.

IV. Applications of rocket power

The role of rockets in space exploration is treated in SPACE EXPLORATION. In addition to the space applications, and the military applications described above, rockets have been used for a wide variety of other purposes.

SOUNDING ROCKETS

Atmospheric sounding--or vertical probe—rockets have become an important research tool for scientific investigation of the upper atmosphere. Despite their short lifetime, as compared with satellites, they have valuable characteristics. They are relatively much cheaper and simpler in design, and they offer flexibility in making meteorological measurements at times and locations desired by the scientist. Of foremost importance, however, is the fact that sounding rockets offer the only means of lifting scientific instruments to altitudes of 50–100 miles—higher than balloons or aircraft can reach and below the altitude at which satellites can function, because of atmospheric drag. Also, these versatile vehicles are used for operational checkout of instruments designed for use in research satellites. Sounding rockets have investigated such phenomena as cosmic rays, solar ultraviolet radiation and X-rays, auroral particles, stellar astronomy, and micrometeorites.

Before 1946 all upper-atmosphere research had been conducted from balloon sondes or aircraft. Beginning in 1946, various kinds of physical sensors were carried aloft in the nose cones of captured and reassembled V-2s to heights above 100 miles. Recovery of the instruments and data was by parachute after separation from the rocket.

More than 60 V-2s were launched, including the Bumper-WAC,a V-2 fitted with a second-stage WAC/Corporal rocket, which reached a 250-mile altitude in 1949. The WAC/Corporal was a liquid-propellant research rocket developed by the Theodore von Kármán research team at the California Institute of Technology near the end of World War II. The WAC/Corporal was 16 feet long; the rocket motor developed 1,500 pounds of thrust for 45 seconds. Propellants were red-fuming nitric acid and aniline mixed with 20 percent furfuryl alcohol.

Features of the Viking sounding rocket

By 1947 the United States Naval Research Laboratory commenced development of the Viking sounding rocket. First launch of this upper-atmosphere research rocket was in September 1949. A single stage vehicle, 41–47 feet long and weighing 9,650 pounds, it was powered by a 20,000-pound-thrust liquid oxygen–alcohol rocket motor. The fully gimbaled motor control system was unique. Another innovation was the integral propellant tank construction, in which the tank walls were the skin of the rocket. Payload weight was 400–500 pounds. Twelve Vikings were built, one reaching a 158-mile altitude in 1955. Some Vikings were launched from the USS "Norton Sound." Others were fired from White Sands Proving Ground in New Mexico.

Another sounding rocket of this period was the Aerobee, a Navy Bureau of Ordnance program. Nearly 19 feet long, this single-stage sounding rocket was an improvement of the WAC/Corporal design and could carry a 150-pound payload to an altitude of 72 miles (116 kilometres).

An exciting period of research developed with these new scientific tools, despite the technical problems associated with the developing rocket technology. Cosmic radiation measurements were made; photographs of the Earth were taken from space; mice were photographed under conditions of weightlessness; and the chemical composition of the upper atmosphere was studied.

Typical of experiments performed by sounding rockets is one begun in 1954 that determined regions of intense turbulence and strong shear winds below the altitude of 60 miles and extremely high winds above that altitude. These conditions were observed by Nike Asp sounding rockets that released trails of sodium vapour into the atmosphere between 60 and 120 miles above the Earth. Temperature measurements at high altitudes are made by timing the arrival of the sound of exploding grenades ejected from sounding rockets at altitudes between 38 and 65 miles. By the early 1970s much of the knowledge of the temperature and pressure of the atmosphere, as well as the composition of the ionosphere, was based on data from sounding rocket experiments.

Sounding rockets range in size, performance, and cost from a simple, single-stage, solid-propellant rocket such as the 66-pound Arcas, which can lift a 12-pound meteorological payload 37 miles, to the two-stage, solid-propellant, 11,500-pound Astrobee 1500, which can lift a 50-pound payload to 2,000 miles.

French rockets in IGY

Liquid-propellant sounding rockets are also used extensively. The French Vdronique rockets proved valuable during and after the International Geophysical Year (IGY). First flight-tested in 1952, these rockets were subsequently flown at the French missile test range at Colomb-Béchar, Algeria. In its IGY flights the VCronique used an 8,820-pound-thrust engine burning nitric oxide and turpentine. The rocket weighed 2,962 pounds and could lift a payload of 166 pounds to an altitude of 140 miles. The VCronique was 24 feet long and fin stabilized. A later version, the Véronique 61, was 30 feet long, weighed 4,310 pounds, and could lift a 440-pound payload to an altitude of 162 miles (261 kilometres).

Cooperative worldwide efforts during the International Years of the Quiet Sun resulted in the 1960s in several dozen coordinated rocket soundings of the upper atmosphere, which yielded new scientific understanding of solar effects upon the Earth.

During the solar eclipse of 1970, 31 sounding rockets were launched to obtain upper-atmosphere data on effects of the eclipse. Nearly half of these rockets were launched during a 20-minute period (see also METEOROLOGICAL MEASUREMENT).

Representative sounding rockets are listed in Table 5.

Table 5: Representative Sounding Rockets

	name	payload (lb)	altitude (mi)
Argentina	Rigel	66	135
Australia	HAD	20	80
Australia	HAT	35	47
Australia	Aero High	45	130
Canada	Black Brant III B	112	146
Canada	Black Brant IV	84	575
Canada	Black Brant V B	300	240
France	VCronique 61	440	162
France	Vesta	1,100	227
France	Emma	11	43
France	Tibere	740	1,200
France	Dauphin	287	93
France	Eridan	287	265
France	Dragon	66	435
West Germany	Project 621	50	47
Great Britain	Skylark	600	200
Great Britain	Skua 2	12	65
Great Britain	Petrel	30	95
Japan	S-300	120	100
Japan	K-8	175	125
Poland	Meteor 2K	22	62
Sweden	SR-1	45	110
U.S.	Nike Apache	50	150
U.S.	Judi-Dart	2	38
U.S.	Argo D-4	100	434
U.S.	Arcas	12	37
U.S.	Archer	40	90
U.S.	Iris	100	200
U.S.	Metroc	3	20
U.S.	Phoenix I	5–20	265–170
U.S.	Aerobee 150	125	186
U.S.	Aerobee 350	600	202
U.S.	Sandhawk	200	110
U.S.	Kangaroo	12	76
U.S.	Astrobee 1500	50	2,000
U.S.S.R.	V2A	4,840	131
U.S.S.R.	V5V	2,860	317

AIRCRAFT PROPULSION

The first aircraft to fly on rocket propulsion alone was a glider built by Fritz von Opel of Germany; it flew on September 30, 1929. Black-powder rockets were used for launching and a brief sustained flight. The first aircraft powered by a liquid-propellant rocket engine was a Heinkel design, in 1937. During World War II the Messerschmitt Me 163 rocket-powered interceptor became operational. It was powered by a hydrogen peroxide engine and achieved speeds of more than 600 miles per hour.

First supersonic flight

After the war the United States built the first supersonic research aircraft, the Bell X-1. It was powered by a four-barrelled, 6,000-pound-thrust, pressurized, liquid oxygen–alcohol engine. Each of the four 1,500-pound-thrust rocket motors could be fired independently or in combination. Launching was from a mother B-29 aircraft, which carried the X-1 underneath. Supersonic flight was achieved October 14, 1947. A later version, the X-1A, had a turbopump rocket engine, carried more propellant, and reached 1,650 miles per hour at 90,000 feet in 1953. Meanwhile, the U.S. Navy, in cooperation with the National Advisory Committee for Aeronautics (NACA) developed the Douglas D-558. This research aircraft reached mach 2 (twice the speed of sound). The Bell X-2 design, with a 15,000-pound-thrust engine, reached mach 3.3 and 126,000 feet. The North American X-15, with a 50,000-pound-thrust engine, flew a new aircraft speed and altitude record—mach 6.7 and 67 miles. Much valuable knowledge and hypersonic flight data were obtained in this program in 199 flights flown by three aircraft.

In postwar France and Great Britain, consideration was given to rocket-powered interceptor aircraft. Although a number of experimental aircraft were built and flown, no nation is known to have military rocket-powered aircraft in operational use.

OTHER APPLICATIONS

During World War II both the United States and Germany developed jet-assisted takeoff units for aircraft. Both solid and liquid propellants enabled land planes and seaplanes to take off with a shorter run or with increased loads, or both. Development continued after the war, particularly in the United States, where the units were also produced for the commercial market. Jet-assisted takeoff units are sometimes designed as a fixed installation in aircraft, but otherwise are intended for parachute drop after takeoff.

Rockets have been used to propel supersonic sleds along a twin-rail track. These sleds serve as beds for acceleration tests of various components such as parachutes, aircraft ejection seats, and nose cones. Important aeromedical tests on men have also been performed with rocket sleds. Accelerations as high as 100 gravity (*g*), decelerations of 150 *g*, and velocities up to 4,600 feet per second are possible. Both liquid- and solid-propellant rockets are used. Braking is accomplished by a parachute or, more often, by extending a scoop beneath the sled into a trough of water between the track rails.

Buried land mines were attacked in World War II by rocket-propelled devices. One of these devices, known as the Infantry Snake, pulled a 100-foot train of skis, linked together and carrying TNT or explosive-filled pipes known as Bangalore torpedoes.

A unique device based upon a German development was the Wedge, also called Donnerkeil. It used a powerful rocket to drive a six-foot-long rod, one inch in diameter, into the ground. By pulling out the rod and inserting and detonating a length of cord filled with explosive (primacord), a hole 10 to 12 inches in diameter was created. Since the hole made would accommodate a telephone pole, it was sometimes called a portable posthole.

Other wartime rocket-propelled devices included smoke bombs that, in barrage firing, could swiftly erect a smoke screen. Antiradar rockets, used in the Allied invasion of Normandy, released metal foil strips, called window, at the peak of their trajectories. These strips caused spurious radar reflections and appeared as swarms of aircraft on radar screens. Grapnels and rope ladders have been boosted over heights by rockets in commando operations. Rockets designed to carry thin cables aloft as anti-aircraft devices and mortar-fired rockets have been developed.

Solid rocket propellants have been used for turbojet engine starters, to furnish gas pressure to power small gyroscopes or electric turbogenerators for guided missiles, and to pressurize flame throwers. Solid-propellant charges have been fired in specially designed tools in oil wells to create high pressures that fracture the earth and increase oil production. Rocket jets have been used to drill holes through earth and rock. Rocket motors provide a high-temperature jet that is useful for materials tests, such as those carried out in the development of ballistic missile nose cones.

The possibility of delivering troops and equipment accurately and at long ranges with long-range ballistic missiles has been proposed. Even rocket boost of individual soldiers for hundreds of yards has been achieved by the use of a device known as a rocket belt.

Rockets have been attached to automobiles, boats, gliders, iceboats, motorcycles, railcars, and even to a man on ice skates. These experiments, many of which were carried out in the 1920s and 1930s, were generally more ingenious and publicity-seeking than efficient. Nevertheless, rocket motor technology, an extremely broad field of engineering involving chemistry, physics, materials technology, and fluid dynamics, has considerable promise in the future of transportation.

BIBLIOGRAPHY

General Reference: K. GATLAND, ***Missiles and Rockets*** (1975); R. PRETTY (ed.), ***Jane's Pocket Book of Missiles,*** new ed. (1978); and M.J.H. TAYLOR, ***Missiles of the World,*** 3rd ed. (1980), provide worldwide coverage of rockets and missiles. M.S. KNAACK, ***Encyclopedia of U.S. Air Force Aircraft and Missile Systems*** (1978–), published by the U.S. Office of Air Force History, is still in progress.

History: W. LEY, ***Rockets, Missiles, and Men in Space*** (1968); and W. VON BRAUN and F.I. ORDWAY III, ***History of Rockets and Space Travel,*** 3rd rev. ed. (1975), are standard histories. Also by Von Braun and Ordway, ***The Rocket's Red Glare*** (1976), is a nontechnical, narrative history. D. BAKER, ***The Rocket*** (1978), is a more technical, topical account. E.C. GODDARD and G.E. PENDRAY (eds.), ***The Papers of Robert H. Goddard*** (1970), includes his research reports to the Smithsonian Institution. W. DORNBERGER, ***V2, der Schuss ins Weltall*** (1952; Eng. trans., *V-2,* 1954); and E. KLEE and O. MERK, ***Damals in Peenemünde*** (1963; Eng. trans., ***The Birth of the Missile,*** 1965), tell the story of rocker development at Peenemünde. J.E. BURCHARD, ***Rockets, Guns and Targets*** (1948), contains accounts of U.S. rocket development during World War II. J.L. CHAPMAN, ***Atlas: The Story of a Missile*** (1960); J. BAAR and W.E. HOWARD, ***Polaris!*** (1960); and E.G. SCHWIEBERT, ***A History of the U.S. Air Force Ballistic Missiles*** (1964), are accounts of postwar ballistic missile developments. M.W. ROSEN, ***Viking Rocket Story*** (1955), relates the development of this significant postwar sounding rocket. K.W. GATLAND, ***Spacecraft and Boosters*** (1964), provides details of postwar developments in the United States and the Soviet Union; M. STOIKO, ***Soviet Rocketry*** (1970), chronicles Soviet rocket development. For current technical information, see publications of the American Institute of Aeronautics and Astronautics, New York City.

(F.C.D.III)

Rock Magnetism

The phenomenon of magnetism was first discovered in rocks. Early inhabitants of the Middle East and China knew of deposits of lodestone, an iron-bearing rock that attracted pieces of similar rock and that pointed toward the north and south if freely suspended. These unusual characteristics gave rise to myths about the heavens, the polar regions, and special lodestone mountains that were not dispelled until the 17th century.

It was also recognized that an unusual "field" existed in the vicinity of magnetic rocks. Iron objects or iron-bearing rocks that were unmagnetized would themselves acquire magnetism when placed in this field, and in many cases they would remain magnetized after removal from the original field. Magnetism, which is created in a substance by the presence of a magnetic field and which is retained after the original field is removed, is called ***remanent magnetism;*** most of the permanent magnets used in modern technology are made in this general manner. Magnets can also be made from coils carrying electric currents; these types are referred to as electromagnets. The main magnetic field of the earth is believed to result from the circulation of electric currents near the earth's core. The crust of the earth is thought to be free of such currents; it appears to have been magnetized by the main field of the earth.

The strength and direction of the remanent magnetism of crustal rocks are related to the strength and direction of the ambient field that induced the magnetic state. Accordingly, the remanent magnetism of rocks can be used to study the magnetic field that originally magnetized them. Some rocks, especially those that are iron bearing, are more strongly magnetized than are others in the same ambient field. This results from differences in the chemical or mineralogical composition of the materials. Temperature plays a key role because at elevated temperatures, the remanent magnetism characteristics will be markedly altered. Upon cooling in a magnetic field, an entirely new set of characteristics may be acquired.

Because magnetism is a quantity posseasing both direction and strength, precise information about the orientation of the field that caused remanent magnetism can be obtained. For example, if the direction of the inducing field is known and a different direction is measured in a rock outcrop, it is possible to reconstruct the motion through which the rock has been mechanically turned since it was magnetized. **A** major interest in rock magnetism is the determination of such rotations by assuming a knowledge of the original-field direction or inferring the direction of the original field by assuming that the rotation is negligible.

Advances in the technology of measuring the remanent magnetic properties of earth materials have permitted earth scientists to work with materials other than those commonly thought of as magnetic rocks. Although very weakly magnetized, many sedimentary rocks and soft sediments can be measured with sensitive instruments.

This article treats the ways in which rocks acquire magnetism, the determination of this property, and some interesting aspects of paleomagnetism, such as sea-floor spreading. For further information on this latter point, see CONTINENTAL DRIFT; and SEA-FLOOR SPREADING. See also MAGNETISM; EARTH, MAGNETIC FIELD OF; EARTH, STRUCTURE AND COMPOSITION OF; and ROCKS, PHYSICAL PROPERTIES OF for discussions of relevant information on magnetism in the earth and in the rocks of the earth's crust.

MAGNETIZATION OF ROCKS

Kinds of magnetization and magnetic materials

The magnetic moment (product of pole strength and distance between poles) of rock material is caused by the orbital and intrinsic spin motion of electrons in constituent atoms. In most materials there is little interaction between the magnetic moments of adjacent atoms. The vector directions of the moment are randomly distributed so that a specimen of material has no net magnetic moment in the absence of an external field. When an external field (H) is applied, however, there will be a weak interaction between the neighbouring atoms, and at this instant vectors will align themselves along the axis of the external field. If the alignment reinforces the field, the material is said to exhibit paramagnetic susceptibility. The magnitude of the susceptibility k (a measure of the ease of magnetization) is on the order of one-millionth of the external field (10^{-6}). If the alignment opposes the field, the susceptibility is diamagnetic. The common rock-forming minerals quartz and feldspar are diamagnetic materials. Pyroxene, biotite, and amphibole are paramagnetic.

In certain other materials, there may be additional effects that give rise to very strong magnetic interactions of neighbouring atoms. For these materials the crystal lattice of the structure is such that atoms with very strong intrinsic magnetic moments (particularly iron) are in a spatial configuration that allows them to become aligned permanently, relative to the crystal lattice. This ordering is due to a very powerful intramolecular magnetic field and results in spontaneous magnetization of the material. The region of the material over which this ordering occurs is called a magnetic domain and is on the order of 10^{-4} centimetres. The energy required to produce the ordering is termed the exchange energy. If the magnetic moment vectors in a domain are all parallel to one another, the material is ferromagnetic. Antiferrornagnetism refers to an arrangement of vectors in two equal antiparallel sublattices so that the net magnetization is zero. If the sublattices are antiparallel but not of equal magnitude, giving rise to net magnetization, the material is termed ferrimagnetic. If the sublattices are of equal magnitude but not exactly antiparallel, the material exhibits parasitic antiferromagnetism. The minerals iron, pyrrhotite, magnetite, and hematite, respectively, are examples of each of these types of domain orderings.

Intensity and components of magnetization

The magnetic moment per unit volume of material is called the strength or intensity of magnetization. It is a vector quantity and usually expressed by the letter J. The total magnetization J may contain both induced I and remanent components $\boldsymbol{R}$. These add vectorally to give $J = \boldsymbol{R} + I$ The induced magnetization is dependent on the direction and strength of the ambient magnetic field and the susceptibility k of the material. The value of k is a function of the mineralogical composition of the material. This relation can be expressed as: $I = \mathrm{k}H$. $\boldsymbol{R}$ is dependent on the above and on the thermal, chemical, and physical history of the material. A specimen that already has a remanent magnetization may acquire a new magnetization, This new magnetization may not erase the original but may add to it, vectorally, to produce a new total magnetization. Thus if R_{A} and I_{A} are the original remanent and induced components and a new remanent component R_{B} is added to it, the resultant total magnetization will be

$$J_{\mathrm{AB}} = R_{\mathrm{A}} + R_{\mathrm{B}} + I_{\mathrm{A}}.$$

If a new magnetization I_{B} is simply induced by an ambient magnetic field H_{B} that is different from the original ambient field H_{A}, then the new magnetization will be

$$I_{\mathrm{B}} = \mathrm{k}H_{\mathrm{B}},$$

and the total magnetization will then be given by

$$Jan = R_{\mathrm{A}} + In,$$

in which the I_{A} is lost and replaced by **In.** A portion of the magnetism induced in this fashion may or may not be retained as a remanent magnetization R_{B}.

A rock from the surface of the earth may have several remanent component magnetizations such as R_{A}, **Rn,** R_{C}, etc. It is possible by various laboratory techniques to remove the magnetization attributable to certain of these components one at a time. This selective removal of magnetic components is termed demagnetization. It can be accomplished by heating the rock or exposing it to strong alternating fields. The values of magnetization found in nature vary from the smallest value that instruments can measure, about 10^{-7} electromagnetic unit per cubic centimetre (emu/cc) for limestones and shales, to about 10^{-2} electomagnetic unit per cubic centimetre for lava flows and iron ores.

Iron oxide or sulfide minerals commonly exhibit properties of ferromagnetism and can be classified into two types. The first type includes the more strongly magnetized cubic oxide minerals: magnetite, maghemite, and the titanomagnetite series. The weakly magnetic rhombohedra] minerals, hematite, ilmenite, and pyrrhotite, constitute the second type.

The major parts of most rocks do not possess remanent magnetism, but in localized regions certain magnetic minerals may form small grains. In magnetite-bearing rocks the grains are usually large enough to constitute domain clusters. For rocks containing hematite, the grains are usually of such small size that each grain represents a single domain that is easily disordered by thermal agitation. Although these rocks may appear to be strongly magnetized in the presence of an applied magnetic field, they show low magnetic remanence when removed from the field.

Temperature and magnetization

The rocks most strongly magnetized are those of igneous origin, namely those rocks thrust up to the earth's surface in a molten state. Volcanic extrusions of lava and intrusive dikes (tabular bodies that penetrate pre-existing rocks and structures) are examples. These materials contain relatively large amounts of iron oxide minerals and are commonly formed at temperatures greater than about 1,000"

C. This temperature is much above the Curie temperature of the magnetite minerals (575" C), above which ferrimagnetic minerals lose their remanent magnetic characteristics. Hematite loses its remanence above *675"* C. This is called the Neel temperature in antiferromagnetic minerals. Upon cooling at the earth's surface these minerals become strongly magnetized in the direction of the ambient earth's magnetic field. The temperature at which the rock acquires its magnetization upon cooling in a magnetic field is the blocking temperature. The blocking temperature depends largely on the size and shape of the magnetic domains. Such rocks are said to have acquired thermoremanent magnetization (TRM) because the magnetization is retained after the temperature falls to that of the surface environment and the field is removed. This magnetization is very stable and subsequent exposure of rocks with TRM to magnetic fields several orders of magnitude stronger than the magnetizing field cannot appreciably change the original magnetization.

Acquisition of magnetization

There are several ways in which a specimen may acquire magnetization in an ambient magnetic field, these include:

1. Thermoremaneut magnetization (TRM) is as described above.

2. Isothermal remanent magnetization (IRM) is a weak magnetization acquired in a strong magnetic field at constant low temperature, much below the Curie temperature, and in only a matter of minutes. This magnetization is considered soft (less stable) as compared to TRM.

3. Chemical remanent magnetization (CRM) is acquired in a weak magnetic field when new magnetic mineral grains are formed during such chemical reactions as oxidation and dehydration. These reactions occur at low temperatures much below the Curie temperature. The CRM is much like TRM in its strength and stability characteristics.

4. Viscous remanent magnetization (VRM) is acquired at constant low temperature in a weak magnetic field over a time period on the order of millions of years. It is a relatively strong and hard magnetization compared to IRM. The regular thermal agitation of the magnetic domains eventually orders the mineral grains to produce a coherent magnetization.

5. Detrital remanent magnetization (DRM) arises when small magnetic grains that already possess remanent magnetization fall through the earth's atmosphere or through water and are subjected to the earth's ambient magnetic field. The grains align themselves to produce an appreciable magnetic effect that is quite stable.

Less frequently encountered types of magnetization include piezoremanent magnetization (PRM), which arises when stress is applied to a specimen in a magnetic field; inverse-type thermoremanent magnetization (ITRM), due to anomalous changes in crystal structure with temperature in the presence of a magnetic field; and anhysteretic remanent magnetization (ARM), which sometimes arises in rock specimens when an external alternating magnetic field is decreased from a maximum value to zero in the presence of a constant magnetic field. In addition many materials acquire a strong magnetization when they are struck by lightening. It is important to note that an existing magnetic field must always be present when a rock is magnetized, regardless of the method.

The total magnetization of a rock may be due to the sum of any or all of these types of remanent magnetization and to a nonremanent (induced) magnetization as well. The ratio of the strength of remanent to induced magnetization is called the (Koningsberger ratio) Q. This ratio has been found to be nearly zero for some sedimentary rocks and characteristically high, approaching 100, for certain marine lava rocks.

The primary or original magnetization of most igneous rocks is generally considered to be TRM. The magnetization of sedimentary rocks is more complex. Certain sediments are believed to contain both CRM and DRM components of primary magnetizations. The former probably results from chemical changes shortly after the sediments are deposited on the sea floor, whereas the latter is acquired as the sediment particles slowly accumulate. The primary magnetization is believed to be acquired in most rocks soon after the rock is formed. Determination of the age of a rock and its primary magnetization therefore permits inferences to be drawn about the ancient geomagnetic field. Such studies of paleomagnetism are of great significance in terms of elucidation of earth history.

Sampling and study of rocks

In the study of rock types on land, a number of practical problems arise. The first is the question of whether the rock formation has been reoriented since it was magnetized. Sedimentary strata may have been tilted or severely folded, for example, by geologic forces. The observed angles for the magnetization vector must be corrected for any postdepositional tilting to obtain information about the original magnetic field direction. If there have been intrusions of magma (molten material within the earth) into a pre-existing rock mass, it is possible that thermal and pressure effects (metamorphism) will alter the magnetic characteristics of the older mass. Such problems can be overcome by careful inspection of the sampling locality and by taking many samples in the general vicinity. It should be recognized that even when all necessary corrections are made, the primary magnetization direction of rocks generally does not agree with the earth's present field direction. It appears that deviation from the present earth's field increases with the age of a given rock.

The sampling of rocks and sediments from the ocean floor is associated with different problems. The chief difficulty in dredging hard rocks at sea is that the recovered rock sample may be part of the sea floor proper but also may be an erratic fragment transported to that locality by icebergs or other natural means. The age and geologic history of the rocks, therefore, are not usually known. Secondly, there is the problem of determining the orientation of the recovered specimen. New types of ocean floor rock drills have permitted some oriented samples to be obtained by a surface ship, and some research submarines carry experimental drills for taking oriented samples. Soft deep-sea sediment samples generally are obtained by a coring device, a hollow tube that penetrates the bottom and is then recovered with the entrapped sediment sample (Figure 1). There is usually some question about the azimuthal orientation of the coring tube and thus some uncertainty about the declination (horizontal

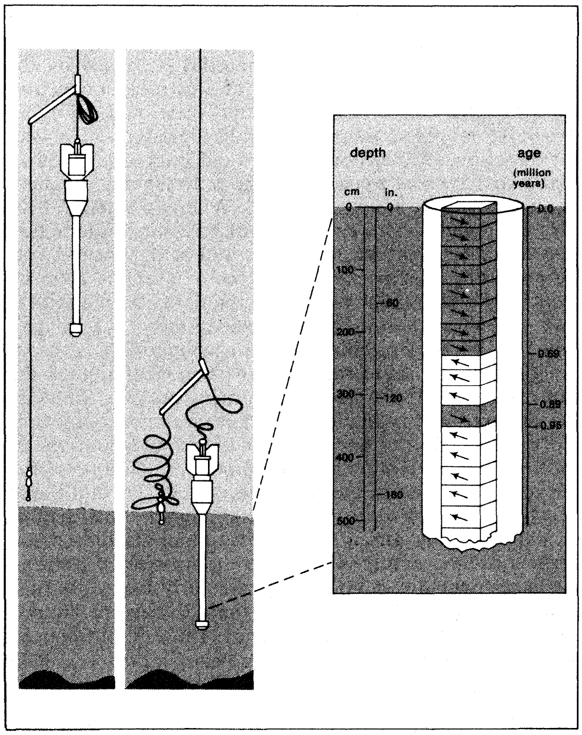


Figure 1: (Left) Recovery of ocean-floor-sediment column by deep-sea coring technique. (Right) Core of sediment. Arrows represent sediment magnetization vector directions. Dark and light cubes show expected sequence of normal and reversed magnetic polarity intervals as a function of depth beneath sea floor.

bearing) of the measured magnetization with respect to north. Fortunately, because the tube falls vertically downward in taking the sample, the inclination (departure from the horizontal plane) does not have to be corrected. Most sedimentary bedding planes on the ocean floor are horizontal or very nearly so. Sea-floor sediment samples also are usually weakly magnetized compared to the hard rocks and this constitutes another difficulty. The use of high sensitivity magnetometers has permitted measurement of the magnetic properties of rocks in recent years, however.

Measuring a rock's magnetism

The magnetization of a rock is measured in either of two ways. When an astatic magnetometer is used the specimen generally is positioned near a standard magnet system that is suspended vertically by a thin fibre. The specimen's external magnetic field causes the magnet to rotate, and the deflection of the magnet is noted, The specimen is then repositioned about a second axis and third axis; the three axes are oriented at right angles to each other. In each case the deflection is noted, and the magnetization is calculated from the three measurements.

In the second method, the sample is spun at high frequency near a signal pickup coil wound of fine wire, and the phase and amplitude of the induced voltage are noted. The sample is then repositioned with respect to the coil face and spun about a second and third set of axes. The magnetization can then be calculated by a procedure somewhat similar to that above. The instrument used is called a rock generator or spinner magnetometer.

THE EARTH'S PALEOMAGNETIC FIELD

The overall strength and direction of the earth's magnetic field can be expressed by its magnetic dipole moment. One hundred and fifty years ago the earth's dipole moment was 6 percent greater than at present. Fifteen hundred years ago it was 50 percent greater, but 5,500 years ago it was only about one-half of the present dipole moment. Various measurements made on samples as old as 500,000,000 years suggest that the field strength at that time was as little as one-quarter of its current value, although the variations through geologic time do not follow a very smooth curve.

In order to determine the strength of the paleomagnetic field, it is necessary first to measure the magnetic properties of an ancient rock sample. The sample is then heated above its Curie point to destroy its magnetism; finally it is cooled in a known magnetic field, and the new thermoremanent magnetization intensity is compared to the value it had before heating. If it is not the same the heating must be repeated followed by cooling in a different field. In this fashion, almost by trial and error, the strength of the original magnetizing paleomagnetic field is determined. The method is fraught with difficulties. Viscous magnetic effects and other irreversible changes, possibly of a chemical nature, cannot be accounted for. Nevertheless, the method has been applied to geologic samples, to the lining of kilns used by the Romans, and to hearths from early civilizations with apparent success.

Geomagnetic polarity reversals. Great strides have been made in understanding in detail how the earth's magnetic field has changed direction in the past. The most remarkable change in the field is the complete reversal of its polarity. It appears that within a relatively short period, probably less than 10,000 years, the field may reduce to a very small intensity and then re-establish itself with approximately equal intensity in exactly the opposite direction.

Naturally occurring rock sequences such as lava flows and sedimentary strata show sections that are reversely magnetized relative to the earth's present field direction. For many years it was believed that a rock could somehow automatically reverse its magnetization and, consequently, that there was little point in studying such an unstable characteristic of rocks. In 1960 it was indeed found that a special ilmenite–hematite mixture could become magnetized in a direction opposite to that of the applied field. Such mixtures are extremely rare in nature, and reversely magnetized rocks are not now believed to be due to any such cause. Rather the reversely magnetized specimens are thought to be due to their having been magnetized at a time in the geologic past when the earth's magnetic field had a polarity opposite from that which it has today. The idea of reversals of the earth's field was long considered hypothetical, but it is now accepted as fact by most geophysicists.

Volcanic rocks and the geomagnetic time scale

The most convincing evidence for polarity reversals of the earth's main magnetic field has come from recent studies of volcanic rocks collected from all parts of the world. Extensive field sampling programs have been carried out in the basalt flow regions of Iceland, Hawaii, California, southern Europe, East Africa, Australia, and India. Thick sheets of Pliocene and Pleistocene basaltic lavas exposed over large outcrop areas in these regions have been particularly useful. Such exposures yield rock samples that possess exceedingly stable remanent magnetism (TRM) in a flow sequence where the relative rock ages are clear.

The potassium–argon radioisotope decay technique is generally employed to determine absolute ages of rock specimens. Because the method is precise to about 3 percent of the age of the rock, a detailed chronology for the absolute age of lava flows extruded over the last 4,000,000–5,000,000 years has been determined. The time scale for geomagnetic polarity reversals can thus be inferred simply by noting the magnetic polarity of a lava flow and its absolute age. By analyzing a number of flows whose ages span appropriate time intervals, a catalog of age versus polarity can be developed. Because a single volcanic region generally contains lava flows extruded over a relatively short time interval, compared to the duration of a polarity state, it has been necessary to sample many volcanic regions to get a complete time se-

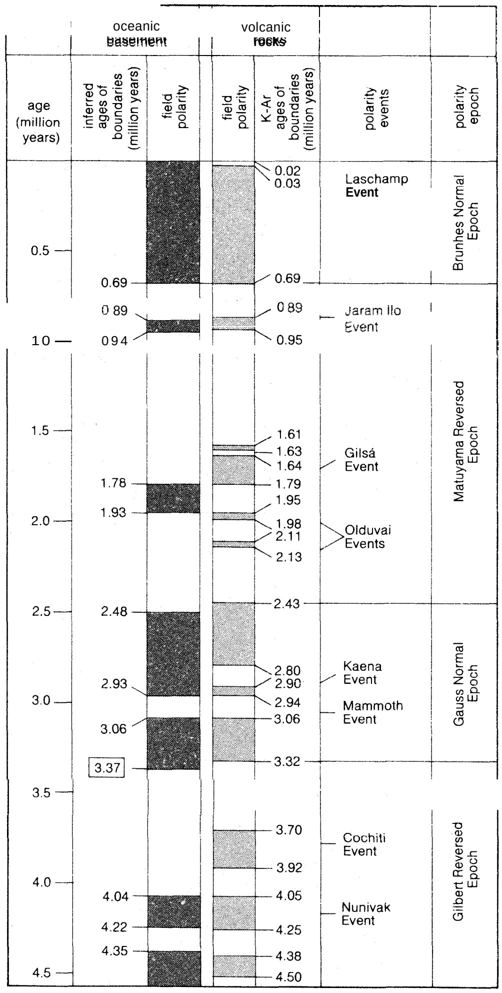


Figure 2: Geomagnetic polarity reversal time scales. Time scale determined from laboratory magnetic and radiometric age dating of volcanic lavas is at right. That from sea-floor-spreading interpretation of magnetic profiles across oceanic ridge crests is at left.

quence. Fortunately, nearly overlapping basalt flow chronologies, each spanning several polarity states, have been developed from the late Tertiary and Pleistocene Icelandic, Columbia River, and Hawaiian lava fields. Comparison of polarities measured from rocks of the same age collected in different parts of the world show remarkable agreement as would be expected if the geomagnetic field polarity does reverse itself through time. In fact, it is now possible to erect a detailed history of geomagnetic polarity reversals for the last 4,500,000 years. Figure 2 shows the sequence of normal and reversed states with the names assigned to the short events and the more comprehensive epochs. When the epochs were originally named and their dates established, they were thought to be long duration unipolar states. It was not known that they would contain several states of both polarities. Accordingly, the resulting system of naming epochs and events is partly redundant and somewhat ambiguous. Although several hypotheses related to changes in the electric-current pattern at the core–mantle interface have been advanced to account for the reversal phenomena, a satisfactory explanation is not available at this time.

Marine sediments and the geomagnetic time scale

The study of the magnetization of sediment layers from the sea floor yields the same sequence of polarity reversals as that found on land. In dating the subbottom horizons, each reversal in magnetization is related to a dated geomagnetic field reversal as determined from terrestrial volcanic rocks (Figure 1). The sedimentation rate can then be inferred and the sediment layer dated at various points on the sea floor. The coring technique, by which undisturbed bottom samples are obtained, has not permitted verification of the reversal time scale or the dating of material older than about 4,500,000 years. An interesting aspect of the reversal chronology observed in deep-sea cores is the apparent correlation of the extinction and first appearance of certain marine fossil organisms with reversal boundaries. It has been speculated that this is a causal relationship resulting from an increased influx of cosmic radiation during a polarity reversal. Such an influx may have accelerated evolutionary processes by causing extinctions and by increasing the rate of gene mutations. Much more work remains to be done in this area before definitive results will be available, however.

It is now believed that the field has had its present polarity for about the last 700,000 years, although there is some very recent evidence to suggest that one or a few very short periods of reversed polarity may have occurred within this 700,000-year period. Each such episode lasted no more than 10,000 to 20,000 years.

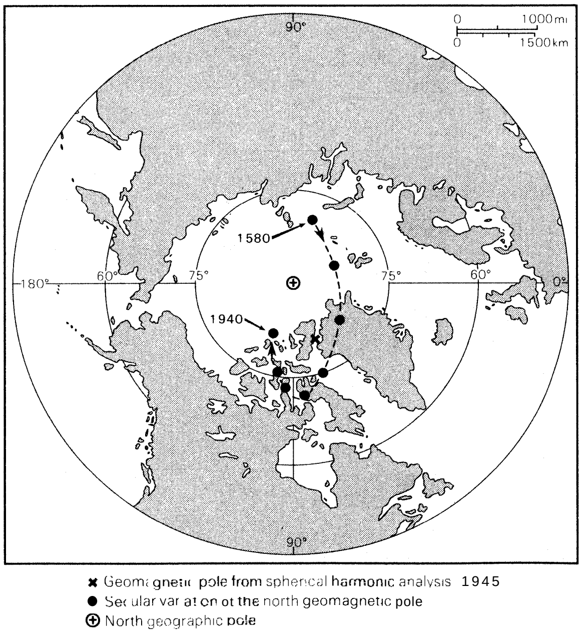

0 1000 mi
0 1500 km
90°
1580
180°
60°
75°
75°
60°
0°
1940
90°
✖ Geomagnetic pole from spherical harmonic analysis 1945
● Secular variation of the north geomagnetic pole
⊕ North geographic pole

Figure 3: Secular variation of the north geomagnetic pole since AD 1580.

Historical departure of the magnetic and geographic poles

Polar wandering. Another important aspect of directional changes of the field is the implied migration of the magnetic pole across the earth's surface. The recent changes appear to be rather smooth and predictable. Current magnetic charts used by navigators show yearly correction factors that must be applied to the declination. Over the last 300 years observatory records at Paris and London show that the position of the magnetic field pole has completed nearly half a circle of rotation about the earth's geographic or rotational pole. The field direction has changed by about 7° in inclination and 35° in declination (Figure 3). Studies on archaeological specimens indicate that over the last 2,000 years the declination and inclination at various sites in England may have varied by as much as 20 to 30 degrees to either side of their present values.

The general configuration of magnetic lines of force of the earth appears to have a westward drift, to judge by the recent observatory records. Whether this trend will continue is problematical, but the magnetic dipole axis, now inclined some 11° to the geographic axis, might be expected to move slowly about the earth so that its average from the geographic or rotational axis will be zero over a few thousand years, say 10,000 years.

On a somewhat larger time scale the declination indeed appears to average near zero over a few thousand years. For example, ocean sediments deposited at a rate of one centimetre in a few thousand years provide core specimens a few centimetres in dimension for measurements on spinner magnetometers. Each specimen thus spans a time interval of 1,000 to 10,000 years and in most cases shows a magnetization close to the present rotational pole. The misalignment of the present geomagnetic dipole and the rotational pole therefore is probably a transient situation on the recent end of the geologic time scale.

Polar departures during geological time

In contrast to these rather minor magnetic and geographic discrepancies of pole position, directional changes observed in rocks on the order of tens and hundreds of millions of years old show marked deviations from the present geographic pole. The remanent directions appear to be offset from the present field because of slow wandering of the magnetic pole with respect to the present rotational pole. The directions may also be different for another very important reason: the entire continent upon which the rock lies may have moved or drifted. Continental drift, once considered extremely controversial, is now accepted by most earth scientists. A most important observation that led to the acceptance of the theory was the smooth angular displacement of the paleomagnetic poles for rocks of different ages taken from the same continent. Also rocks of the same age from other continents showed a consistently different displacement path. By shifting the continents, the displacement paths could be superimposed, and similar geologic features along the margins of the two continents would also match. In any event, the resulting magnetic-pole path during geologic time does not match the present rotational pole. In fact, the indicated north magnetic pole appears to have moved slowly across the earth's surface from equatorial latitudes 200,000,000–300,000,000 years ago to its present position in the Arctic Ocean. Some earth scientists believe that the ancient rotational poles also have wandered over the earth's surface with the magnetic poles.

Magnetic anomalies over the continents and oceans. Variations in the magnetic properties of the earth's crust are commonly used to locate and define ore bodies of economic value. Iron ore deposits are particularly well suited to magnetic-exploration methods. Because the hard rock basement in most areas is more magnetic than the overlying sediments (a magnetization contrast on the order of 1,000 usually exists), the sharpness of the magnetic effects on the surface also gives an estimate of the depth of burial of the basement and information about the configuration of the basement. Quantitative procedures exist for obtaining the most likely geometry of buried structures, but it is never possible to determine from the magnetic data alone the exact configuration of the basement. The configurations of many magnetic bodies can give rise to the same anomalous values on the surface.

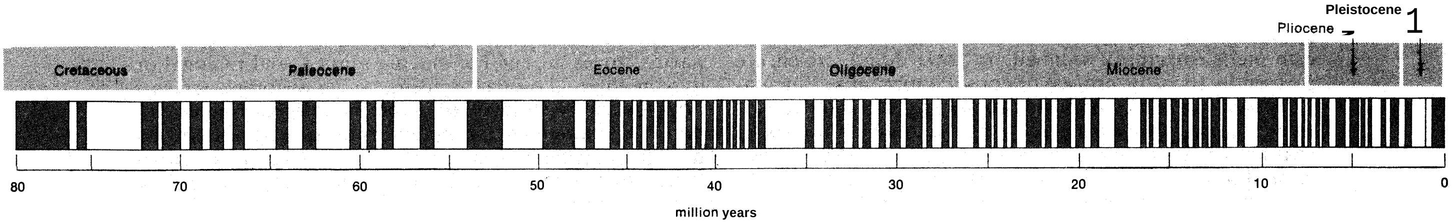


Figure 4: Geomagnetic polarity reversals time scale for the last 80,000,000 years as inferred from magnetic anomaly profiles across oceanic ridge crests. Black and white bands represent periods of normal and reversed polarity respectively.

Nevertheless, magnetic surveys are useful for estimating the depth of sedimentary basins.

Anomalies over oceanic ridges

An analysis of the spatial fluctuation of magnetic values shows that wavelengths of up to about 100 to 200 miles are due to the. earth's crustal features. Wavelengths greater than 1,000 to 2,000 miles are part of the earth's main field and arise at the core-mantle interface within the earth. There are relatively few wavelengths between 200 and 1,000 miles.

The pattern of anomalous magnetic values (magnetic anomalies) is, in general, quite irregular over the continents, the continental shelves, and marginal seas such as the North Sea and the Mediterranean Sea. Over much of the ocean, on the other hand, the pattern is more simplified and exhibits characteristic linear trends. Except for relatively isolated disturbances caused by volcanic seamounts, the ocean is dominated by linear magnetic anomalies that are related to the midocean-ridge system in all the oceans. Basaltic submarine lava flows near the sea floor are believed to be responsible for most oceanic anomalies. These rocks typically show a very high Q-ratio (ratio of remanence to induced magnetization).

The midocean-ridge system is an enormous mountain range that extends down the full length of the Atlantic Ocean, continues around Africa into the Indian Ocean, and, after branching, goes south of Australia and up the eastern part of the Pacific Ocean. The strength of the magnetic field is alternately anomalously high and low with increasing distance out from the axis of this ridge system. The anomalous features are almost symmetrically arranged on both sides of the axis and parallel the axis, creating bands of parallel anomalies.

Sea-floor spreading

According to the sea-floor-spreading theory, these linear anomaly trends are underlain by alternating bands of normally and reversely magnetized basaltic rocks on the sea floor. These rocks are thought to have been created at the axis of the midocean ridge, where they acquired a thermoremanent magnetization as they cooled. The polarity of their magnetization depended upon the polarity of the geomagnetic field at the time they cooled. It appears that the older rocks spread out to each side of the ridge as new igneous material continuously wells up along the ridge axis. Because the sea floor is thought to spread at a relatively constant rate, the history of geomagnetic field reversals is believed to be recorded in rocks of the ocean floor. If spreading rates are on the order of one to ten centimetres per year (10 to 100 kilometres per 1,000,000 years), then the history of reversals matches the geomagnetic polarity time scale determined from terrestrial volcanic rocks very closely. Observed magnetic profiles are calibrated by matching them with computer-simulated anomaly profiles formed by spreading during the Gauss epoch, 3,370,000 years ago. The spreading rate is considered constant throughout the last 4,500,000 years. A comparison of the polarity time scales determined by this method and the volcanic-rock method is shown in Figure 2.

When the different spreading rates in the oceans of the world are considered, the sequence of normal and reversely magnetized rocks is remarkably similar in all the ocean floors. Initially an arbitrary number was assigned to the anomalies for identification, but absolute dates now have been assigned to the magnetic anomalies; these signify both the age of the ocean floor and the date of the magnetic reversals of the field.

The assignment of ages to magnetic anomalies has been made by assuming that the particularly distinctive magnetic anomaly associated with the Gauss geomagnetic normal polarity epoch about 3,370,000 years ago can be recognized in the magnetic profiles. Measurement of the distance of this anomaly feature from ridge crests thus permits determination of the spreading rate over the last 3,370,000 years. By extrapolating this rate to the outer ridge flanks, anomalies older than 3,370,000 years can be dated. On the basis of profiles across the South Pacific and South Atlantic oceanic ridges, a time scale for the last 80,000,000 years, approximately, has been developed (Figure 4); its verification was provided by recent deep-sea drilling in the south Atlantic Ocean. The age of sea-floor lava rock obtained at varying distances from the ridge crests showed that the spreading rate inferred from magnetic profiles is essentially correct.

BIBLIOGRAPHY. Readable accounts written for the layman and nonspecialist include: R.M. BOZORTH, *Ferromagnetism*, p. 423–475 (1951), a basic text on the origin of ferromagnetic properties of materials; A. COX, R.R. DOELL, and G.B. DALRYMPLE, "Reversals of the Earth's Magnetic Field," *Scient. Am.*, 216:44–54 (1967), *an* excellent account of geomagnetic-field reversals and observations in volcanic rocks and a brief discussion of rock magnetism; J.R. HEIRTZLER, "Sea Floor Spreading," *Scient. Am.*, 219:60–70 (1968), a discussion of the origin of magnetic anomalies at sea and the motion of the sea floor; E. IRVING, ***Paleomagnetism and its Application to Geological and Geophysical Problems*** (1964), a broad overview of most phases of rock magnetism and paleomagnetism; J.H. NELSON, L. HURWITZ, and D.G. KNAPP, "Magnetism of the Earth," ***Publs. U.S. Cst. Geod. Surv. No. 40-41*** (1962), a brief presentation of basic information about the geomagnetic field; N.D. OPDYKE, "Paleomagnetism of Oceanic Cores," in R.A. PHINNEY (ed.), ***History of the Earth's Crust***, pp. 61–72 (1968), a brief account of the recent advance in studies of magnetic reversals in deep-sea sediments; A.D. RAFF, "Magnetism of the Ocean Floor," *Scient. Am.*, 205:146–156 (1961), a presentation of the magnetic anomaly characteristics over marine areas; s . ~RUNCORN, "Magnetization of Rocks," in ***Handbuch der Physik***, vol. 47, *Geophysics* I, pp. 470–497 (1956), a discussion of the magnetic properties of rock; F.D. STACEY, ***Physics oj the Earth***, pp. 125–191 (1969), a comprehensive treatment of the nature of the earth's magnetic field, rock magnetism, and paleomagnetism, directed primarily to the intermediate-level student; and F.J. VINE, "Magnetic Anomalies Associated with Mid-Ocean Ridges," also in ***History of the Earth's Crust***, pp. 73–89, a presentation of the sea-floor spreading interpretation of marine magnetic anomalies. Two works primarily for the specialist are A. COX and R.R. DOELL, "Review of Paleomagnetism," ***Bull. Geol. Soc. Am.***, 71:645–768 (1960), an excellent review of paleomagnetic principles and techniques; and T. NAGATA, *Rock Magnetism*, rev. ed. (1961), a classic text.

(J.R.H./J.D.P.)

Rock Metamorphism, Principles of

Metamorphism means change, and in petrology the term refers to the physical and chemical alteration of rocks that are subjected to the elevated temperatures and high mechanical stresses that occur in the earth's crust. This article describes the kinds and degrees of metamorphism that are known and the relations between the important causative variables and rock compositions and structures. For information on the specific rock types involved see METAMORPHIC ROCKS; for information on the conditions that lead to metamorphism see EARTH, STRUCTURE AND COMPOSITION OF and MOUNTAIN BUILDING PROCESSES; see GEOCHEMICAL EQUILIBRIA AT HIGH TEMPERATURES AND PRESSURES for the geochemistry involved.

Petrologists divide rocks into three broad classifications: igneous rocks (*q.v.*), those that have solidified from silicate melts (magmas); sedimentary rocks (*q.v.*), which are formed by the slow process of consolidation of sedimentary deposits; and metamorphic rocks (*q.v.*), which constitute a major part of the earth's crust and which are rocks that originally belonged to the other groups but that were altered in various ways. The alteration consists in part of chemical reactions among the minerals in the rocks or between incoming fluids and the original minerals. These changes in physical and chemical properties of rocks, under the new environmental pressures, transform them into new species. For example, clay sediments change to shales and slates at temperatures around 100°–200° C. At higher temperatures in the earth's crust, clay may change to mica schists or various types of gneisses. Limestones, originally formed as sedimentary rocks from the calcareous shells of primitive animals, change to marbles by the chemical reactions and recrystallization that occur during metamorphism.

Mechanical deformation of rocks is another common occurrence in metamorphism. This deformation is partly plastic; the rock yields more or less like metal under crustal stresses. These crustal stresses can sometimes crush rocks to fragments varying from pieces a foot long to fine powder (see ROCKS, PHYSICAL PROPERTIES OF).

Alterations of a physical and chemical nature are also caused by weathering (*q.v.*) of rocks in·contact with the atmosphere and surface water and by diagenesis of sediments, but such changes are generally not included in the term metamorphism. Weathering and diagenesis occur at temperatures below those at which true metamorphism takes place. On the other hand, temperatures above 600°–900" C, depending on rock type, cause silicate rocks to melt and thus the realm of magmatism and magmatic rocks is created. Rocks exposed to these extreme physical conditions are called ultrametamorphic.

METAMORPHIC ALTERATIONS

All kinds of rocks may be exposed to metamorphic alterations. Alterations without change of bulk chemical composition of the rocks are called isochemical metamorphism. If change of bulk chemical composition is involved, the term allochemical metamorphism applies.

Some chemical constituents are very mobile in rocks undergoing metamorphism. The content of mobile constituents such as water and carbon dioxide is rarely constant in a rock during metamorphic alterations. Thus, a gabbro that consists of two anhydrous minerals, plagioclase feldspar and pyroxene, picks up several percent of water if metamorphosed at low temperatures because hydrous minerals such as chlorite and hornblende are formed. On the other hand, a sedimentary rock containing hydrous clay minerals loses large amounts of water when metamorphosed at high temperatures (above 500" C), at which essentially anhydrous minerals become stabilized. Similarly, rocks may readily gain or lose carbon dioxide because at high temperatures, calcite and other carbonate minerals react with quartz or silicates to form calc-silicates and carbon dioxide gas, which escape from the altered rocks.

Rocks containing calc-silicates such as wollastonite, diopside, and anorthite often develop carbonates at low-temnerature metamoryhism. This means considerable introduction and gain of carbon dioxide. An example of reactions involving carbon dioxide is as follows:

low temperature		high temperature	
$CaCO_3$ +	$SiO_2 \rightleftharpoons$	$CaSiO_3$ +	CO,.
calcite	quartz	wollastonite	carbon dioxide

The right-hand side of this metamorphic mineral reaction corresponds to high temperature and loss of carbon dioxide, the left-hand side to low temperature and gain of carbon dioxide.

Many rock-making chemical constituents other than water and carbon dioxide are mobilized during metamorphism, thus making change of bulk chemical composition a common phenomenon.

Types of metamorphism. Metamorphism is conveniently classified as contact metamorphism, which occurs in narrow zones around hot magma masses, and regional or dynamothermal metamorphism, which encompasses vast complexes of rocks involved in the evolution of fold mountains. The latter type of metamorphism is not caused by heat from molten magmas but rather by tectonic, or structural, stresses in the earth's crust and a regional rise in crustal temperature. The reason for such regional rises in crustal temperature is not known for certain, but it is probable that it is caused by the dissipation of the potential energies that create fold mountains (see ROCK DEFORMATION).

Contact metamorphism. Rocks altered by contact metamorphism show little or no sign of strain such as plastic flowage or mylonitization or crushing of minerals. In contact metamorphic rocks, recrystallization (for example, formation of coarse-grained marble from limestone or quartzite from sandstone) and neomineralization (the formation of new mineral assemblages more stable than the old ones) are the most prominent features. Such rocks are generally called hornfels.

Regional metamorphism. Regional metamorphic rocks bear signs of mechanical strain in addition to recrystallization and neomineralization. The strain is seen as schistosity or foliation and other kinds of anisotropic structures in the rock bodies, much like the structures found in rolled metal plates and drawn wires. Crystalline schists such as mica schists and greenschists, many marbles, most gneisses, certain soapstones, amphibolites, and granulites are regional metamorphic rocks.

Progressive and regressive metamorphism. Rock alteration taking place at increasing temperature (and pressure) is called progressive metamorphism (for example, clay sediment altered to mica schist). Alteration at decreasing temperature is called regressive metamorphism or diaphoresis (for example, greenschist formed from basic lava).

Polymetamorphism. A polymetamorphic rock is one that has been exposed to two or more periods of metamorphism. Such rocks may carry proof of their complex history in the form of structural or mineralogical relics from earlier states in their evolution. The term palimpsest (from the Greek for scratched or scraped, that is, marked again) structure was coined for features of this kind.

GRADES OF METAMORPHISM

Metamorphic rocks provide an illustration of the fact, established by atomic science, that the concept of "dead matter" is an illusory one. Exposed to changing geological conditions in the lithosphere, rocks respond by mineralogical, structural, and chemical alterations and, thus, tend to develop mineral assemblages or recurring combinations and structures best suited—*i.e.*, most stable—under the conditions of temperature, pressure, and anisotropic stress that affect rocks during their evolution.

The principal processes taking place in rock bodies during metamorphism are recrystallization and neomineralization. referred to above. The change of mineral assemblages in response to varying temperature and pressure conditions makes it possible for geologists to distinguish among various grades or degrees of metamorphism. The different grades are determined by the occurrence of characteristic minerals or mineral assemblages that represent, and presumably are stable at, unlike pressure and temperature (P,T) conditions.

Grades and depths. The Swiss petrologist Ulrich Grubenmann (1850–1924), in his classification of crystalline schists, distinguished among three grades of metamorphism, the epi (upper), meso (middle), and kata (downward) zones. Grubenmann believed that depth of burial in the earth's crust was the chief reason for unlike grades of metamorphism because pressure obviously increases with depth and, generally, so does temperature, as shown by the geothermal gradient, or increase in temperature of the earth from the surface toward the interior, averaging about 3° C per 100 metres (see EARTH, HEAT FLOW IN).

Epizonal rocks were regarded as products of shallow depths because their mineral content was characteristic of low temperature (chlorite, talc, serpentine, and other hydrous minerals).

Mesozonal rocks carry minerals generally formed at moderate temperature (300" C to 400" C). Such rocks were consequently believed to have developed at intermediate depth in the crust.

Katazonal rocks consisting mostly of anhydrous minerals such as garnet, kyanite, hypersthene, cordierite, feldspar, etc., were considered ambassadors from the deepest portions of the earth's crust, now exposed by erosion.

Unlike grades and constant composition. G. Barrow, working in Scotland, demonstrated that the grade of regional metamorphism changed in horizontal direction across the Caledonian fold-mountain chain, which extends in a northeast-southwest direction through the Highlands. Zones of rocks exposed to the same grade of metamorphism are approximately parallel to the axes of the mountain chain, low-grade rocks generally occupying marginal zones and high-grade rocks central zones, with rocks of intermediate grade between. Similar conditions are encountered in other areas of more or less eroded fold-mountain chains, such as the Caledonian chain in Scandinavia and many Precambrian, deeply eroded mountain chains in Fennoscandia (the Baltic Shield), Canada, west Greenland, and other areas. Barrow chiefly studied alterations of argillaceous (clay-rich) sediments, which are very responsive to changing pressure and temperature. Certain index minerals were used to identify certain grades of metamorphism; one index mineral would change over to another along a more or less curved boundary line in the field. Such a line is termed an isograde and represents the intersection between the earth's surface and a boundary surface separating two neighbouring zones of metamorphism. Such boundary surfaces between adjacent but unlike zones of regional metamorphism are curved surfaces that often but not always show fold axes more or less parallel to the axis of the mountain chain. Chlorite, biotite, garnet, kyanite, staurolite, and sillimanite were minerals used by Barrow to indicate increasing grades of regional metamorphism.

Goldschmidt's studies

Constant grade and unlike compositions. V.M. Goldschmidt was the first to apply extensively physicochemical principles in a study of metamorphic rocks. His treatise of the Oslo (Kristiania) area in Norway has become classic. In this area numerous magmatic bodies of Permian age have intruded a sedmentary complex of sandstone, limestone, and argillaceous schists. These sediments have been exposed to high temperature in a contact zone of varying thickness around the intrusive magmatic rocks. At a constant distance out from the rim of a given magma body the temperature in the surrounding sediments has been constant. Therefore Goldschmidt was able to study in great detail how rocks of unlike bulk chemical composition (sandstone, limestone, claystone, etc.) responded to a given temperature as determined by the solidifying magma. In other words, Goldschmidt studied the effect of a given grade of metamorphism (the grade of contact metamorphism, or the pyroxene hornfels facies of P. Eskola, discussed below) on a number of rocks of unlike chemical composition. Barrow had studied the effect of unlike metamorphic grade on rocks of nearly constant composition. Later, P. Eskola combined both by studying the effect of various grades of metamorphism on rocks of various bulk composition.

Chemical equilibrium and the phase rule. The important Oslo-area work led to the conclusion that chemical equilibrium had been approached rather closely in sediments baked in the heat from intrusive magmas. Subsequent studies have shown that chemical equilibrium among minerals in rocks generally is attained also during regional metamorphism, probably more commonly than under contact metamorphism. The laws of chemical equilibrium in heterogeneous systems are consequently applicable to metamorphic rocks, a fact largely responsible for the recent progress in metamorphic petrology. Metamorphic rocks frequently obey the chemical phase rule: $P + f = C + 2$, in which P stands for phases, which in petrology correspond to minerals, f means degrees of freedom, and C components. In silicate rocks simple oxides are conveniently counted as components; if sulfides are present sulfur is an additional component.

The chemical phase rule

Under isochemical metamorphism, rocks are closed systems not permitting communication of chemical constituents with the environment. Two degrees of freedom have then generally prevailed inasmuch as pressure and temperature are independent variables, and the so-called mineralogical phase rule is valid: $P = C$. That is, a metamorphic rock formed under equilibrium conditions should have the same number of different minerals as it has components. It is noteworthy that metamorphic rocks very rarely have more minerals than components, counted in terms of simple oxides, such as those of potassium, calcium, magnesium, silicon, etc. Indeed, in most metamorphic rocks the number of minerals is fewer than the number of components. For example, many amphibolites, which are common metamorphic rocks in terrains of gneisses and crystalline schists, contain only the two minerals plagioclase and hornblende, yet chemical analyses may show ten or more simple oxides. The reason for this apparent discrepancy is in part that many minerals are mixed crystals in which several atoms or ions can substitute for one another in identical structural positions. Sodium and calcium are interchangeable in plagioclase and so are silicon and aluminum to some extent. In hornblende a great number of cations substitute for one another. Magnesium, iron, titanium, aluminum, manganese, nickel, and cobalt are all encountered in identical structural sites. This means that several unlike chemical elements must be counted as one component because they are identical from a mineral–chemical point of view.

The mineralogical phase rule

Mineral assemblages. The general approach to chemical equilibrium among minerals in metamorphic rocks results in an order and regularity that invite natural classifications of such rocks. Contact metamorphic rocks can, for example, be placed in phase diagrams of the type shown in Figure 1.

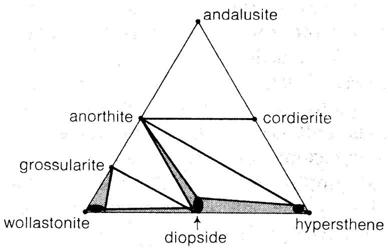


Figure 1: Mineral assemblages of contact metamorphic rocks.

This particular three-component diagram with Al_2SiO_5 (andalusite), $CaSiO_3$ (wollastonite) and $(Mg,Fe)SiO_3$ (hypersthene) at the corners is only applicable to rocks with excess silica metamorphosed in the contact zones around magmas. Any one of the possible mineral assemblages shown in Figure 1 may develop. There are, however, restrictions as to the number as well as types of associated minerals at equilibrium. If chemical analysis of a hornfels places it on a join line between two minerals in Figure 1, then only these two minerals are permitted at stable equilibrium. (Minerals containing elements not shown in the diagram may certainly occur, in addition to those indicated by the phase diagram.) A hornfels falling within one of the three angles in the figure must contain the three minerals at the corners. Only one mineral can be present stably in a hornfels with bulk chemical composition coinciding with that of the mineral itself. Incidentally, such monomineralic rocks are not as rare as one may assume in view of the small chance that bulk composition of sediments should coincide with that of a pure mineral. Pure diopside hornfels and wollastonite hornfels are described from various places.

The Oslo hornfels adjacent to the magmatic intrusives follow very strictly the rules of permitted mineral assemblages. Ten classes of hornfels have been recognized in the field:

Classes of hornfels

Class 1 rocks with andalusite and cordierite (albite);
Class 2 rocks with andalusite, cordierite, and anorthite;

Class 3 rocks with anorthite and cordierite;
Class 4 rocks with anorthite, cordierite, and enstatite;
Class 5 rocks with anorthite and enstatite;
Class *6* rocks with anorthite, enstatite, and diopside;
Class 7 rocks with anorthite and diopside;
Class 8 rocks with anorthite, diopside, and grossularite;
Class 9 rocks with grossularite and diopside;
Class 10 rocks with grossularite (vesuvianite), diopside, and wollastonite.

This sequence of classes of hornfels corresponds to increasing amounts of calcite (calcium) in an argillaceous sediment. Andalusite is the metamorphic product of the typical clay mineral kaolinite, and wollastonite forms at high temperature by calcite and silica interaction.

Metamorphic facies. At grades of metamorphism different from contact metamorphism other minerals and mineral associations occur. These assemblages can be shown in diagrams similar to Figure 1, with calcium, aluminum, and magnesium + iron silicates at the corners. Eight grades, each with its characteristic phase diagram (called ACF diagram), have been recognized and described in studies of areas in Finland and and Norway. The characteristic mineral association of each grade is called mineral facies or metamorphic facies, with the name of each facies chosen from the kind of rock most dominant within that grade. Basic rocks rich in calcium and magnesium are most sensitive to changes of metamorphic grades. The various facies are named after such rocks rather than acidic and alkali-rich ones. This does not mean, however, that acidic and alkaline rocks are not found in the unlike facies or grades; it only emphasizes that it is difficult to distinguish among unlike metamorphic grades in rocks such as granitic gneisses, quartzites, nepheline syenites, and the like. Quartz, potash feldspar, and acidic plagioclase, the chief minerals in such rocks, are stable over the whole pressure-temperature field in which most metamorphism takes place. Thus a granitic gneiss consisting of quartz, potash feldspar, and albite remains a granitic gneiss with the same minerals at all grades of common metamorphism. On the other hand, minerals occurring in basic rocks, being of magmatic or sedimentary origin, such as pyroxenes, hornblendes, chlorites, talc, serpentine, olivine, anorthite, and epidote, have narrow fields of stability. A basic rock of given bulk composition may, for example, yield a gabbro (pyroxene and plagioclase) at high metamorphic grade, an amphibolite (hornblende and plagioclase) at intermediate grade, and a greenschist (chlorite epidote and albite) at low grade.

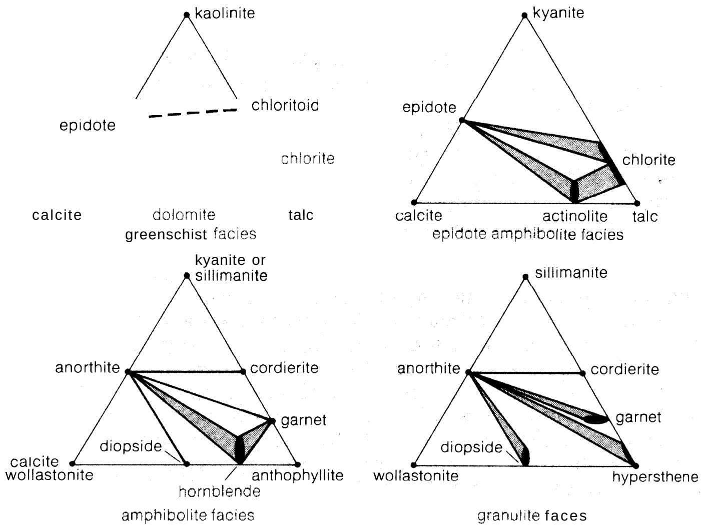


Figure 2: Various grades or facies of regional metamorphism.

The sequence of metamorphic grades from low to high grade of regional metamorphism is greenschist facies, epidote amphibolite facies, amphibolite facies, and granulite facies. Corresponding ACF diagrams are shown in Figure 2. Contact metamorphism belongs to the pyroxene hornfels facies.

Other facies less commonly encountered are eclogite facies, glaucophane schist facies and sanidinite facies. Figure 3 shows the approximate positions of the various facies in a pressure-temperature diagram.

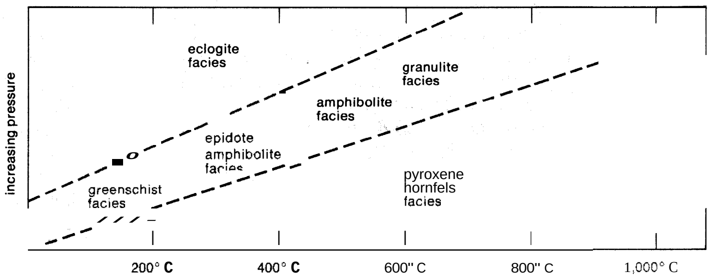


Figure 3: Facies of contact metamorphism.

TEMPERATURE AND PRESSURE OF METAMORPHISM

Knowledge of the temperature and pressure represented by the various metamorphic facies is limited. It seems certain, however, that granulite facies is the highest of regional metamorphic grades and corresponds to about 600"–700" C. Greenschists must have developed at much lower temperatures, perhaps 200" C.

Temperature. Assuming that chemical equilibrium was attained in metamorphic rocks at the time of recrystallization, their mineral assemblages should reveal the correct pressure-temperature conditions of metamorphism. Laboratory studies have given some information on stability relations of mineral assemblages. Such studies indicate that a temperature of 600"–700" C can be assigned to the granulite facies. The two alkali feldspars, microcline and albite, constitute a complete mixed-crystal series above 600" C. At lower temperatures a gradual unmixing takes place, the two coexisting feldspar phases becoming more and more pure. At room temperature only negligible amounts of soda in microcline or potash in albite can exist stably.

Laboratory studies of pressure-temperature conditions of metamorphism

In granulites, albite and microcline are often found as one single phase of solid solution whereas in lower-grade rocks albite with varying amounts of dissolved potash and microcline with varying amounts of dissolved soda are found as two independent minerals (see also FELDSPARS).

A similar situation exists in the calcite-magnesite (calcium carbonate and magnesium carbonate) system. Calcite takes varying amounts of magnesium carbonate in solid solution, depending upon prevailing temperature during recrystallization. It has been determined that magnesian marbles metamorphosed in granulite facies often carry calcite with concentrations of magnesium corresponding to temperatures around 600" C.

Muscovite is a hydrous silicate occurring in many metamorphic rocks. At somewhat elevated temperature the mineral reacts with quartz and loses water, whereby sillimanite and potash feldspar are produced:

$$\underset{\text{muscovite}}{KAl_3Si_3O_{10}(OH)_2} + \underset{\text{silica}}{SiO_2} \rightleftharpoons \underset{\text{feldspar}}{KAlSi_3O_8} + \underset{\text{sillimanite}}{AlSiO_5} + \underset{\text{water}}{H_2O}.$$

Muscovite is never found in true granulites but sillimanite and potash felspar occur frequently together. In lower-grade rocks, muscovite and quartz are commonly found together. Thus the pressure-temperature conditions at which the above reaction takes place should give the lowest boundary of granulite facies—*i.e.*, the boundary line between granulite facies and amphibolite facies. At low pressures this reaction is sensitive to pressure variations because water vapour is involved. At higher pressures, however, laboratory investigations have shown that the reaction temperature at equilibrium is close to 600" C over a wide pressure range. This is then the third mineral system that indicates that the highest temperature in regional metamorphism is not much higher than 600° to 700" C (see MICAS). At higher temperatures silicate rocks begin to melt and magmas are born, a process called anatexis or palingenesis.

A mineral system often applied as a geologic thermometer in metamorphism is calcite + quartz and wollastonite + carbon dioxide.

Pressure. Pressure also varies from one facies to another. In a general way it is known, for example, that pyroxene hornfels facies represent much lower pressures than granulite facies, the latter not only corresponding to the highest temperature in regional metamorphism but also the highest pressure. Reactions among minerals with unlike molal (gram-molecular) volumes are useful geologic barometers. The mineral jadeite, $NaAlSi_2O_6$, which mostly is found in rocks altered or formed at high pressures, is an example. Jadeite has much higher density than albite although the composition of the two minerals is nearly the same, albite having the formula $NaAlSi_3O_8$. Jadeite can form from albite as follows:

$$\underset{\text{albite}}{NAlSi_3O_8} \rightleftharpoons \underset{\text{jadeite}}{NaAlSi_2O_6} + \underset{\text{quartz}}{SiO_2}.$$

The jadeite–quartz combination has considerably less volume than albite so that high pressure favours the jadeite–quartz assemblage. Within the temperature range in which this reaction occurs stably, albite is stable at low pressures and jadeite plus quartz at high pressures. Thus eclogites, which are considered high-pressure rocks by most petrologists, have a high percentage of jadeite molecules (see PYROXENES).

TEXTURE AND STRUCTURE OF METAMORPHIC ROCKS

The mineral assemblages of metamorphic rocks have developed by recrystallization and chemical reactions in the solid state, often under conditions of tectonic stress. Such rocks have therefore texture and structure different from those of igneous and sedimentary rocks.

Mineral grains in metamorphic rocks are mostly without crystal faces and border against their neighbours along irregularly curved surfaces not related to their internal crystalline structure. Such texture is called crystalloblastic. An idioblast is a mineral grain with well-developed crystal faces. During metamorphism certain minerals show greater tendency to form idioblasts than others. Garnet often forms perfect idioblaats in mica schists. whereas quartz never does and feldspar only rarely.

The idioblastic series

Minerals may be placed in the so-called idioblastic series according to their tendency to develop crystal faces during metamorphism. The major members of this series are as follows: sphene, rutile, garnet, kyanite, epidote, pyroxene, hornblende, albite, chlorite, quartz, microcline. Sphene almost invariably forms idioblasts, quartz and microcline very seldom.

Riecke's principle and force of crystallization. A mineral growing in a metamorphic rock has to produce space for itself in a solid environment. Surrounding minerals must then either be pushed aside bodily or disintegrate chemically (for example, dissolve in a pore moisture). The so-called force of crystallization or pressure of growth is of importance in this connection. Any crystal is able to grow against excess pressure if the system is supersaturated with references to that crystal; the more supersaturated the system, the higher the pressure of growth. On the other hand, if a crystal is exposed to excess pressure in a saturated system, it becomes unstable and tends to disintegrate chemically. (Excess pressure here means pressure on the crystal higher than in the rest of the system.) In solid rocks such disintegration does not necessarily mean dissolution in a fluid, because liquid solutions are not always present. It generally means disintegration into individual ions, atoms, or molecules, which are carried away by diffusion along mineral interfaces or even through their crystalline body. This phenomenon of chemical disintegration under excess pressure is referred to as Riecke's principle. It not only explains plastic flow of solid rocks under stress but is also important in explaining recrystallization and neomineralization. A newly stabilized mineral grows in a metamorphic rock because the rock has become somewhat supersaturated with reference to the new mineral owing, for example, to changing pressure-temperature conditions. The growing mineral surrounds itself by a field of pressure in excess of that prevailing elsewhere in the rock. Minerals present in this field of excess pressure disintegrate according to Riecke's principle and thus provide space for the newly stabilized growing mineral. Thermodynamic analysis shows that under otherwise like conditions (same pressure-temperature values, same degree of supersaturation) dense minerals develop higher pressure of growth than less dense minerals. In this connection it may be noted that the above-mentioned idioblastic series of minerals as determined empirically from field studies also is a density sequence, sphene and garnet being dense structures, quartz and feldspar loose, open structures.

Analysis of .structure. The structure of rocks is called isotropic or massive if it is identical in all directions in space. Some contact metamorphic rocks may show such structure, but practically all regional metamorphic rocks have anisotropic structure—*i.e.*, structures with planar and linear features. Planar structure shows up well in most mica schists in which the majority of mica flakes may be parallel to a given plane—the plane of foliation or schistosity. In ideal planar structure all directions in the schistosity plane are identical. Needle-shaped minerals such as siilimanite and hornblende show statistic parallelism to the schistosity plane, but they are oriented at random within that plane. In linear structure needle-shaped minerals show preferred parallelism to a line. This line is called lineation. In most natural rocks linear and planar structures occur together.

The origin of anisotropic structure is twofold:

1. The structure may be relic (a remnant of the original structure) after premetamorphic anisotropic structure of various kinds. Sediments with bedding structure and lavas with flow structure may preserve their anisotropic structure as relic during metamorphic recrystallization even if not exposed to anisotropic stress. Structures formed in this manner are termed mimeoblastic. In polymetamorphic rocks anisotropic structures formed in earlier periods of metamorphism may remain as relic mimeoblastic structure through later recrystallization. In contact metamorphic rocks mimeoblastic structure is common because anisotropic stress is not likely to occur in contact zones around magmas.

2. Anisotropic structure develops in rocks recrystallized under anisotropic stress. The stress causes strain of the rock bodies (cataclastic, with fractures, and plastic, without fractures) and consequent rearrangement of mineral grains into various types of anisotropic structures. The structure of most regional metamorphic rocks is of this kind.

Stress mechanisms in rocks

The mechanism causing such structures to form is little known and probably very complex. It is assumed that the following processes occur in a rock under stress: (1) rotation and relative movement of mineral grains by slip along interfaces; (2) slip along certain crystallographic planes and lines inside mineral grains (so-called glide planes and lines); (3) dissolution (or other kind of chemical disintegration) at points of high compressive stress and redeposition at points of low stress in accordance with Riecke's principle; and (4) fracturing and crushing of mineral grains—so-called cataclasis.

Crushing. Evidence of any one of these processes may be found in metamorphic rocks. If fracturing is dominating the phenomenon is called mylonitization and the produced rock a mylonite. Such mylonites occur usually only in narrow zones in metamorphic terrains, demonstrating intensive shear movements along these zones with corresponding crustal stresses exceeding the crushing strength of minerals. However, many metamorphic rocks show no sign of mineral crushing.

Rotation. Evidence of rotation of mineral grains during metamorphism is not uncommon, the most cited example being the so-called snowball garnet with inclusions arranged in a spiral as a result of rotation during growth.

Slip. Calcite and other carbonates yield readily by slip along certain lattice planes often resulting in polysynthetic twinning and bending of grains. This is observed in many natural marbles in thin sections under the microscope. Laboratory experiments with marble blocks under stress at elevated temperature also show that gliding along crystallographic glide planes is important in rock deformation. It seems, however, that most silicate minerals deform via chemical processes more or less in harmony with Riecke's principle during regional metamorphism. Tectonic stress is generally too weak to produce crushing and lattice slip. Slow creep due to recrystallization seems more important in rock deformation.

Orientation. The statistic parallelism of mineral grains in rocks with anisotropic structure may be dimensional orientation or lattice orientation. Dimensional orientation refers to the external shape of mineral grains, whereas lattice orientation refers to the internal structure of minerals. Quartz grains in schists and gneisses may, for example, have lenticular shapes unrelated to the internal crystal structure. Such quartz lenses are often dimensionally oriented parallel to the schistosity plane. This kind of anisotropic structure is most likely a result of dissolution at points of high compressive stress (*i.e.*, at mineral interfaces that are perpendicular to the highest compressive stress) and redeposition at points of low compressive stress (*i.e.*, at interfaces or surfaces more or less perpendicular to the direction of the least compressive stress or maximum tensile stress).

Minerals with equidimensional outline such as those belonging to the regular crystallographic system may show preferred lattice orientation that can be detected only by accurate microscopic studies.

METAMORPHIC DIFFERENTIATION

During metamorphism, originally homogeneous rock bodies may become heterogeneous in the sense that schlieren (alternating bands of contrasted composition) and veinlets develop. Such phenomena are called metamorphic differentiation. For reasons not clearly understood minerals disintegrate chemically at certain sites in rocks and redevelop at other sites, the transfer probably being in the form of diffusion. Quartz veinlets and clusters commonly form in crystalline schists, "knots" of epidote in green stone, and segregations of diopside or other calc-silicates in crystalline limestone. Such metamorphic differentiation is most prominent in strongly deformed rocks. It is therefore concluded that stress and strain are instrumental in metamorphic differentiation. Minerals that dissolve more readily than others in zones of intensive stress and shearing are leached out from such localities and may be deposited anew where stress is low, thus causing metamorphic differentiation. Interfacial tension, energy of nucleation, and force of crystallization are properties that seem significant in metamorphic differentiation.

Many features in metamorphic terrains caused by metamorphic differentiation are difficult to distinguish from features resulting from magmatic activity. Pegmatite dikes and mineral veins in crystalline schists are commonly products of metamorphic differentiation although similar rocks may also form from magmas. Some workable ore deposits (*q.v.*) in regional metamorphic terrains are probably also products of some sort of metamorphic differentiation.

BIBLIOGRAPHY. T.F.W. BARTH, C.W. CORRENS, and P. ESKOLA, *Die Entstehung der Gesteine* (1939), a classic work in which Eskola gives a textbook version of his important mineral facies principle; T.F.W. BARTH, *Theoretical Petrology* (1962), a textbook treating relatively advanced materials from an elementary point of view; BRIAN BAYLY, *Introduction to Petrology* (1968), a textbook giving an easy-to-understand introduction to the evolution of rocks, including metamorphic~RAYMOND KERN and ALAIN WEISBROD, *Thermodynamics for Geologists* (1964), an advanced work discussing the applications of thermodynamics to metamorphic processes; HANS RAMBERG, *The Origin of Metamorphic and Metasomatic Rocks* (1962), the earliest comprehensive book in which thermodynamic principles are applied to metamorphic problems; F.J. TURNER and J. VERHOOGEN, *Igneous and Metamorphic Petrology* (1960), a general text covering petrology, including metamorphic rocks.

(Ha.R.)

Rocks, Physical Properties of

The response of rocks to physical forces and conditions is impottant in many practical undertakings as well a5 in scientific study of the Earth's structure and development. Mineral exploration and extraction, groundwater supply, excavation, foundation selection and preparation (for dams, bridges, buildings), earthquake safety, construction of rock fill and concrete, and uses of ornamental and building stone are all dependent on physical properties of rock materials. Scientists deduce the Earth's internal structure and composition from indirect measurements of physical properties of the interior; and the processes of past and present development of the Earth, such as the formation of the continents and cordillera (mountain ranges) and the occurrence of earthquakes and volcanism, depend on physical properties of the rocks of which the Earth is made. This article treats the physical properties important in the above subjects, namely, the volumetric, mechanical, thermal, electrical, magnetic, and optical properties of rocks. The significance of these properties in scientific and engineering fields is explored further in the articles EARTH, STRUCTURE AND COMPOSITION OF; EARTH, MECHANICAL PROPERTIES OF; EARTH, HEAT FLOW IN; ROCK DEFORMATION; EARTHQUAKES; ROCK MAGNETISM; MINERALS; GROUNDWATER; MINING AND QUARRYING; SOIL MECHANICS, APPLICATIONS OF; and ICE SHEETS AND GLACIERS. The mineral and chemical compositions of rocks and the textures and structures that influence their physical properties are dealt with in the articles IGNEOUS ROCKS; METAMORPHIC ROCKS; and SEDIMENTARY ROCKS.

FACTORS THAT AFFECT ROCK PROPERTIES

Rocks are aggregates of mineral grains or crystals, and their physical properties are to a large extent governed by the properties of the individual minerals that compose them. In a rock, the individual mineral properties are averaged in accordance with the relative proportions and orientations of the various mineral crystals present. Because of this averaging, the properties of rocks can usually be considered as isotropic (uniform in all directions), although most minerals individually are in fact anisotropic. Rocks in which the crystals are not oriented at random, such as schist and slate, show a definite anisotropy reflecting that of the individual minerals.

The averaging of mineral properties provides a valid basis for most of the bulk properties of crystalline rocks —that is, rocks composed of a dense, interlocking aggregate of mineral crystals. Most igneous and metamorphic rocks, such as granite, basalt, marble, and gneiss, are of this type. Glassy rocks, such as volcanic glass, are noncrystalline and have properties similar to those of artificial glasses. For clastic rocks, composed of an aggregate of originally loose mineral or rock fragments subsequently cemented together, important additional factors governing the physical properties are the grain packing and contact geometry, the amount and distribution of void space (porosity), and the nature of the cementation.

Clastic rocks

Physical properties of rocks are influenced by pressure and temperature. In addition to effects characteristic of the individual mineral components, there are distinctive phenomena related to the granular and void structure of rocks. In particular, the presence of water or other pore fluids has an important influence on some prop-rties.

Although most of the physical properties described here can be measured with reasonable precision for an individual rock specimen, the measured values vary substantially, often by as much as 50 percent, from specimen to specimen of a given rock type and even from specimen to specimen of rock from a single, apparently homogeneous parent mass or formation. This statistical variability is an inherent attribute of rocks, resulting from their complex constitution and the consequently large number of variable details of chemical composition and granular structure that can influence the physical properties. It contrasts with the better defined properties of individual minerals, although these, too, are subject to some variability resulting from variations in chemical composition and internal structure. Because of variability, precise figures for the physical properties of a particular rock type cannot be stated: instead only representative values that convey the general magnitude to be expected. Whenever precise figures are needed, a thorough statistical study of the rock mass of interest is called for. Some of the aggregate physical properties of large rock masses in situ (in their natural locations) in the Earth are significantly affected by gross inhomogeneities, such as fractures that occur on a scale so coarse that they cannot readily be sampled in specimens appropriate for laboratory tests. Also, the act of sampling itself may disturb the rock so as to alter some of its physical properties, In such cases, the physical properties of interest must be measured in *situ* by field methods.

VOLUMETRIC PROPERTIES

Density. Table 1 gives ranges of observed density for several important rock types. For a pore-free rock, the density p is a volume-weighted average of the densities of the individual mineral components—that is, the sum of the volume fraction of each mineral times its density. Among the plutonic rocks—*i.e.*, crystalline rocks of deep-seated igneous origin—the density increases systematically from granite, composed mostly of the minerals feldspar (density 2.56 grams per cubic centimetre, abbreviated g/cm^3) and quartz (2.65 g/cm^3), to peridotite and eclogite, composed mostly of the minerals pyroxene (3.4 g/cm^3), olivine (3.5 g/cm^3), and garnet (3.8 g/cm^3). This density increase plays an important role in the structure of the Earth: plutonic rocks ranging in composition from granite to diorite form the continental masses and essentially float in a dense substratum of peridotite and eclogite, which lies below depths of 10 to 50 kilometres (six to 30 miles).

The range in densities indicated for the individual crystalline rocks in Table 1 mainly reflects variations in mineral and chemical composition among rocks to which the specific names are applied. The average density of fresh rock samples of the types listed is near the middle of the ranges given. Glasses are generally less dense than crystalline solids of the same composition, as illustrated by obsidian, which is volcanic glass of granitic composition.

Porosity. The bulk densities of clastic rocks and of vesicular lavas (the name given to lavas with frozen-in gas bubbles) are reduced significantly by the void space present. The resulting porosity is important in providing storage space for water or petroleum in the rocks. Porosity η is defined as the ratio of void volume to bulk volume (solid plus voids). Typical values are given in Table 1. If the void-free solid material is of density p,, then the porous solid has bulk density (ρ_b) equal to the void-free density times the quantity one minus porosity—*i.e.*, pa $=$ ρ_s $(1 - \eta)$, or $\rho_b = \rho_s$ $(1 - \eta/100)$, if η is expressed in percent, as in Table 1.

Density range in basalt

The large range of densities for basalt in Table 1 reflects variations in porosity attributable to gas vesicles. The porosity of vesicular lavas increases with the amount of volatile material (such as water) dissolved in the original rock melt. Separation of gas bubbles is inhibited by solidification at depth, under pressure. Thus, Hawaiian basalts solidified on the ocean floor at depths less than 800 metres (2,600 feet) are vesicular and have densities less than 2.8 g/cm^3, whereas those formed deeper than 4,000 metres (13,000 feet) are almost vesicle free and have densities close to 3.0 g/cm^3. The most highly vesicular volcanic rocks, called pumice, have densities so low that they float on water.

Table 1: Densities and Porosities of Important Rock Types

	characteristics	density ρ g/cm^3	porosity η (percentage)
Granite	crystalline, plutonic, sialic	2.5–2.8	0.3–1.5
Diorite	crystalline, plutonic, intermediate	2.7–3.0	0.5 ±
Gabbro	crystalline, plutonic, mafic	2.9–3.1	0.5 ±
Peridotite	crystalline, plutonic, ultramafic	3.1–3.3	0.5 ±
Eclogite	crystalline, plutonic, ultramafic	3.3–3.6	0.5 ±
Gneiss	crystalline, metamorphic	2.6–3.1	0.5 ±
Schist	crystalline, metamorphic	2.7–3.0	0.5 ±
Slate	crystalline, metamorphic	2.7–2.85	4 1
Quartzite	crystalline, metamorphic	2.641	0.5 ±
Marble	crystalline, metamorphic	2.6–2.8	0.4–2
Obsidian	glassy, volcanic	2.3–2.4	low
Basalt glass	glassy, volcanic	2.7–2.85	low
Basalt	crystalline–glassy, volcanic	2.2–3.0	1–30
Scoria	crystalline–glassy, volcanic, vesicular	1.4–2.4	10–50
Pumice	glassy, volcanic, vesicular	0.5–1.1	60–90
Sandstone	clastic, young (Tertiary)	1.9–2.2	10–35
Sandstone	clastic, old (Paleozoic)	2.3–2.5	5–25
Shale	clastic, young, shallow (<0.5 km)	1.8–2.2	15–35
Shale	clastic, young, deep (>2 km)	2.4–2.5	9–11
Shale	clastic, old, deep	2.6–2.7	4–7
Chalk; coquina	clastic, shell aggregates	1.2–2.2	15–55
Limestone	crystalline–clastic, shell fragments	2.5–2.7	0.1–15
Tuff	clastic, volcanic shards	1.4–1.5	35–40

The porosity of clastic rocks is highly variable and depends on the original shape and packing arrangement of the fragments, the amount of subsequent compaction, and the amount of void space filled by cementation. Rounded mineral or rock fragments, such as sand or gravel, pack initially to a porosity of 30–50 percent, which is reduced to 25–35 percent by compaction. Lower porosities are achieved when a wide range of particle sizes is present, so that the finer particles can fill the interstices among the coarser ones. Unconsolidated (*i.e.*, loose) sand and gravel are converted into sandstone and conglomerate by cementation, and the porosity is thereby reduced to 5–30 percent. As the rocks become older, additional cementation accumulates and the porosity decreases further (Table 1).

The packing properties of animal shells and shell fragments give high initial porosities for coquina (limestone composed of coarse shell debris) and chalk (limestone composed of shells of micro-organisms). In ordinary limestones, void spaces among shell fragments tend to be filled with fine calcite or aragonite detritus. Compaction can occur through solution and reprecipitation of these readily soluble minerals, leading in some cases to a rock of very low porosity.

Porous clastic rocks in situ in the Earth are generally saturated with water, the upper limit of saturation being called the water table. If the pores are completely filled with water, the bulk density pa is increased by an amount equal to η, the porosity, over the density of the dry rock, which is the density given in Table 1. The volume fraction of water that is absorbed by an initially dry rock specimen is called the sorption or apparent porosity. In the laboratory, sorption is usually about half the total porosity. Under the pressures at depth in the Earth, however, the water content probably approaches the total porosity.

MECHANICAL PROPERTIES

When an external force acts on a body, such as a rock, and there is a change of volume or shape, called a strain, internal forces will be transmitted throughout its interior. An internal force and its force of reaction acting across an imaginary plane that separates two parts of the body is known as stress, and the magnitude of the stress is measured in force per unit area of the imaginary plane. The unit of stress is the bar, defined as 10^6 dynes per square centimetre; to a good approximation it equals one atmosphere (14.7 pounds force per square inch) or one kilogram weight per square centimetre. Often used is the kilobar (kbar), which is 1,000 bars.

Among types of response to externally applied stress, distinction is made between elastic properties, in which the strain is reversible (recoverable on unloading), and nonelastic properties, in which it is not. In elasticity, strain is usually proportional to stress, and the constant of proportionality is called an elastic constant: specific examples discussed below are the elastic compressibility, bulk modulus, Young's modulus, and shear modulus. The response under the simplest type of stress, hydrostatic pressure, is usually considered separately from the response to directed, nonhydrostatic stress; hydrostatic behaviour is in principle only a special case of nonhydrostatic, but phenomena of different types are important under the two different types of stress.

Elastic properties. Compressibility. The density of rocks under the high pressures at depth in the Earth is increased as a result of compressibility (reduction in volume). Bulk compressibility β is defined in terms of the increase of bulk density ($\Delta\rho_b$) that occurs on raising the pressure by a small amount (ΔP), at constant temperature, and is equal to the fractional bulk density increase

Table 2: Elastic Properties of Rocks at Low and High Confining Pressure*

	compressibility β (10^{-6} bar^{-1})		moduli at low P (10^6 bar)			moduli at high P (10^6 bar)			Poisson's ratio ν	
	low P	high P	K	E	G	K	E	G	low P	high P
Granite	8	2.0	0.1	0.3	0.2	0.5	0.6	0.4	0.05	0.25
Gabbro	4	1.1	0.3	0.9	0.6	0.9	0.8	0.5	0.1	0.2
Dunite	—	0.8	1.1	1.5	0.5	1.2	1.7	0.7	0.3	0.27
Obsidian	2.8	2.8	0.4	0.7	0.3	—	—	—	0.08	—
Basalt	2.2	1.3	0.5	0.8	0.3	0.8	1.2	0.4	0.23	0.25
Gneiss (granitic)	11	1.9	0.1	0.2	0.1	0.5	0.7	0.3	0.05	—
Marble	9	1.3	0.1	0.4	0.2	0.8	0.7	0.3	0.1	0.3
Quartzite	7.6	2.4	—	—	—	0.5	1.0	0.4	—	0.07
Sandstone (2.3 g cm^{-3})	14	—	0.07	0.2	0.08	—	—	—	0.1	—
Shale	–25	–5	0.04	0.1	0.05	—	—	—	0.04	—
Limestone	1.3	1.3	0.8	0.6	0.2	—	—	—	0.30	—

'Low pressure—approximately one atmosphere; high pressure—approximately 3,000 atmospheres.

divided by the pressure increase. The density increase considered in this definition is that which results from elastic compression; any loss of volume that is not recovered on releasing the pressure represents compaction (see below).

Compressibility of jacketed rock

When pressure is applied to a rock specimen by means of a fluid that permeates the pores, there is no compaction, and the compressibility observed is an average of the elastic compressibilities of the individual mineral components. On the other hand, when the rock is kept dry by a fluid-impermeable jacket enclosing the specimen, compaction may occur, and, in addition, the elastic compressibility is usually much greater than an average taken of the mineral components. The same behaviour is observed also for jacketed wet specimens vented so that the pore fluid can escape at low pressure. This arrangement simulates the conditions of compression experienced by rocks at depth in the Earth's crust, because rock permeability (discussed later) allows the pore fluids to escape, and the pore fluid pressure is usually only about a third the total overburden pressure.

The high compressibility of jacketed rock specimens is caused by closure of pore space under pressure. This effect is marked even for crystalline rocks of low porosity. As the pressure on a specimen is raised and the pores close, the compressibility decreases markedly. Low- and high-pressure compressibilities measured for rocks of several types are given in Table 2. Above a pressure of about 2,000 atmospheres, the compressibility becomes nearly constant, because substantially all of the pores that can close are already closed. Pores capable of closing over the pressure range zero to 2,000 atmospheres must be narrow cracks; hence the great increase in compressibility at low pressure is said to be due to crack porosity. The cracks probably lie mainly along the boundaries between mineral grains. Pores of roughly equidimensional shape (such as vesicles and interstitial spaces between clastic grains) do not close up rapidly with pressure and contribute only moderately to the compressibility, but the contribution must continue to much higher pressures than that caused by the crack porosity. The drop in compressibility with pressure is much larger in the case of crystalline rocks that formed at depth than it is in those that formed near the surface, suggesting that the cracks opened as a result of the pressure release that followed formation of the deep-seated rocks,.

Once any initial crack porosity has been removed by application of pressure, the compressibility of a rock is an average of the compressibilities of the mineral components. The averaging process is complex, however, because an aggregate of grains having different and generally anisotropic compressibilities cannot compress uniformly but instead develops a complex pattern of stresses and strains varying from grain to grain. An exact method of calculating the compressibility of a rock from the cornpressibilities of its mineral components is not known, but limits on the compressibility can be set by two types of average: (1) the so-called Reuss average (after the German physicist A. Reuss), which is the volumetrically weighted average of the compressibilities of the mineral components; and (2) the Voigt average (after Woldemar Voigt), in which the reciprocal compressibility (called the bulk modulus) is calculated as the volumetrically weighted average of the bulk moduli of the components. The true compressibility must lie between these two averages. The Voigt average can be applied to estimate the compressibility of a rock with quasi-spherical pores by treating the pores as components of zero bulk modulus.

Nonlinear compressibility. Over a pressure range from about 2,000 to 10,000 atmospheres (or zero to 10,000 atmospheres for a rock with no initial crack porosity), the compressibility can be considered to be linear in the sense that the compression increases in direct proportion to the pressure, and the compressibility coefficient β is practically constant. Over a wider pressure range, however, such as the range zero to 2,000,000 atmospheres that occurs within the Earth, deviations from linearity become significant. Theories of finite elastic strain have been developed to deal with compressibility as a function of pressure over this wide pressure range. It is found empirically that for many rocks the bulk modulus K (reciprocal compressibility $1/\beta$) is to a good approximation a linearly increasing function of pressure: K is equal to the bulk modulus at zero pressure (K_0) plus a constant (K') times the pressure P, or $K = K_0 + K'P$. The corresponding rock density (ρ) as a function of pressure is given by the density (ρ_0) at zero pressure times the quantity 1 plus the ratio of pressure to zero-pressure bulk modulus raised to the power $1/K'$, or

$$\rho = \rho_0(1 + P/K_0)^{1/K'}.$$

The dimensionless quantity K' (an arithmetic value without units) has a value in the range four to six for most rocks. It plays an important role in reasoning about the internal density and composition of the Earth.

Nonhydrostatic stress. When a cylindrical specimen is compressed along its length, while laterally unconfined, the compressive stress σ is related to longitudinal elastic strain e_{long} by Young's modulus (after Thomas Young, an English physicist). Young's modulus (E) is equal to the ratio of stress to strain, or $E = \sigma/e_{long}$. The ratio of lateral expansion e_{lat} to longitudinal compression $-e_{long}$ is Poisson's ratio (after a French mathematician, Siméon-Denis Poisson). Poisson's ratio (ν) is thus $\nu = -e_{lat}/e_{long}$. (The longitudinal or lateral strain e is the ratio of change in specimen length, Δl, or width, Δw, to the corresponding initial dimension l or w: el,,, $= \Delta l/l$, $e_{lat} = \Delta w/w$.) When a specimen is loaded in shear—that is, with stress parallel to a particular plane—the ratio of shear stress (τ) to shear strain (y) is the shear modulus, $G = \tau/\gamma$ (see ELASTICITY). Values of the moduli and Poisson's ratio for several rock types are listed in Table 2. Because of the large typical variation of measured moduli from specimen to specimen, the figures in Table 2 are given only to one decimal and convey only the general magnitudes to be expected for the given rock types. The greatest source of variation is crack porosity. As the stress is increased, some of the cracks become closed; hence, Young's modulus increases with compressive stress σ. The cracks that tend to close are aligned transverse to the direction of compression; hence, Poisson's ratio is low (about 0.1) initially, at low σ, and increases with σ, approaching generally a value near 0.25 when the cracks become closed. Similar effects are produced by hydrostatic confining pressure. The high pressure values given in Table 2 represent rocks from which crack porosity has been eliminated.

For isotropic materials, Young's modulus, the shear modulus, and Poisson's ratio are, in principle, related by the condition that Poisson's ratio (ν) is equal to the ratio of Young's modulus (E) to twice the shear modulus (G), less 1, or v $= E/2G - 1$. Similarly, the moduli E and G are related to the bulk modulus K and Poisson's ratio ν by: $E = 3K(1 - 2\nu)$ and $G = 1.5\ K(1 - 2\nu)/(1 + \nu)$. Tabulated values do not always satisfy these relations because the values may have been measured on different specimens, under different conditions, or by different techniques or because the specimens may not have been fully isotropic. For $\nu = 0.25$, it is expected that $E = 1.5$ K and $G = 0.6$ K; these expectations are realized to a rough approximation by the modulus values in Table 2. Limits on the moduli E and G can be estimated from the elastic properties of the mineral components by methods similar to those already discussed for the compressibility.

Cause of elastic anisotropy

Elastic anisotropy occurs when the mineral components have a preferential alignment, as in gneiss (a type of metamorphic or altered rock), schist, and slate. In a schist with highly aligned flakes of mica, Young's modulus measured parallel to the plane of alignment can be as much as three times as large as when it is measured perpendicular to this plane, and an even larger anisotropy in the shear modulus can occur. Actually observed elastic anisotropies of rocks are, however, generally much smaller. Preferential alignment of unclosed microscopic cracks can also cause elastic anisotropy.

Elastic **waves.** When a rock is subjected to oscillatory loading, waves of oscillating displacement and stress, much like sound waves, are propagated through it. These waves are of two types: longitudinal or compressional waves (P waves), in which the displacements occur along a direction parallel to the direction of propagation; and the transverse or shear waves (S waves), in which the displacements occur at right angles to the direction of propagation. P waves propagate with a velocity (v_P) that is equal to the square root of the density (ρ) divided into the sum of the bulk modulus and 4/3 times the shear modulus $v_P = \sqrt{(K + \frac{4}{3}G)/\rho}$ whereas S waves propagate with a velocity (v_S) that is equal to simply the square root of the ratio of the modulus G to the density: us $= \sqrt{G/\rho}$. The velocities can be calculated from static test data such as those in Table 2, but they are often measured directly for comparison with velocities at depth in the Earth as inferred by seismology. At low pressures, crack porosity causes the moduli, and hence the velocities, to be low and erratic, but above a confining pressure of about 2,000 atmospheres the intrinsic velocity, representing an average over the mineral components, is realized. Values for several rock types are listed in Table 3. Higher values are shown by the denser rock types, reflecting the fact that denser minerals generally have elastic moduli that are increased much more than in proportion to their increased density. When, however, the density increase is due primarily to a substitution of iron for magnesium or aluminum in the mineral composition, relatively little increase in elastic moduli occurs, and the elastic wave velocities drop. Empirical observations are summarized by the following approximate formula relating P-wave velocity (in kilometres per second) to rock density p (grams per cubic centimetre) and the average value of the atomic weights of the constituent atoms (symbolized A): or, $v_P \approx 7.3 + 3.1\ (\rho - 3.0) - 0.5(\bar{A} - 21)$. For most igneous rocks A is near 21, except for rocks notably enriched in iron, for which it can approach 30. For mica-rich rocks in which the mica flakes are strongly aligned, such as mica schist, an appreciable velocity anisotropy is expected; for example, the P-wave velocity parallel to the foliation (plane of mica alignment) can be as much as 1.8 times higher than that perpendicular to the foliation of the schist.

Elastic waves are subject to attenuation, a decrease in wave amplitude, which results from imperfections in elasticity, called internal friction, delayed elasticity, or viscoelasticity. The attenuation is measured in terms of the specific damping capacity b, which is the fraction of the elastic energy lost from the wave per cycle of oscillation. (It is often reported in terms of the quality factor [Q], defined as equal to the numerical factor 2π divided by the specific damping capacity, *i.e.*, $Q = 2\pi/b$.) For crystalline rocks, specific damping capacities are typically in the range from 0.5 to 10 percent per cycle under ordinary conditions. Porous clastic rocks such as sandstone and shale often show high attenuation (with a specific damping capacity *[h]*of about 10 to 50 percent per cycle). The attenuation under ordinary conditions is dominated by frictional losses that arise from sliding across microscopic cracks; variability in characteristics of crack porosity is responsible for the wide scatter in observed attenuations. About 5,000 atmospheres of confining pressure inhibits attenuation due to crack sliding and reduces b to less than 2 percent per cycle. The remaining attenuation is due to a variety of mechanisms, which have not yet been thoroughly elucidated, because of the complex composition and structure of rocks. The known individual loss mechanisms have a marked dependence on frequency of the elastic wave, but the aggregate effect caused by superposition of many mechanisms is an observed specific damping capacity that shows no definite dependence on frequency over a wide range of frequencies, from those used in typical laboratory experiments, 10^2 to 10^6 hertz (cycles per second), to those encountered in observed seismic waves, 10^{-1} to 10^{-3} hertz.

Table 3: Elastic Wave Velocities for Rocks at Low and High Confining Pressure (km/sec)

	low P*		high *P*†	
	V_P	V_S	V_P	V_S
Granite	5.3	2.3	6.3	3.6
Diorite	5.4	3.1	6.6	3.7
Gabbro	6.5	3.5	7.1	3.8
Dunite	7.4	4.2	8.0	4.5
Eclogite	7.0	4.0	7.8	4.5
Diabase	6.3	3.2	6.8	3.8
Basalt	5.6	3.0	—	—
Marble	5.8	3.2	6.7	3.5
Quartzite	5.6	—	6.2	4.0
Slate	4.3	2.9	—	—
Sandstone:	2.2	—	4.5	—
Shale‡	2.1	—	4.4	—
Limestone:	2.8	1.1	6.4	3.0
Clay	1.6	0.7	—	—

*Approximately one atmosphere. †Approximately 4,000 atmospheres. ‡Velocity for young, porous sediments at shallow depth is given in low-P column; those for old, dense sediments at substantial depth (—4 km) are given in the high-P column.

Temperature and pressure effects

Increase of temperature generally causes an increase in attenuation, but at depth in the earth's mantle the inhibiting effect of pressure dominates, and b is low (approximately 0.5 percent per cycle). A conspicuous increase in attenuation, by as much as a factor of 2, occurs when a rock begins to melt. Partial melting is probably responsible for the relatively high attenuation (about 5 percent per cycle) for seismic P waves traversing the low-velocity zone in the Earth at a depth of about 200 kilometres (100 miles). In general, attenuation for S waves appears to be larger than for P waves by about 50 percent.

Nonelastic properties. Permanent deformations, nonrecoverable on unloading, become detectable normally at stresses of 10 to 100 bars in mechanical tests on rocks. At the much lower stresses in typical elastic waves, there is some deviation from reversible elastic behaviour, as indicated by attenuation, but permanent deformations are small and generally remain undetected. Except for compaction, the nonelastic properties of rocks appear only under directed stress. **A** representative curve of stress σ

as a function of strain *e* in a mechanical test is shown in Figure 1. In the elastic range of the curve, from *O* to Y, some deviation of the curve from an ideal straight line usually occurs, as shown, because of crack closure. The knee in the curve at Y represents the onset of permanent, nonrecoverable strain and is called the yield point. The permanent strain is evident on unloading, which follows the path from W to *S*. At point R, rupture occurs, and the load that the rock can support (stress) drops abruptly to zero. The amount of permanent strain accumulated between Y and R is a measure of the ductility of the specimen. Under ordinary conditions, most rocks are brittle—*i.e.*, rupture with little or no ductile deformation—and the rupture point R essentially coincides with the yield point Y.

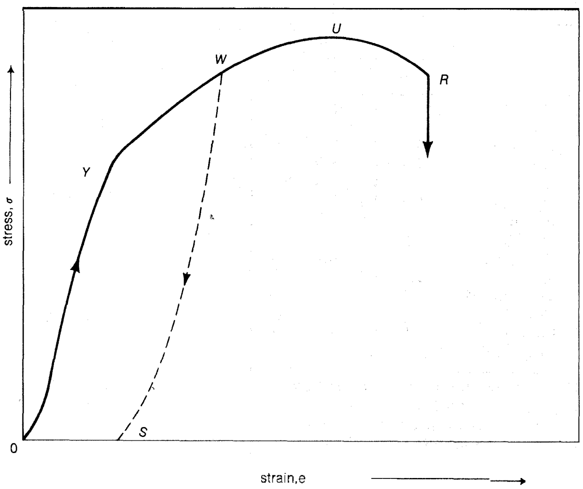


Figure 1: A typical stress-strain curve for rock. *Y* is the yield point (beginning of plastic failure), *U* the point of ultimate strength, and *R* the point of rupture. WS represents unloading before rupture (see text).

Compaction. Porous rocks composed of or cemented by relatively weak minerals, such as clay, undergo significant compaction under pressure, the weak grains deforming inelastically into the pore spaces. Thus the porosity of shale decreases substantially with depth of burial (Table 1). For porous sandstones, compaction is normally minor because the tight packing of strong grains (quartz or feldspar) provides mechanical support for the cementing material, which is itself often a strong mineral (calcite or quartz). Under high pressures (approximately 10,000 atmospheres), however, the strong grains themselves fail and porosity is eliminated. No standard measure of compaction as a physical property is available.

Brittle rupture. In most practical situations in which rocks are used as building or foundation materials, their strength is limited by brittle failure. The sudden release of stored elastic energy that occurs in brittle failure of rocks below the surface of the Earth is the source of the seismic disturbances in earthquakes, and in deep mines it takes place in "rock bursts," causing violent damage. In brittle failure, the rock loses cohesion and splits into two or more pieces along fracture surfaces. Two types of fracture are distinguished (see Figure 2): (1) extension fracture, in which perpendicular separation occurs across a surface oriented at right angles to the direction of greatest tensional (pulling apart) stress; and (2) shear fracture, or faulting, in which there is lateral (shear) displacement across a fracture surface inclined at about 25° to the direction of greatest compressive stress. In extension fracture, the stress supported by the rock always drops abruptly to zero. In shear fracture, an abrupt stress drop usually occurs, but the stress does not drop entirely to zero because of sliding friction across the fracture surface; under some conditions the friction is large enough to prevent any stress drop when the fracture or fault develops. Cylindrical rock specimens fail by extension fracture when loaded in unconfined tension and by shear fracture in unconfined compression, or in compression with superimposed confining pressure. Under intermediate states of stress there is a transition from shear to extension fracture, in which the orientation and nature of the fracture surfaces change progressively.

Crushing strength

The stress required to cause brittle failure in an unconfined compression test is called the compressive strength or crushing strength and, in a tensile test, the tensile strength. Representative values are given in Table 4. The tensile strength of a rock is typically only one-tenth to one-twentieth the compressive strength. Because of effects of internal friction, the strength increases substantially with pressure. In compression tests with superimposed pressure, the strength, measured as the difference between axial compressive stress σ and lateral confining pressure P at failure, increases linearly with the pressure: the strength $(\sigma - P)$ is equal to the unconfined crushing strength (C_0), plus the product of a coefficient (s) times the bulk confining pressure (P); *i.e.*, $(\sigma - P) = C_0 + sP$. The coefficient (s) determines the frictional contribution to the strength and commonly has a value in the range three to 10 (Table 4). At a depth of five kilometres (three miles) in the Earth, where pressure is of the order of 1,500 atmospheres, the frictional contribution (*sP* approximately 9,000 bars) greatly exceeds the contribution from internal cohesion (c, is approximately 2,000 bars).

If a rock contains pore fluid under a pressure *p*, the strength is reduced accordingly: the pressure quantity that is effective in governing the strength according to the formula above is the confining pressure (P) less the pore pressure (*p*), thus $(\sigma - P) =$ C, $+ s(P - p)$. The quantity $(P - p)$ is called the effective confining pressure and plays an important role in the mechanical properties of rocks. It is thought that the pumping of water into the ground under high pressure has, by decreasing the effective confining pressure, been responsible in some instances for reducing the strength of rocks to the point where faulting and earthquakes have occurred. There is evidence that, even in the absence of significant pore pressure, the presence of pore water reduces the strength of sandstones about 50 percent below that of the dry rocks.

Brittle failure of solid rock is controlled by the presence of microscopic flaws that reduce the strength greatly from what is theoretically possible for perfect mineral crystals. These flaws are doubtless closely related to the cracks responsible for anomalies in compressibility and elastic moduli, discussed above. The Griffith theory of failure proposed by the English engineer A.A. Griffith supposes that at the microscopic level the cracks grow by tensile fracture. The theory is applicable to rocks if the additional assumption is made that slippage across the

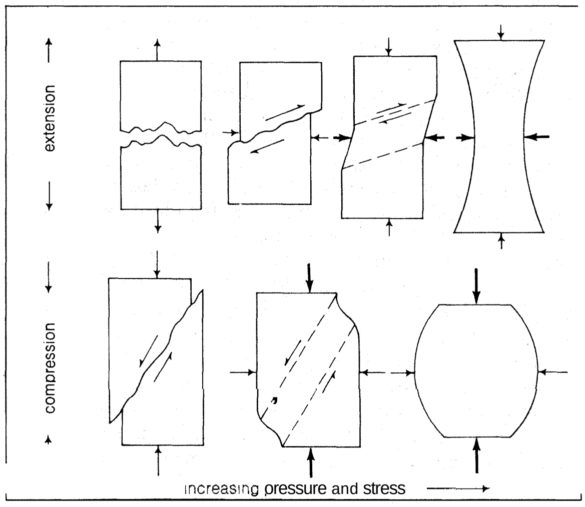


Figure 2: Stages of transition from brittle to ductile failure, as seen in laboratory tests on cylindrical specimens. (Top) The vertical axis is of relative extension, (bottom) of compression. (Left) Extension and shear fracture are illustrated, and (left to right) a progressive increase in confining pressure is shown. Bold arrows indicate directions of greater relative stress.

Table 4: Brittle Strength Parameters for Rocks

	tensile strength T_0 (kbar)	crushing strength C_0 (kbar)	pressure dependence s	cohesion τ_0 (kbar)	internal friction μ	fracture angle θ (in degrees)	
						predicted	observed
Granite	0.2	2.3	7.3	0.39	1.3	19	26
Granodiorite	—	1.1	9.7	0.17	1.5	17	—
Didhase	0.4	4.9	3.8	0.87	0.9	24	26
Basalt	—	2.2	5.6	0.44	1.1	21	—
Gneiss (dioritic)	–	1.1	9.0	0.18	1.4	18	—
Dolomite marble	0.1	1.5	2.1	0.38	0.8	26	17
Quartzite	0.3	4.6	8.4	0.76	1.4	18	25
Limestone	—	1.1	11.6	0.15	1.6	16	—
Sandstone	—	0.5	4.9	0.11	1.0	23	—
Siltstone	—	0.03	8.3	0.005	1.4	18	—

cracks is restrained by friction. If the coefficient of friction is p, the theory predicts a linear dependence of strength on effective confining pressure, with coefficient $s = 2\mu/(\sqrt{1+\mu^2} - \mu)$. In the Coulomb-Navier-Mohr theory of failure resulting from the work of a French physicist, Charles-Augustin Coulomb, a French engineer, Claude-Louis-Marie Navier, and a German engineer, Otto Mohr, which predicts the same relationship but is conceptually less satisfactory, the quantity μ is called the coefficient of internal friction. (It should not be simply equated with the internal friction that causes elastic wave attenuation.) The theory predicts that the fracture surface should be oriented relative to the compression axis at an angle (θ) that is 45° minus half the angle whose tangent is μ, thus $\theta = 45" - \frac{1}{2} \tan^{-1}\mu$. It also assigns to each rock a quantity called the cohesive strength, which is the shear stress necessary to cause failure across a potential fracture surface in the absence of any frictional constraint; the cohesive strength (τ_0) is related to the crushing strength (C,) as follows: $\tau_0 = \frac{1}{2}C_0(\sqrt{1+\mu^2} - \mu)$. The cohesion τ_0 is of the order 100 bars for clastic rocks and 500 bars for crystalline rocks. The values of μ needed to account for s in the observed pressure dependence of rock strength are reasonable as coefficients of friction, being near unity (Table 4). The predicted fracture orientation angles (Table 4) are in rough agreement with observation, except in the transition from shear to extension fracture. The strength quantities C_0 and τ_0 vary approximately as the inverse square root of the size of mineral grains (the individual crystals or particles of which the rock is made) as is predicted by the Griffith theory: in finer grained rocks microscopic cracks tend to be shorter than in coarser grained rocks; hence the finer grained rocks are stronger.

In rocks with planar anisotropy, such as slate, fractures tend to develop preferentially parallel to the plane of weakness, even when it is misoriented by as much as 30" from the fracture orientation predicted from the theory for isotropic materials. Such fracture, along planes of a definite orientation, is called cleavage. Even apparently isotropic rocks such as granite show a significant tendency in quarrying to split preferentially along certain planes, called the "rift" and "grain" of the rock.

Individual measurements of the crushing and tensile strengths of rocks often show wide fluctuations, typically by as much as 50 percent from the average values for rock from a given source. This wide scatter is probably a result of uncontrollable variations in characteristics of crack porosity from sample to sample. Variations in average values among rocks of the same type from different sources are also large; hence the figures given in Table 4 should be viewed as representative examples only.

Repeated brittle failure. In the Earth, brittle failure commonly occurs in fault zones, in which the rocks have a previous history of failure and are pervaded by pre-existing fractures; this is a situation distinctly different from the failure of virgin rock, described above. If the fractures are not healed by cementation, the rock in bulk has already lost cohesion and has no tensile strength. In compression, in which the rock fails largely by slipping on the pre-existing fractures, however, the friction across these surfaces determines the strength; at depths of several kilometres it is substantial (thousands of bars), unless the effective pressure is reduced by an abnormally high pore pressure. The pre-existing fractures contain comminuted (pulverized) rock debris produced by previous slippage; hence the failure properties are related to those of incoherent granular materials such as sand or soil (see SOIL MECHANICS, APPLICATIONS OF). Of great importance is the distinction between two contrasting types of frictional behaviour: (1) stable sliding, in which the coefficient of friction (a measure of the resistance to sliding) remains essentially constant as slip progresses across the fracture surface; and *(2)* stick-slip friction, in which the initiation of sliding is accompanied by a large and abrupt drop in the coefficient of friction. Stick-slip friction results in sudden stress release, the cause of earthquakes, whereas stable sliding allows a slow creeping motion across the faults, without earthquakes. Increased temperature and decreased pressure promote stable sliding instead of stick-slip friction. The physical conditions controlling the two types of frictional behaviour have important influence on the causes, locations, and hazards caused by earthquakes.

Types of friction

Brittle-duetiletransition. By raising the fracture strength, confining pressure inhibits brittle failure and allows other mechanisms of yielding to occur, which produce a more or less homogeneous permanent deformation and constitute ductile failure (Figure 2). The pressure required to effect the transition from brittle to ductile failure is relatively small (1,000 atmospheres or less) for rock materials of low plastic yield strength, such as limestone, shale, and rock salt. At ordinary temperatures, most igneous and metamorphic rocks and also clastic rocks well cemented by silica begin to show ductile behaviour only at pressures of 10,000 to 30,000 atmospheres, because the directed stress needed to cause plastic deformation in the silicate minerals of these rocks is high. The pressure variable pertinent to the brittle-ductile transition is the effective pressure, as defined above, rather than the total pressure. Intermediate between brittle failure at low pressure and true plasticity at high pressure is a type of deformation behaviour called cataclastic, in which the rock fails in a ductile manner as viewed macroscopically but in which on the microscopic scale there is progressive comminution of the mineral grains by local fracturing distributed microscopically throughout the specimen. Cataclastic deformation occurs under conditions where the internal friction of virgin rock is enough lower than the sliding friction on already formed fracture surfaces that slippage is inhibited on fractures once formed, and brittle failure therefore spreads pervasively through the rock. Increase of temperature promotes true rock plasticity at the expense of cataclastic flow and lowers the brittle-ductile transition pressure. The explanation for this is that temperature has a much greater effect in lowering the plastic yield stresses of minerals than in lowering the cohesions or internal frictions.

Plasticity. In true plastic yielding of rock, intra-crystalline plastic deformation of the individual mineral grains takes place. Crystals, which are composed of uniform layers, or planes, of atoms, deform plastically by two distinct mechanisms: (1) translation gliding, in which individual atomic planes appear to slip past one another, leaving the basic crystal structure unchanged; and (2) mechanical twinning, in which the atomic planes slip through a definite fraction of the crystallographic

Types of crystal deformation

repeat distance, in such a way as to produce a new crystal in twinned orientation relative to the original. The displacements of the atomic planes are actually made possible by the motion of crystal dislocations and are crystallographically controlled: each mineral has certain definite crystallographic planes (called slip planes) across which translation gliding or mechanical twinning can take place by slip in a definite crystallographic direction. Mechanical twinning produces visible twin lamellae (layers) in the individual crystals, which are a common feature of deformed marbles. Translation gliding leaves less obvious traces but can result in a bending of the individual crystals, which is often found in the quartz grains of deformed rocks.

Critical resolved shear stress

For each set of potential slip planes in a given mineral under given conditions, a certain definite shear stress acting across the planes in the direction of slip is necessary to initiate slip; it is called the critical resolved shear stress. For the appropriate planes in calcite this stress is about 80 bars for mechanical twinning and about 1,200 bars for translation gliding at 24° C (75" F) and 3,000 atmospheres confining pressure. With the exception of sheet silicates (clays and micas), silicate minerals generally have high critical shear stresses, in the range 2,000 to 30,000 bars.

Cause of hardening

Once the critical shear stress for a set of slip planes is reached, substantial plastic shear strain can occur by slip across these planes; ideally there is only a modest further increase in stress (called work hardening), so that the stress-strain curve has a sharp knee at the yield point Y (Figure 1), with a low slope above that point. In a rock, the individual crystals cannot yield simultaneously because, with their various orientations, the potential slip planes in different grains have widely different resolved shear stresses for a given stress applied to the rock as a whole. Moreover, the slip systems of adjacent grains in a rock are not generally compatible, and the crystals therefore tend to interfere with one another during plastic flow, disrupting one another and causing increased work hardening. The result of these effects is that the plastic yielding of a rock occurs less sharply and at substantially higher stresses than that of individual crystals of its component minerals. The stress-strain curve has a broad, rounded knee at Y (Figure 1) and a substantial slope above Y. Because of this, a meaningful statement of the plastic yield strength of a rock requires specification of the strain to which the quoted strength refers. Strengths observed for several rock types at 2 percent strain are listed in Table 5. Because of substantial variation in results from one experiment to another, only rough, generalized figures are given. The ultimate strength, which is the maximum stress reached in a loading test (*U* in Figure 1), is often substantially higher than the yield strength because of work hardening. The strain range of ductile behaviour is usually terminated by brittle failure (*R* in Figure 1), which occurs when the deforming stress has risen to the point at which the confining pressure is no longer able to inhibit fracture. Contributing to it is a progressive deterioration of the rock texture caused by structural flaws generated in plastic flow; these flaws reduce the cohesion.

Table 5: Plastic Yield Strength of Rocks

	temperature (°C)	pressure (kbar)	plastic yield strength at 2% strain (kbar)	ultimate strength (kbar)
Granite	500	5	10	11.5
	800	5	5	6
Gabbro	500	5	4	8
Peridotite	500	5	8	9
	800	5	5.5	8
Basalt	500	5	8	10
	800	5	2	2.5
Marble	24	2	2.5	5.5
	500	3	1	2
Quartzite	500	8	21	22
	1,000	8	7*	10
	1,000	8	4†	5
Limestone	24	2	4.5	5.5
	500	3	2.5	3
Dolomite	24	2	6	7
	500	5	4	6.5
Shale	24	2	1.5	2.5
Rock salt	24	1	0.5	1

*At strain rate of 2.4 percent per minute. †At strain rate of 0.2 percent per minute.

In conformity with the fact sliding friction is not involved in the plastic yield process, pressure has only a slight effect on the plastic yield strengths of minerals and rocks. The effect is a modest increase in strength, comparable to the increase in elastic bulk modulus with pressure, discussed earlier. Increase of temperature, on the other hand, lowers the yield strength substantially (Table 5). Truly plastic behaviour of igneous and metamorphic rocks containing strong silicate minerals is generally observed only above temperatures of about 600° C (1,100" F). The effect of temperature is even more dramatic when water is present: the yield strength of quartz at 600" C drops from 20,000 to 1,500 bars when water is made available. It is thought that water causes this effect by breaking some of the strong silicon–oxygen bonds in the silicate mineials. The phenomenon of water weakening explains the seeming contradiction between the great rigidity of quartz as observed in most laboratory tests and the abundant evidence of plasticity in the quartz of natural rocks, even in rocks that must have been deformed at only relatively low temperatures.

Time-dependent deformation: creep. When a rock placed under load experiences a strain that does not appear immediately but instead appears gradually as time progresses, the phenomenon is called creep. The slow, drifting motion of the earth's continents and oceanic plates (rigid, slablike regions of the ocean floor) is a surface indication of creep taking place in the rocks at depth. At ordinary temperatures, the creep rate immediately after the initial loading of a specimen may be relatively large, but it decreases with time. Transient creep, as this is called, yields only a limited total strain; it may, however, affect the use of rock as a building material under high loads. Transient creep of this kind is due primarily to a gradual loosening of the granular texture of the rock. At high temperatures (within a few hundred degrees Celsius of the onset of melting) steady-state creep makes its appearance; this is a type of creep in which, after possible initial transients, the creep settles down to a steady rate under fixed stresses. Although the rate may be low, it can lead cumulatively to large deformations when continued over the millions of years of geologic time. The required stress need not exceed the plastic yield stress, and the strain rate may be so low as to be undetectable in laboratory tests on rocks.

Effective viscosity

For creep, the physically significant relationship is not between stress and strain but between stress and strain rate. If the relationship is linear, so that strain rate is directly proportional to stress, the creep behaviour is similar to the flow of a viscous fluid except that the deforming material is a crystalline solid rather than an amorphous liquid. The ratio of shear stress to shear strain rate is called the effective viscosity. Certain creep phenomena in the Earth, such as the slow rise of Scandinavia due to removal of the continental ice-sheet load at the end of the Ice Age (called postglacial rebound), have been interpreted in terms of linear creep and imply an effective viscosity of about 10^{22} poise (dyne sec cm^{-2} unit) for rocks in the outer part of the Earth's mantle. The flow of glacier ice corresponds to an effective viscosity of roughly 10^{14} poise. For comparison, the viscosity of water is 0.01 poise, and the viscosity of molten lava is in the range 104 to 10^5 poise at temperatures of 1,000"–1,100" C (1,800°–2,000°F). With an effective creep viscosity of 10^{22} poise, a rock specimen loaded in compression at 1,000 bar would undergo a total strain of 10^{-6} (one millionth) in a year's time.

Laboratory experiments on limestone, marble, dunite (a kind of igneous rock), rock salt, and ice have shown that over the range of measurable strain rates, the creep behaviour is markedly nonlinear. The same is observed in the high-temperature creep of metals. The type of nonlinearity commonly found in the relation between stress σ and strain rate (*i.e.*, change of strain with time, symbolized $\dot{e}$) has the form, strain rate equals a constant (k) times stress raised to a power n, or $\dot{e} = k\sigma^n$. Values commonly found for the exponent n range from 3 to 7, rather different from a value of 1, which represents linear creep. Linear and nonlinear creep behaviour are compared in Figure 3. Because of the observed creep nonlinearity, strain rates increase much more rapidly than in simple proportionality to stress. The effective creep viscosity decreases as the stress increases. At low stresses, there is hardly any creep, and as the stress rises above a level of order σ_0 in Figure 3, the strain rate increases rapidly, so that the behaviour is somewhat like that of a plastic material with yield stress σ_0. Creep rates in the Earth are probably less, in general, than the experimentally studied range, and it is possible that at the lower stresses the creep behaviour tends toward linearity.

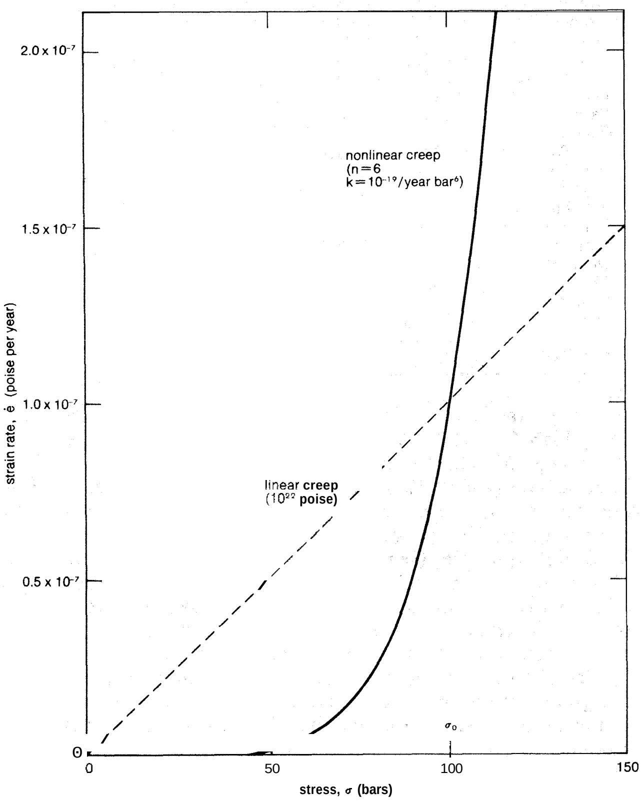


Figure 3: Comparison of stress versus strain-rate curves for linear and nonlinear creep. Linear creep is shown for a rock that deforms to a substantial strain ($e \sim 1$)under a stress of 100 bars applied for 10,000,000 years, a geologically typical period of rock deformation. Nonlinear creep is plotted for a material that behaves approximately as **a** plastic material with yield stress $\sigma_0 \sim 100$ bars.

Rock creep can take place by intra-crystalline deformation of the individual mineral grains, made possible by a motion of crystal dislocations that is similar to what occurs in crystal plasticity, except that the dislocation motion is time dependent, and the dislocations can move out of the plastic slip planes, with the help of atomic diffusion. This type of creep may be called dislocation creep. (In technical practice, distinctions between different types of dislocation creep are made.)

Dislocation creep is usually accompanied by re-crystallization, an important process that modifies the crystalline texture through lateral migration of the grain boundaries and through nucleation and growth of new crystal grains in the rock. By allowing new fresh crystals to grow at the expense of old deformed ones, re-crystallization serves to anneal the disturbances in the crystalline texture generated by intra-crystalline creep or plasticity. As a result of re-crystallization, deformed rocks commonly have a texture much different from the parent rock.

A second creep mechanism, which has been found to be important in the creep of ceramic materials at high temperature, is the diffusion of atoms along the grain boundaries, oı from one grain boundary to another through the intervening crystal grain. This diffusion has an effect much like dissolution of the grains on some sides and growth of the grains by precipitation of dissolved material on other sides. It allows the grains to change their external shape and thus allows the polycrystalline aggregate as a whole to deform, without any intra-crystalline deformation of the grains themselves. The resulting progressive deformation under stress is called diffusion creep. Because the creep rate is limited by the distance that the atoms must diffuse, it is inversely proportional to the grain size; fine-grained materials, with grain sizes in the micron range, are most prone to show diffusion creep.

It is not yet known whether the creep of rocks at depth in the Earth is primarily due to dislocation creep or to diffusion creep. The distinction is important, because diffusion creep is a strictly linear process giving a stress-independent creep viscosity, whereas most types of dislocation creep are markedly nonlinear (exponent n in the range **3** to **7**).

Creep of Earth's mantle

The effects of pressure and temperature on creep have an important influence on the flow of the Earth's mantle at depth, by means of which continental drift and sea-floor spreading take place. Dislocation motion in nonlinear creep is controlled by atomic diffusion around the dislocations, and it therefore follows that regardless of whether the operative mechanism is dislocation creep or diffusion creep, the effects of pressure and temperature on creep rates should be the same as the effects of these variables on the diffusion constant, for which there is experimental information and theoretical reasoning.

The diffusion constant (D) varies with temperature (T) and pressure (P) according to the exponential equation $D = D_o \exp [-(Q, + V_a P)/RT]$, in which Q_a is the activation energy, V_a is the activation volume, and R is the gas constant. For known activation energies and volumes, the diffusion constant increases strongly with temperature and decreases weakly with pressure. On this basis, and from the estimated variation of temperature and pressure with depth in the Earth, it is found that the effect of increasing temperature dominates the creep viscosity down to a depth of about 500 kilometres (300 miles). The predicted viscosity decreases by many orders of magnitude (multiples of 10) down to this depth, at which it has a broad minimum. Below this depth the effect of pressure begins to dominate and the predicted viscosity rises. It is thought that, at depths of 200 to 400 kilometres (100 to 200 miles), the creep viscosity is further reduced by partial melting of the rocks.

Response to shock waves. Impacts of meteoroids on the Earth, Moon, and other planets subject rock to severe stresses on short time scales (fractions of a second). Similar effects are produced by artificially generated shock waves, for which the duration is measured in thousandths of a second; shock pressures as high as 1,000,000 atmospheres have been generated. The point of mechanical failure for a rock is reached early in the shock event. Failure by brittle fracture is abundant both in shock-wave experiments and in cratering impact events: the rock is shattered into many fragments. There is often evidence, however, of plastic deformation (translation gliding, bending, twinning) of the mineral components. A novel and diagnostic type of shock failure is by partial or complete phase change: either by melting or by conversion to glass in the solid state or by transformation to high-pressure mineral phases. As a result of failure, rocks approach a hydrostatic condition at the peak pressure of the shock wave. Hence shock experiments can be used to obtain compressibilities up to very high pressures (see further METEORITE CRATERS).

Hardness and friability. The characteristics of hardness and friability (the resistance of a rock to crumbling into grains) are important in the practical usage of rock materials. Although the hardness of the individual minerals in a rock is a well-defined property, the aggregate, made up of grains of differing hardness, has no single, definite hardness. A rock such as granite, composed mostly of minerals of hardness 6 to 7 on the Mohs scale (in which talc is 1, quartz is 7, and diamond is 10), has a relatively definite, high hardness, whereas a sandy shale has attributes of hardness varying from 1 to 7. The ability of a rock to scratch or abrade other materials is a reflection, by and large, of the hardest abundant mineral component, whereas the extent to which the rock is scratched or abraded by other objects reflects the softest abundant component. Friability, the readiness with which mineral grains are separated from the rock, is also related to the weakest abundant component, usually the cementing mineral in a clastic rock.

THERMAL PROPERTIES

Specific heat. The heat capacity of a rock is a mass-weighted average of the heat capacities of its mineral constituents (*i.e.*, sum of the heat capacity of each mineral times its mass fraction). Figure 4 shows heat capacity as a function of temperature for important rock-forming silicate minerals and thus outlines the range of possible variation of the heat capacities of common silicate rocks. Because the mean atomic weight $\bar{A}$ of most rock-forming minerals is near 21, the high-temperature specific heat (Dulong-Petit value) is near 0.3 calorie per gram degree Celsius. This value is reached only above temperatures of about 1,000" C (1,800" F), because the strong interatomic bonds, especially silicon–oxygen, inhibit excitation of some of the thermal vibrations at lower temperatures. The heat capacity is roughly half the limiting Dulong-Petit value at room temperature. The presence of water, either mechanically included (in pores) or in chemical combination (hydrous minerals), adds appreciably to the heat capacity; anomalously high apparent heat capacities are shown over ranges of temperature in which water is driven off on heating, either by volatilization or by thermal breakdown of hydrous minerals.

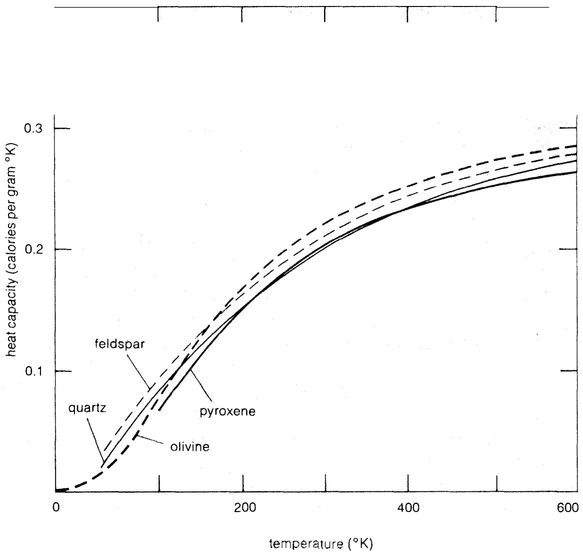


Figure 4: Specific heats (at constant pressure) of rock-forming minerals, as a function of temperature. Weighted averages of these curves give the specific heats of common rocks.

Thermal conductivity. The rate at which heat escapes from the Earth's interior to the surface is controlled by thermal conductivity, and this property affects the suitability of rock as a building material, for which its insulating capability is important. The conductivity of a rock is a somewhat complex average over the conductivities (generally anisotropic) of its component minerals modified by effects of porosity. As shown in Table 6, crystalline silicate rocks have conductivities generally in the range from about 0.004 to about 0.008 calorie per centimetre degree Celsius second (abbreviated cal/cm°C sec), the lower values being typical of silicate rocks relatively rich in magnesium and iron, such as basalt and gabbro, whereas the higher values are typical of rocks rich in silica and alumina, such as granite and granodiorite. The variations reflect the fact that the conductivities of quartz are relatively high (—0.020 cal/cm°C sec), those of feldspar are low (~0.005), and those of chain and sheet silicates are intermediate (—0.007). Porosity reduces the thermal conductivity by interfering with the thermal contact between mineral grains; thus porous sandstones have conductivities lower than quartzite (see Table *6*), and the conductivities of dry soils are especially low (—0.0004). Conductivities of individual specimens vary by as much as 20 percent from the average values in Table 6, probably to a large extent because of variations in the effects of crack porosity. By eliminating crack porosity of rocks and bringing the grains into more intimate contact, confining pressure causes an increase in conductivity, which for most rocks amounts to 10 to 20 percent for a pressure of 1,000 atmospheres. Beyond this effect, pressure has little effect on conductivity. Saturation of the pores with water causes an increase in conductivity similar to that obtained by closing the pores under pressure. The conductivity of mica is about six times greater parallel to the micaceous sheets (mica layers) than perpendicular to them; hence, rocks with abundant, well-aligned mica flakes have anisotropic conductivity; this is illustrated by gneiss and phyllite in Table 6.

Porosity and thermal conductivity

Table 6: Thermal Conductivities of Rocks
(mean values)

	thermal conductivity k(cal cm^{-1} °C^{-1} sec^{-1})	
	at 20° C	at 200° C
Granite	0.0078	0.0066
Granodiorite	0.0071	0.0057
Diabase	0.0053	0.0051
Gabbro	0.0051	0.0050
Anorthosite	0.0042	0.0044
Dunite	0.012	0.0081
Basalt	0.004	0.004
Gneiss		
∥ Foliation	0.0082	0.0074*
⊥ Foliation	0.0059	0.0055*
Phyllite		
∥ Foliation	0.0183	—
⊥ Foliation	0.0079	—
Marble	0.0073	0.0052
Quartzite	0.015	0.009
Limestone	0.006	—
Sandstone	0.006 ±	—
Shale	0.0045 ±	—

*At 100° C.

Glassy rocks, like glasses generally, have a low thermal conductivity (—0.002 cal/cm° C), which increases with temperature. In contrast, increasing temperature reduces the conductivities of most crystalline materials, as shown by the data for rocks in Table 6. The decrease in conductivity is large for quartz, but for feldspars it is small, and, in fact, calcium-rich feldspar shows a small increase, indicated by the data for anorthosite in Table *6.* These differences in behaviour account for the fact that sialic rocks show a larger drop in conductivity with temperature than do mafic rocks. The drop in conductivity tends to cause the geothermal gradient to increase with depth in the Earth, for a constant flux of heat escaping outward. Above temperatures of about 1,600" C (2,900" F), however, conductivities begin to increase with temperature, because transport of heat by thermal radiation through the solid becomes significant; radiative transfer probably makes an important contribution to heat conduction in the Earth's mantle.

Thermal expansion. The increase in volume of a rock specimen with temperature is expressed in terms of the coefficient of volumetric thermal expansion. The coefficient α is equal to the reciprocal of the specimen volume (V) times the ratio of the increase in specimen volume (ΔV) that occurs for a small temperature increase (AT) to that temperature increase, or $\alpha = (1/V)\ (\Delta V/\Delta T)$. Similarly the increase AL in linear dimension (L) of the specimen is expressed in terms of the coefficient of linear thermal expansion $\Delta L/L\Delta T$. It is equal to $\alpha/3$, for isotropic specimens. The volumetric expansion is of particular importance in estimating the density of rocks at the high temperatures within the Earth and in assessing the possibility that temperature gradients in the Earth will cause convective motion analogous to the convection of a viscous liquid. The linear expansion is important where rock must remain fitted to other structural materials over wide variations in ambient temperature. Most rocks have a volumetric expansion coefficient in the range 1.5–3.3 × 10^{-5} per degree Celsius under ordinary conditions. Quartz-rich rocks, especially sandstone and quartzite, have the highest values, indicating a relatively high thermal expansion of quartz itself. Thermal expansion coefficients increase substantially with temperature. At the microscopic level, the thermal expansion of a rock is a complex affair, because of the generally anisotropic character of the expansion of the individual minerals and the differences in expansion from grain to grain. A complex pattern of stresses is set up in the grains, controlled by the thermal expansions and the various elastic properties. To some extent the stresses are relieved by opening up of microscopic cracks, and a part of the measured expansion, probably about 10–20 percent, thus represents increase in crack porosity. Crack opening is not entirely reversible, so that the contraction on cooling generally is less than the expansion during the previous heating, and the expansions on successive heatings may differ. Because of the cracking, piolonged thermal cycling causes mechanical disintegration of rocks into their granular components. These effects are inhibited by confining pressures sufficient to keep the cracks closed.

Thermal cycling

Radioactive heat productivity. Of importance in determining the temperature and thermal evolution of the Earth's interior is the heat produced by absorption of radiation that is emitted from radioactive elements in the rocks. Significant contributions come from the nuclear decay of potassium-40, thorium-232, uranium-235, and uranium-238. In Table 7, average rates of heat production are given for several major rock types, including chondritic meteorites (meteorites imbedded with nodules of certain materials), which are considered to represent a possible composition for the Earth's mantle. The great enrichment of the radioactive heat producing elements in the more sialic rock types of the Earth's crust is evident in Table 7 and has important consequences in the thermal history of the Earth, the generation of magmas, and volcanism. The average geothermal heat flow at the Earth's surface (1.2 × 10^{-6} calorie per square centimetre per second) could be generated by a layer of granite only 16 kilometres (ten miles) thick, somewhat thinner than the typical thickness of the continental crust, whereas some 300 kilometres of chondritic material would be required.

Table 7: Radioactive Heat Generation by Rocks

	heat production (10^{-6} cal g^{-1} $year^{-1}$)			
	from U	from Th	from K	total
Granite	3.4	4.0	1.1	8.5
Granodiorite	1.9	1.8	0.7	4.4
Diorite	1.5	1.7	0.3	3.5
Gahbro, basalt	0.7	0.5	0.1	1.3
Dunite, eclogite*	0.001– 0.04	0.0002– 0.04	0.0002– 0.01	0.001– 0.09
Chondrite	0.009	0.009	0.023	0.04
Sandstone*	2.2	1.2	0.4	3.8
Shale	2.7	2.4	0.7	5.8
Limestone	1.6	0.3	0.1	2.0

*Shows very wide variation.

Melting. Rocks begin to melt at temperatures somewhat lower than the melting points of any of their constituent minerals, and melting occurs over a range of temperatures, from the first appearance of liquid to the final disappearance of the last crystals. Melting points of the common anhydrous (non-water-bearing) rock-forming silicates lie between 1,100" and 1,800" C (2,000" and 3,300" F). The lowest melting mixture of these has the composition of granite (mainly quartz and potassium-sodium feldspar) and begins to melt at 950" C (1,750" F). The temperature of first melting does not depend on the relative proportions of the different minerals present and hence is the same for granitic rocks of diverse compositions, representing different proportions of quartz, potassium feldspar, and sodium feldspar. The temperature of first melting does depend, however, on which minerals are present and on their individual compositions. Thus the more mafic igneous rocks such as diorite and basalt, which lack quartz and potassium feldspar and which contain a calcium-rich sodium feldspar, begin to melt at higher temperatures, approaching 1,100" *C* (2,000" F). Ultramafic rocks such as peridotite, containing only pyroxene and olivine, begin to melt at 1,350"–1,550" C (2,450"–2,800" F). The completion of melting is sensitive to the mineral proportions in the rock. Most granites have a narrow melting range, because their bulk compositions are close to that of the first-formed melt. Increasing the proportion of a mineral conrtituent of high melting point widens the melting range by raising the completion of melting. Thus mafic rocks, containing substantial amounts of high-melting iron-magnesium silicates, have melting ranges 200"–300" C (400"–500" F) wide.

Melting of rocks is greatly affected by the presence of water vapour under pressure, because water can dissolve in silicate melts. Increasing the water-vapour pressure to a few thousand bars lowers the first melting of granite from 950" to 650" C (1,750" to 1,200" F) and causes water to dissolve in the melt to the extent of several percent. The effect of water-vapour pressure in lowering the melting range of basalt and the first melting temperature of granite is shown in Figure 5. Similar considerations apply to the melting of metamorphic and sedimentary rocks of similar compositions.

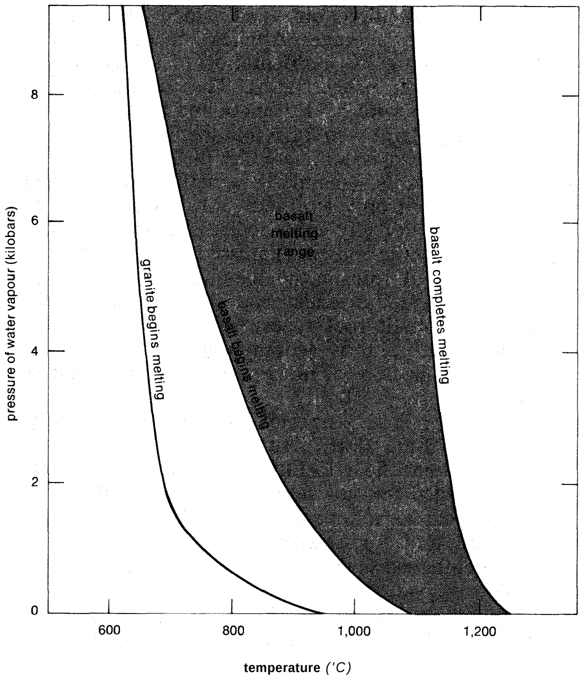


Figure 5: Melting range for basalt (shaded) and temperature of beginning of melting for granite, as a function of water-vapour pressure.

Importance of water vapour under pressure

Another important effect of water vapour under pressure is to increase the stability of hydrous minerals, such as mica and amphibole. When heated at atmospheric pressure, these minerals break down to anhydrous minerals (with release of water) at temperatures of only 500"–700" C (900"–1,300" F), well below the start of melting. Raising the water-vapour pressure both lowers the onset of melting and raises the stability of the hydrous minerals against breakdown, so that it becomes possible for them to melt or, on cooling, to crystallize from a rock melt. The fact that these minerals actually formed in many igneous rocks at the time of initial crystallization shows that substantial pressures of water vapour must have operated on the parent magmas.

ELECTRIC AND MAGNETIC PROPERTIES

In deciphering the history of the Earth's magnetic field and the wandering motions of the continents and in estimating temperatures in the Earth's interior, electric and magnetic properties of rocks play an important role. They are also utilized in methods of geophysical exploration.

Magnetic susceptibility. Application of a magnetic field (H) to a rock specimen induces a magnetization (M) that is generally proportional to the field; the proportionality constant (symbolized by χ, the Greek letter chi) in the equation $M = \chi H$ is the magnetic susceptibility (see MAGNETISM). Substances are classified as paramagnetic or diamagnetic accordingly as χ is positive or negative. In a paramagnetic material, the directions of the inducing magnetic field and the induced magnetization are the same, whereas in a diamagnetic material they are opposite. Most minerals are weakly diamagnetic, with magnetic susceptibilities of about —10-6. Minerals containing iron, manganese, and related elements (transition elements in the periodic table), whose atoms act as individual little magnets, are paramagnetic and typically have magnetic susceptibilities of about $+10^{-4}$. Certain paramagnetic materials typified by iron and the common magnetic mineral magnetite have abnormally large susceptibilities (χ is approximately 1 to 10^{+3}) and are called ferromagnetic. In these materials, the individual atomic magnets (*e.g.*, iron atoms) align themselves spontaneously to produce a magnetization even in the absence of an inducing field. A ferromagnetic mineral normally shows little or no net magnetization in bulk because the spontaneous magnetization occurs in microscopic regions, called domains, the magnetic orientations of which point in diverse directions, and which therefore cancel out on the average. Application of a magnetic field causes progressive reorientation of the domains, inducing a net bulk magnetization. The corresponding susceptibility varies with field strength and magnetic history, but it is always large—so large that the bulk magnetic susceptibility of rocks is dominated by their content of ferromagnetic minerals (mainly magnetite), even though these are present only as minor constituents. The magnetic susceptibility χ of most rocks is approximately 0.3 times the magnetite content expressed as volume fraction. Because of regularities in the chemical processes of rock formation, magnetite tends generally to be more abundant in more mafic rock types; hence the susceptibilities of basalts and gabbros (commonly $\chi \sim 10^{-3}$) are as a rule larger than those of most granites, granitic gneisses, and sandstones $(\chi < 10^{-4})$.

Rocks of higher than normal magnetic susceptibility at depth beneath the Earth's surface tend to enhance the Earth's magnetic field locally in the same way that an iron core enhances the field of an electromagnet. The resulting local variations in the Earth's field provide a geophysical tool for assessing the distribution of rock types at depth, especially ore bodies containing magnetic minerals.

Remanent magnetization. The variety of magnetite called lodestone is the oldest known example of a ferromagnetic material in which remanent magnetization is retained in the absence of an inducing magnetic field, forming a permanent magnet. Remanent magnetization results from some non-randomness in the orientations of magnetic domains. Rocks containing ferromagnetic minerals can acquire a remanent magnetization that results from the magnetizations of the individual grains (see ROCK MAGNETISM).

Significant remanent magnetism in igneous rocks is acquired by a process called thermoremanent magnetization, which develops as the rock cools from high temperature. Above a critical temperature, called the Curie temperature (578° C [1,072° F] for magnetite), thermal agitation disrupts the orderly alignment of atomic magnets in a ferromagnetic mineral, so that the ferromagnetism is destroyed and χ decreases abruptly to values of about $+10^{-4}$, typical of paramagnetic minerals. At the Curie temperature, a remanent magnetization develops parallel to and proportional to the ambient magnetic field. Because the bulk magnetism of rocks is contributed by several different minerals and because grain size and shape also influence magnetic behaviour, the thermoremanence is actually acquired over a wide temperature range from 600" C (1,100" F) down to about 300" C (600° F). An important empirical principle is that the partial thermoremanent magnetization acquired over any particular cooling interval is independent of that from other intervals and is destroyed by heating over the original cooling interval. Thermoremanent magnetizations of basalts are typically about 1 percent of the ambient magnetic field. For the same composition, coarser grained rock should have lower remanence.

Depositional magnetism

Sedimentary rocks can acquire magnetization during their original deposition, by the settling of magnetized particles from water suspension, the particle orientation being controlled by the Earth's field. This is called depositional magnetization. The particles are commonly elongated grains of magnetite, magnetized parallel to their long dimension. Because the particles tend to settle with their long axes horizontal, depositional magnetization tends to ignore the dip of the magnetic field, but it records the direction of the field in the horizontal plane at the time of deposition.

Rock magnetization provides a record of the Earth's magnetic field as it existed at the time of rock formation. This record tends to be obscured by magnetization components impressed later, under changed magnetic fields. Fortunately, however, these later components can often be removed by suitable heat and alternating-field demagnetization treatments without destroying the original magnetization. The permanence of a given magnetization is measured in terms of the corresponding coercive field, which is the strength of the reversed magnetic field required to reduce the bulk magnetization to nil. For the thermoremanent magnetization in bulk samples of magnetite, the coercive field is about 20 oersted units, which, although substantially larger than the Earth's present field (about 0.5 oersted), is considered not stable by comparison with what is achieved in other materials. It appears nevertheless that the magnetite in rocks contributes to the stable thermoremanent magnetization, probably because much of the magnetite is present as grains smaller than the domain size; if elongated in shape, these one-domain grains have a high coercive field (about 500 oersted).

The thermoremanent magnetization acquired by large masses of basalt in the oceanic crust is strong enough to modify measurably the Earth's magnetic field in their vicinity. In certain strip-shaped areas of the ocean floor, the basalts of the oceanic crust are magnetized essentially parallel to the Earth's present field direction, whereas in intervening strips they have the reversed direction of magnetization because they formed at times when the Earth's field was reversed. The magnetic fields from these oppositely magnetized rock masses set up local variations amounting to about 1 percent of the total field. These variations, readily detected by sensitive ship- or air-borne magnetometers, provide a basis for mapping the magnetized strips and inferring the history of formation of the oceanic crust.

Electrical conductivity and dielectric behaviour. The common rock-forming minerals (silicates, oxides, carbonates, sulfates) are electrical insulators, having resistivities higher than about 10^8 ohm metres (units of ohm times metre) at ordinary temperatures. Many ore minerals (sulfides and sulfosalts) and some oxides (notably magnetite and ilmenite, which are common accessory minerals in rocks) are electronic semiconductors, with resistivities in the range of 10^{-4} to 10^4 ohm metres. Metallic conductors are rare in nature, with the exception of graphite. The electrical conductivity of rocks depends on mineral content, texture, and porosity in a way similar in most respects to the thermal conductivity. Crack porosity increases the resistivity of dry rocks. A conductive minor mineral component such as graphite, magnetite, or pyrite gives only a modest increase in bulk conductivity if present as isolated grains but gives a large increase if distributed in interconnected networks, such as interstitial cementation in a clastic texture, intersecting fracture fillings, or dendritic (branching) growth structures. Resistivities of slates, for example, are normally several thousand ohm metres but can be reduced to less than one ohm metre by a content of graphite and pyrite amounting to only a few percent.

Resistivities of dry igneous rocks approximate 10^{10} ohm metres at room temperature. Above about 500" C (900" F), the resistivity begins to drop rapidly with temperature. Atoms displaced by thermal agitation out of their stable sites in mineral structures can diffuse through the structures under the force of the electric field, and at high temperatures this type of ionic or electrolytic conduction dominates the electrical behaviour of insulating minerals. Its temperature dependence follows the exponential form for thermally activated processes: $1/\rho =$ A $\exp(-Q/RT)$, in which $1/\rho$ is conductivity [reciprocal resistivity, expressed as $\text{ohm}^{-1}\ \text{m}^{-1}$—that is, 1/(ohm metre)], A is a constant, Q is the activation energy for the conduction process, and R is the gas constant (2 calories per mole per degree Kelvin). For olivine, which is thought to be a major constituent of the outer part of the Earth's mantle (in the rock peridotite), the constant A is 5 $\text{ohm}^{-1}\text{m}^{-1}$ and the activation energy is 16 kilocalories per mole; the resistivity falls from about 10^6 ohm metres at 200" C (400" F) to about 10^2 ohm metres at 1,000" C (1,800" F). The resistivities of rocks in the Earth's interior can be calculated from the magnetic effects of currents induced in the interior by current flow in the ionosphere. Resistivities so determined decrease from about 103 ohm metres in the lower part of the crust to one ohm metre in the upper mantle, at a depth of about 600 kilometres (400 miles), and to the order of 10^{-2} ohm metres in the lower mantle. This is doubtless the effect of temperature increase inward. Because ionic conductivity, like diffusion, has a substantial activation volume, it is inhibited by the high pressures in the mantle, and it therefore seems likely that the high conductivities of rocks in the deep mantle are due to appreciable electronic semiconduction at the high temperatures prevailing there.

Resistivity in the Earth's interior

In the shallower parts of the crust, at depths less than about six kilometres, where because of the relatively low confining pressure (less than 2,000 atmospheres) the rocks retain some crack porosity, their electrical properties are dominated by the presence of water. Pure water is a reasonably good insulator (its resistivity is 2×10^5 ohm metres), but the water present in rocks (groundwater, connate water [water trapped during rock formation]), is moderately conductive due to the presence of dissolved salts; resistivities are commonly in the range one to ten ohm metres and less commonly as low as that of seawater (0.2 ohm metre) or even lower. Except for rare rock types containing substantial amounts of conductive minerals, the resistivities of water-saturated rocks do not depend on the mineral resistivities but are instead proportional to the resistivities of the contained water and depend sensitively on the amount and nature of the pore space. It is found empirically that for a given type of porosity, the resistivity p varies approximately as the inverse square of the porosity η: p $=$ a $\rho_w\eta^{-2}$. Here a is a constant and ρ_w is the resistivity of the contained water. The empirical constant a depends on the type of porosity: for clastic rocks (interstitial porosity) it is about 0.7, whereas for volcanic rocks it is larger ($\sim$1.5), indicating the relatively poor connectivity of the vesicular pore space. Crack porosity is the most effective source of rock conductivity, the value of the constant a for crystalline rocks being about 0.3. It has been found that the increase in crack porosity that occurs in stressed rocks just before brittle failure can be detected by a corresponding decrease in resistivity, thus providing a possible method for anticipating the occurrence of fracture in the Earth, and hence the resultant earthquakes and earth tremors.

Earthquakes predicted by resistivity

For rocks only partially saturated with water, which occur above the water table, the resistivity increases as expected for a reduced porosity corresponding to that portion of the pore space actually occupied by water. Beyond a certain reduction in water content, corresponding to about half of the total porosity, the resistivity increases more rapidly, probably because the connectivity of thin films of water between the mineral grains becomes severed as the water content becomes increasingly lower.

Information about the average resistivities of rocks down to depths of about 100 metres (300 feet) in the Earth is obtained by electromagnetic measurements at radio frequencies. It is found that younger sedimentary rocks generally show resistivities less than 40 ohm metres, older sedimentary rocks and volcanic rocks are intermediate (40–100 ohm metres), and crystalline rocks of low porosity show resistivities greater than 100 ohm metres.

In addition to setting up current flow, an electric field (E) causes the displacement of bound charges in a material medium, creating a state of electric charge unbalance called the polarization (P). It is analogous to the magnetization produced by a magnetic field and obeys a similar linear relation between polarization and the electric field producing it: P $= \alpha E$, in which α is the polarizability of the medium, analogous to the magnetic susceptibility χ. The constant a may be calculated directly from a measured quantity, the dielectric constant, E $= 1 + 4\pi\alpha$. The dielectric constant is highest in a static (steady state) electric field and generally decreases in alternating fields of progressively higher frequency. At the frequencies of visible light, the dielectric constant equals the square of the refractive index and ranges from 2.3 to 3 for the common rock-forming minerals; at radio frequencies, E is generally two to three times larger. The corresponding bulk dielectric constants for rocks are given to a reasonable approximation by a volume-weighted average over the mineral components. The effect of porosity can be included by treating the pore space as a component with dielectric constant 1. Water-saturated pore space can also be included, provided account is taken of the great dielectric dispersion (change of ε with frequency) of water near 10^{11} hertz, the dielectric constant dropping from a value of about 85 at lower frequencies to about 5 at higher frequencies.

Rocks containing appreciable amounts of moisture, even absorbed moisture, begin to show departures from the above type of behaviour at frequencies less than about 10^6 hertz. Below about 10^4 hertz, the apparent dielectric constant measured by standard techniques begins to increase rapidly with decreasing frequency; the values reported at low frequencies, below one hertz, are prodigious (10^5 to 10^8), higher by many orders of magnitude than the static (zero hertz) dielectric constants of any of the mineral components or water. The high apparent values are probably caused by the flow of electric currents that are unable to pass freely through the rock, owing to internal obstructions along the conductive water films in the rock and to electrochemical blocking either internally or at the measuring electrodes. When the currents are blocked in this way, they result in what appear to be displacements of bound charge and give a large apparent contribution to the dielectric constant E. The product $\varepsilon\rho\omega$ (in which ρ is the bulk resistivity and ω is the frequency), is found to be nearly frequency-independent over a wide frequency range, from 10^3 hertz to as low as 10-3 hertz. Because the resistivity appears frequency-independent, the apparent dielectric constant at low frequencies thus varies inversely as the frequency. The reciprocal product $1/\varepsilon\rho\omega$ is known as the loss tangent and is the fractional power dissipated per cycle of oscillation of the electric field; it is thus analogous to the specific damping capacity b for elastic waves, which for rocks shows a similar frequency independence. Loss tangents are more useful than dielectric constants for characterizing rocks electrically at frequencies below 1,000 hertz.

HYDRAULIC CONDUCTIVITY

Although porosity indicates the capacity of a rock to store fluids such as water or petroleum, the usefulness of this storage capacity depends on the readiness of extraction or recharge, which is controlled by the permeability

Darcy's law

(hydraulic conductivity). This is defined in terms of the volume of fluid that will flow per unit time through a rock cross-sectional area under a specified pressure gradient driving the fluid forward. The driving pressure is the excess over the simple hydrostatic pressure that would be present in the fluid at rest and is expressed in terms of the excess fluid head (h). Darcy's law (after Henri Darcy, a French engineer) states a linear relationship between the volume of fluid and the pressure gradient (the decrease in pressure with distance x in the direction of flow, —*dh*/*dx*) along the fluid flow path: the volume (Q) of fluid flowing through a rock per unit time is equal to the cross-sectional area (A) times the permeability (K) times the driving pressure gradient—*i.e.*, Q — — AK *dh*/*dx*. In engineering practice, ***Q*** is expressed in gallons per day and ***A*** in square feet; so that, if h and x are taken in the same units, the permeability K has dimensions gallons per day per square foot, a unit also called the Meinzer (after Oscar Meinzer, an American hydrologist). The values of K in these units vary from less than 10^{-5} for crystalline rocks to 10^{6} for gravel; some intermediate values for the permeability of various rock materials to water at *25"* C (77" F) are sandstone 10^{-2}–102, shale 10^{-6}–10^{-4}, limestone 10^{-4}–10, and sand 10 2–104. Permeabilities for other fluids, or for water at other temperatures, can be obtained from the fact that the permeability is inversely proportional to the fluid viscosity.

Although the more porous rocks generally have higher permeability, the relationship is greatly affected by the type of porosity: interstitial porosity in a clastic texture, where the pore spaces interconnect, gives a much greater permeability than does vesicular porosity. For porosities of a given geometrical type, the permeability varies directly as the square of the pore dimensions, hence coarse detritus and coarse-grained clastic rocks have much higher permeabilities than their finer-grained counterparts. The strong dependence on pore size distinguishes hydraulic conductivity sharply from electrical conductivity. The wide range of observed permeability values for rock materials of a given type (*e.g.*, 10^{-2} — 10^{4} for sand) indicates the influence on pore sizes of various factors in addition to clast size: size sorting, compaction, clast shapes, cementation, etc. Thus clean sands have much higher permeabilities than clay-rich sands, whose interstices tend to be blocked by clay. Crystalline rocks have not only small total porosity, but also thin pores (cracks), and the permeability of sound specimens is for practical purposes nil. Masses of crystalline rock undergound, however, usually show some bulk permeability because of the presence of macroscopic fractures (joints, faults) on a coarse scale; this type of fracture permeability cannot be assessed primarily from laboratory tests on small rock samples.

OPTICAL PROPERTIES

Colour effects

The effects of rocks on visible light are best considered in terms of the optical properties of the individual component minerals, and are best observed in slices of rock about 30 microns (3 × 10^{-5} metre) thick. To the naked eye, the striking optical feature of many rocks is their colour, which derives from the colour of one or more mineral constituents. Striking colours are often caused by minor and petrologically insignificant mineral components, as, for example, in the pink colour of some granites. In most common rocks the colour effects are dominated by iron-containing minerals. Reds, browns, and ochre yellows are caused by oxides or hydrous oxides of ferric iron, and greens by silicates of ferrous iron or by finely divided silicates or hydrous silicates of both ferrous and ferric iron. Less commonly, other metal ions with distinctive absorption spectra in the visible range are responsible for striking rock colourations, when there is local enrichment of the comparatively rare minerals containing these ions: thus blue from copper, pink from manganese or cobalt, green from chromium or nickel, yellow from cadmium, orange from chromium or molybdenum, red from mercury, etc.

The lustre of rock surfaces depends on the lustres of the component minerals, modified by the effects of grain size: coarse-grained minerals display their inherent lustre, whereas fine-grained aggregates generally have a dull lustre. Rock weathering produces a dull lustre because it involves chemical alteration of the original coarse-grained minerals to aggregates of clay and other fine particles.

BIBLIOGRAPHY. S.P. CLARK, JR. (ed.), *Handbook of Physical Constants* (1966), is a basic reference that contains an exhaustive compilation of measured data on physical properties of rocks and minerals, with a brief introductory discussion of each of the properties treated. J.C. JAEGER, *Elasticity, Fracture and Flow, with Engineering and Geological Applications,* 3rd ed. (1969), gives a general discussion of mechanical properties and their application to the calculation of stresses and deformations in the Earth. **A** more detailed treatment of mechanical properties, with emphasis on current experimental studies and theoretical concepts of brittle failure, is contained in C. FAIRHURST (ed.), *Failure and Breakage of Rock* (1967); and in K.G. STAGG and O.C. ZIENKIEWICZ (eds.), *Rock Mechanics in Engineering Practice* (1968). These works also discuss engineering applications of the mechanical properties, and methods of testing, particularly of rock masses *in situ.* Experimental studies of rock plasticity and creep pertinent to geological phenomena are collected in *Rock Deformation,* ed. by D.T. GRIGGS and J. HANDIN (1960). A recent survey of creep properties of rocks in relation to the flow of the Earth's mantle is given by J. WEERTMAN, "The Creep Strength of the Earth's Mantle," *Rev. Geophys. Space Physics,* 8:145–168 (1970); the approach is primarily theoretical, but contains references to recent experimental studies. The behaviour of rocks under shock loading and the structural changes produced in rocks by shock waves are treated by B.M. FRENCH and N.M. SHORT (eds.), *Shock Metamorphism of Natural Materials* (1968). Magnetic properties of rocks and some of their applications are discussed by D.W. STRANGWAY, *History of the Earth's Magnetic Field* (1970). Electrical properties, particularly in relation to their use in exploration geophysics, are discussed by G.V. KELLER, "Electrical Characteristics of the Earth's Crust," in J.R. WAIT (ed.), *Electrontagnetic Probing in Geophysics* (1971).

(W.B.K.)

Rocky Mountains

The major section of the great upland system that dominates the western North American continent, the Rocky Mountains stretch from northern Alberta and British Columbia southward through the western United States to Mexico, a distance of some 3,000 miles. In places the system is several hundred miles wide. Limits are mostly arbitrary, especially on the western side, where other mountain systems, generally excluded, exist (see Figure 1). The explorations of Silvestre Vélez de Escalante, Lewis and Clark, and John Wesley Powell caught the imagination of the world, and through their reports the knowledge of the Rockies began to unfold. The snow-capped peaks, the conifer forests, the wide intermountain valleys, the crystal-clear streams, the big sky, and a vast mineral resource provide homes and work for about 5,000,000 people, and millions more come each year to tour and play. The Rockies comprise one of the country's most popular tourist attractions.

The basic divisions

To the west, especially in Nevada, western Utah, and Arizona, lies the Great Basin, or the Basin and Range Province (*q.v.*). Here, the earth's crust has been broken by numerous faults, with the blocks between faults uplifted, depressed, and tilted. This manner of deformation, geologically speaking, has been a rather recent affair, and has produced a kind of relief and drainage entirely different from that of the more typical Rocky Mountains to the east. As a result, it is necessary to recognize that there are two basic divisions of the Rocky Mountains. The first is that of the map area (see Figure **2**), in which the ranges retain characteristics of their original structure and shape, and the second, that west of the map area, where the original structures and forms have been broken and much defaced by block faults. In fact, there is much controversy about what the original structures of the western division were like.

This article deals with fundamental themes in the geological evolution of the Rockies, with particular emphasis on the basic structural elements from which the mountains of today have been formed. The article NORTH AMERICA treats the development of the Rockies as a factor in continental evolution, and the articles BASIN AND RANGE PROVINCE; SIERRA NEVADA RANGE; PACIFIC COAST RANGE; and ALASKAN MOUNTAINS describe other component ranges of the western mountain system. Articles on the GRAND CANYON and the major rivers of the area provide more detailed information regarding individual physical features of the Rockies. The article NORTH AMERICAN DESERT contains descriptions of the plant and animal life of the region. Full treatment of the contemporary landscapes of the region will be found in the article UNITED STATES: *Landscape* and in the articles on individual states.

Ranges and basins of the Eastern division

Eastern division. A study of Figure 2 gives an indication of the number of ranges in the eastern division. Each one is a large uplifted mass; many have high peaks and dramatic scenery. The uplifts or ranges are named on the map (Figure 2), as are the better known basins. The Front Range of Colorado and the central Arizona uplift are the two most massive uplifts, but most of the rest have a somewhat uniform area, about 62 miles (100 kilometres) long and 15 miles (25 kilometres) wide. The Front Range supports peaks over 14,000 feet high. including Mt. Elbert, which at 14,431 feet is the highest point in the Rockies. Wyoming has peaks in the Wind River and Teton Ranges well over 13,000 feet high, as does Utah in the Uinta Mountains. Some of the high mountain areas have remained virtually untouched by man and have been set aside as national wildernesses. All have national forests.

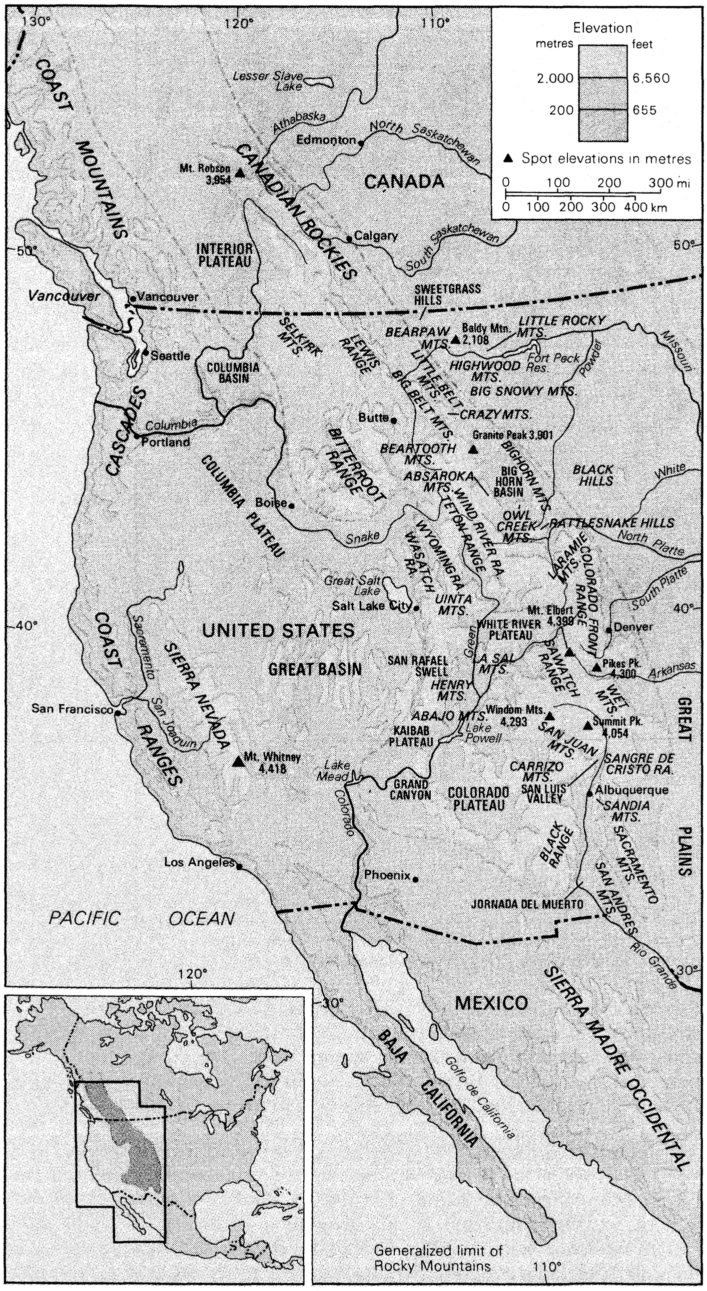


Figure 1: The Rocky Mountain ranges in the United States and Canada.

The basins of the eastern division are broad and generally from 4,000 to 7,000 feet high. During the period of the uplift of the ranges and for some time afterward the basins were the receptacles of the rock debris eroded from the uplifts. Later, as drainage became well established to the Missouri and Colorado rivers, the basin sediments were cut into by water action and were in some spots deeply eroded.

Western division. The Canadian and northwestern Montana Rockies, including the mountains of Glacier National Park and the Lewis Range of northwestern Montana, are a subdivision of the western division. They are characterized by a series of parallel ridges, which resulted generally when thick sections of sedimentary rocks were thrust on top of each other. The same tyye of linear mountain ridges is noted in western Wyoming and southeastern Idaho (see Figure 2), but from the southwestern corner of Wyoming to the southwestern corner of Utah and southern Nevada the western division is masked by the later Basin and Range block faulting. How far west in western Utah and Nevada the true Rocky Mountains once extended is still regarded as controversial.

In summary, therefore, the eastern division of the Rockies is marked by blister-like uplifts and large intermontane basins and the western division by thrust faults and folds.

HISTORY OF SCIENTIFIC STUDY

Early explorers

Catholic missionaries had worked their way northward from Mexico into New Mexico by the middle of the 18th century, and in 1776 Padre Escalante and his party explored and documented their travels into what is now Utah, reaching almost to the Great Salt Lake. The Lewis and Clark expedition in 1803–06 explored and charted a route up the Missouri River into Montana and thence across Idaho and Oregon to the Pacific. Following the Missouri River into east-central Montana, Jedediah Smith worked his way southward into the Big Horn Basin and thence into southeastern Idaho, northern and southwestern Utah, southern Nevada, around the Sierra Nevada, and back to the Great Salt Lake across the Great Basin. His journeys occurred in the years 1822 to 1831 and were possibly the most remarkable of all western explorations, but unfortunately Smith wrote very little. John Frtmont's explorations occurred in 1846–48: he followed the North Platte River into Wyoming, up the Sweetwater River to the south end of the Wind River Range (South Pass), and thence southwestward into Utah. He then explored northward into Idaho and around the broad north end of the Great Basin into Oregon and down into California. This was an important scientific survey because he charted distances, determined latitudes, longitudes, and elevations and recorded objectively in some detail what he saw.

Fdur great western surveys were then organized by the federal government, namely that of Clarence King (the 40th Parallel Survey of 1867–78), that of Ferdinand V. Hayden (geological survey of Nebraska and Wyoming 1867–78), that of George M. Wheeler (100th meridian, 1872–79), and that of John Wesley Powell (exploration of the Colorado River and of Utah, Arizona, and southern Nevada, 1869–78). The maps and preliminary observations of these important surveys laid the groundwork for a great mass of research that followed. The Rocky Mountains, with their abundant coalfields, numerous oil, gas and uranium prospects, and a wealth of metal prospects have provided the U.S. Geological Survey with much work over the past century. Many of the prospects have turned into large and small mines, making a substantial contribution to the regional and national economies. The mountains yield lumber and; by the second half of the 20th century, were offering summer and winter recreation facilities of increasing importance. They also yield water, of paramount importance in this semi-arid and arid land.

GEOLOGICAL EVOLUTION

Sedimentary provinces. The eastern division as defined above is one in which the layers of sedimentary rocks deposited in seas of the Paleozoic and Early Mesozoic Era (from about 550,000,000 to 225,000,000 years ago) are relatively thin, while those of the western division are thick. The western division is characterized as geosynclinal (resulting from a vast downwarping in the earth's surface), whereas the thinner eastern deposits are of a peripheral (shelf) type. The sedimentary layers of the uplifted ranges of the shelf vary in thickness from a few feet to 6,000–7,000 feet, whereas those of the western geosyncline exceed 40,000 feet in places.

Age and thickness of deposits

As the beds of the downwarp began to be uplifted, folded, and thrust faulted under the pressure of earth movements in Late Mesozoic time (during the Jurassic and Cretaceous periods of 136,000,000 to 190,000,000 years ago), their waste products, formed by the erosive action of wind and water, were carried eastward and deposited on the adjacent shelf in considerable volume. In fact, a foredeep basin (*i.e.*, one in proximity to the great downwarp) in western Wyoming and central Utah subsided and received over 10,000 feet (3,080 metres) of the sediments, which consequently spread eastward in diminishing amounts.

Structures of the western divisions. The sediments of the geosyncline were first deformed by vast earth movements in Jurassic time; with the deformation, a broad upland was created where marine waters had once existed. As has been noted, the sediments eroded from the upland were deposited in the foredeep basin to the east, where marine seas advanced and retreated and in which marine animals lived. It is the fossils of these animals, trapped in the sediments, that indicate to geologists the age of the foredeep beds and thus, by inference, of the uplift to the west.

The exposed rocks of the ranges of the Great Basin reveal that the chief structures formed at this time of massive changes in the earth's crust were great folds and sheets of rock, with some beds thrust and folded over each other. The thrust sheets of this type in southeastern Idaho and western Wyoming appear to be large slices of the sedimentary rocks that started on the Cache uplift (Figure 2) and slid by gravity eastward 20 to 30 miles to their present position. The Canadian and northwestern Montana Rockies consist of a similar succession of thrust sheets, and the base of this enormous assembly is visible at the surface for a number of miles. Its width and near flatness attest to the long-distance transport of the thrust sheets and strengthen the theory that gravity caused the sheets to slide at a low angle down the east slope of a west-lying uplift. The Lewis Thrust along the east front of Glacier National Park is the best known of the thrust sheets. One has only to view these sheets on the ground to become aware of their vast magnitude. They are several thousand feet thick, probably ten to 20 miles wide, and 50 to 150 miles long and can be visualized as gigantic landslides moving down a slope of only 2° to 5°. The Charleston–Strawberry–Neb Thrust (see Figure 2) is interesting inasmuch as a sequence of rocks from its geosyncline, 25,000 feet in depth, has been moved eastward against, and probably over, a rock sequence of the shelf, itself more than 3,000 feet thick. There is no doubt as to the size and importance of these huge sheets of rock in the formation of the Rockies. In Canada and Montana, as in southeastern Idaho and western Wyoming, their mechanical aspects can be fairly well analyzed, but in the Great Basin of Utah and southern Nevada their sizes, extent, and connections remain controversial.

One school of thought suggests that the thrust sheets arose from an uplift in eastern Nevada and that the sheets then rode eastward 120 miles to central Utah. Another hypothesis posits an uplift like the Cache only 25 or so miles to the west and has the sheets coming from this nearby source. Still a third hypothesis submits that stresses occasioned by the weight of sea-floor deposits, and associated earth movements, at and under the continental margin of the Pacific 500 miles to the west, may have been transmitted to the western division of the Rocky Mountains thrusting the sheets eastward.

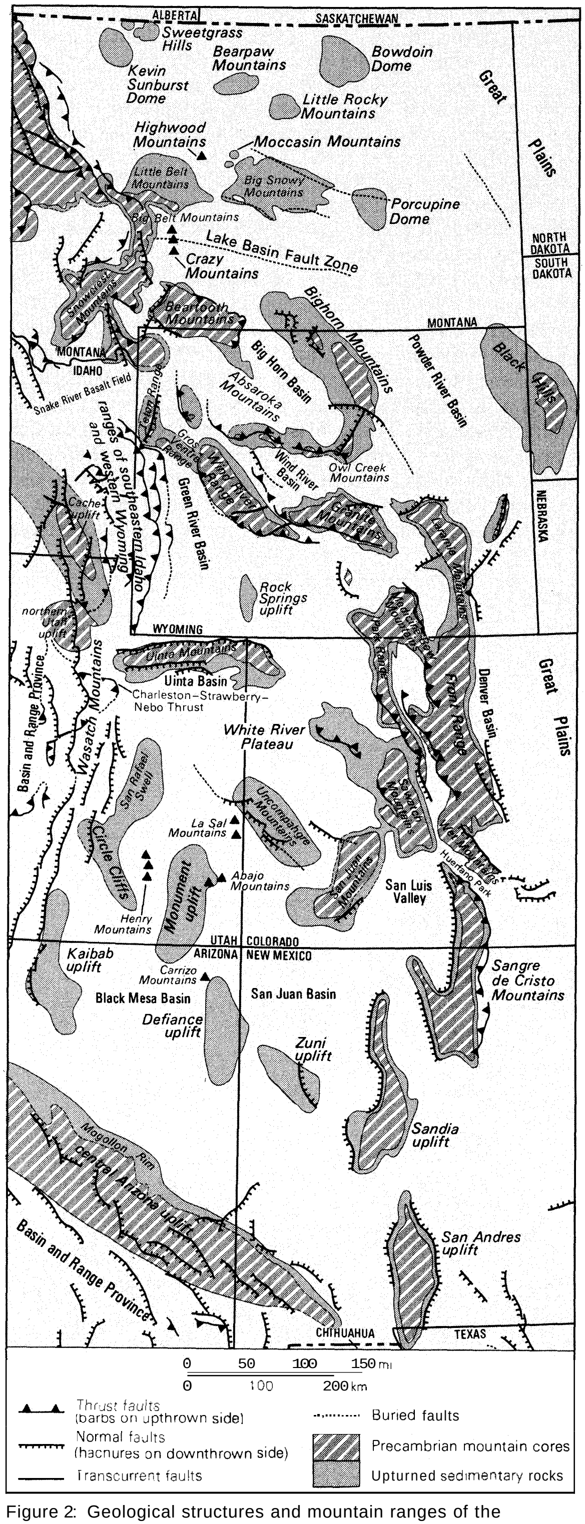


Figure 2: Geological structures and mountain ranges of the eastern division of the Rocky Mountains within the U.S.

Structures of the shelf. The uplifts of the Eastern, or shelf, division of the Rocky Mountains may conveniently be considered state by state, although state boundaries are, of course, in no sense geologically significant.

In Montana a belt of uplifts extends eastward through the centre of the state and is composed principally of the Little Belt and Big Snowy Mountains. Figure 2 indicates the oldest rocks exposed, which are the cores of the

Ages of the rocks in the shelf area

uplifts. The oldest rocks are those of Precambrian age, which have been dated by radioisotope samples from various locations as being from 1,100,000,000 to 2,700,000,000 years old. Cambrian time began about 570,000,000 years ago, and all Precambrian rocks lie below the Cambrian strata. As one of these great uplifts (known technically as a dome or anticline) is elevated, the highest part is attacked by erosion first, and thus, as the strata, or beds of rock, are eroded away by wind, rain, and ice, the oldest rock in the core ultimately is exposed. The Little Belt Mountains are asymmetrical, that is, with the core of Precambrian rocks exposed near the southern margin rather than at the centre. The Big Belt Mountains uplift extends southward from the Little Belt and must be partly involved in the multiple thrust faulting of the geosyncline to the west. The Porcupine Dome is a gentle but broad uplift along the easterly trend.

North of the east–west belt of uplifts are a number of scattered uplifts, the Bearpaw and Little Rocky supporting the highest peaks. In the Bearpaw Mountains is Baldy Mountain at 6,956 feet. The Sweetgrass, Kevin–Sunburst, and Bowdoin uplifts are gentle domes. There are a number of volcanic fields in Montana, but the Highwood and Crazy mountains are the areas that stand conspicuously apart by virtue of their Alpine relief. In this area, the dike swarms — networks of rock figures filled with solidified lava — are classical examples of the type.

In Wyoming the Big Horn Basin is surrounded by the Bighorn Mountains, the Owl Creek Mountains, the volcanic Absaroka Mountains, and the Beartooth Mountains. The Laramie, Sweetwater, and Wind River uplifts form a barrier across central Wyoming. The Sweetwater uplift was originally as high and imposing as the Wind River. After much erosion it foundered until almost all the peaks were covered with sediments, some of volcanic origin. In the current cycle of erosion these ancient peaks are now being exhumed.

The Black Hills uplift is large but was not raised to the same height as some of its neighbours. The Precambrian granites are, nevertheless, exposed, and it was from these that the gigantic Rushmore Memorial was carved. The newest and most rewarding oil field area of the Rockies is located in the Powder River Basin, close to the northwest extremity of the Black Hills.

Special features of the Uintas

The Uinta Mountains uplift in northeastern Utah and northwestern Colorado follows an easterly bend and is structurally linked to the White River Plateau, which is a projection of the Sawatch uplift. The Uintas are interesting for several reasons. They seem to have inherited their position from a Late Precambrian basin that extended eastward from the main north–south basin. The broad anticlinal, or upward, bending nature of the rock strata of the Uintas gives strong support to the theory that they are a vast blister-like upwarp. The upwarp is also bordered by upthrust faults, which accentuate the anticlinal structure and confirm that the uplift was caused by vertical, upward-directed forces. It may be contended that the Uintas are typical of the mode of formation of all the uplifts of the eastern division of the Rockies, although it must be noted that some uplifts are asymmetrical with a border upthrust fault only on one side, as in the case of the Wind River mentioned above.

The large Front Range uplift is terminated on the south by Huerfano Park, with the Wet Mountains making up an arm linked to the north end of the Sangre de Cristo uplift. The Front Range passes into the Laramie uplift on the north by way of narrowing and a lower relief area. On the west it is crowded against the Sawatch uplift, and in this zone the sediments are compressed together, with marked folding and thrust faulting.

Part of the San Juan uplift and part of the San Luis Valley, or depression region, are covered by a large volcanic field. Some of the highest peaks in the San Juans are volcanic rocks. These include Windom Mountain 14,166 feet (4,293 metres) and Summit Peak 13,377 feet (4,054 metres).

In New Mexico the most significant geological aspect is that a post-Rocky Mountain rift, or fault-bounded, valley system developed in Late Cenozoic time (about 30,000,000 years ago), proceeding in a northerly direction through the central part of the state. This rift valley complex now conducts the Rio Grande (*q.v.*) down to west Texas and Mexico. A major block of the valley flow, dropped down between the great faults on either side, has, in fact, cut the San Andres uplift into two opposite facing escarpments, or rock outcrops, the Sacramento Mountains on the east and the San Andres Mountains on the west. The Jornado del Muerto Valley west of the San Andres is another downfaulted block. Farther north the hypothetically reconstructed Sandia uplift has been shortened by the earth movements associated with the rifting process. The rifting is regarded as a north-extending arm of the Basin and Range Province (*q.v.*).

Rift valley of New Mexico

The physiographic province called the Colorado Plateau in southeastern Utah, western Colorado, northern Arizona, and northwestern New Mexico is also basically part of the Rocky Mountains. The several uplifts shown on the map in the Colorado Plateau are comparable in size (but not in height) to the other uplifts of the eastern division. They have not been domed up as much as in the uplifts to the north, and consequently less erosion has occurred, for in none of them are Precambrian rocks exposed. The beauties of the wilderness of the Colorado Plateau area had, by the 1970s. become more accessible to tourists. The Canyonlands National Park over part of the Monument uplift was being opened up by roads, and a new interstate highway had been completed across the San Rafael Swell.

Volcanic activity in the south

The Grand Canyon of the Colorado (*q.v.*) cuts across the southern end of the Kaibab uplift in this region. In addition to the uplifts, four mountain groups — the La Sal, the Henry, the Abajo, and Carrizo mountains — are notable. These are noted examples of laccolith-type mountains, in which, from a central pipelike intrusion reaching deep down into the earth's crust, molten magma has been injected between the layers of sedimentary rocks, causing the overlying beds to bulge up in small domes about one mile across. The domes are called laccoliths, and each mountain group is made up of a cluster or group of laccoliths.

In summarizing overall trends in the complex formation of the Rocky Mountain uplifts, it may be postulated that, in those uplifts in which erosion has exposed Precambrian rocks in the core, border upthrusts have developed on one side or both. In other words, when the uplift, from which mass the present-day mountains have been formed, has developed to a height greater than 20,000 feet above the adjacent basin, a border upthrust has also developed. Authoritative opinion would now suggest that the border thrust dips into the uplifted mass, steepens in depth, and eventually becomes vertical. This aspect of the uplifts results in a further postulate, that the force and mechanism responsible for each uplift is a blister-like magma injection rather deep in the Precambrian crust. The magma is supposedly of basaltic composition and comes from the upper mantle of the Earth's molten interior.

Relation to Sierra Nevada. The Basin and Range Province extends from central Utah to the Sierra Nevada, and although the geological relations in this wide region are not unknown, about three-fourths of the bedrock geology is covered and concealed by volcanic fields and alluvium (erosion waste material). In consequence, there is scope for various interpretations.

It is well established that a belt of deformation and mountain building extended northward through central Nevada. It evolved from Late Devonian time to Permian time (about 345,000,000 to 280,000,000 years ago) and thus preceded the growth of the Sierra Nevada and Rocky Mountains. In the Sierra Nevada (*q.v.*) the earliest intrusions occurred in Triassic time (some 225,000,000 years ago) and seem to have been a continuation of the earlier central Nevada mountain-building activity.

Commencing in Jurassic time (190,000,000 years ago), crustal unrest occurred to the east of the central Nevada belt of deformation. There is controversy as to its nature and extent. It appears that eastern Nevada and western Utah became land, mostly mountainous, in Jurassic time, and this region probably attained its highest elevations in the mid-Cretaceous period (100,000,000 years ago). This was the time of folding and thrusting in west central Utah and in western Wyoming, when, as has been already noted, the western segment of the Rocky Mountains was formed.

Relation to Basin and Range Province. Since part of the western division of the Rocky Mountains now has the Basin and Range faulting superimposed on it and thus has had its bedrock geology partly obscured, the western limit of the Rocky Mountain system is not readily defined or clear. Some aspects of eastern Nevada seem to suggest a different structure from that in western and central Utah, and thus the western limit may lie just east of the Nevada line.

Igneous rocks. Large volcanic fields occur in many parts of the Rocky Mountains. Some of these have already been mentioned, as have the laccolithic mountains of the Colorado Plateau, a further manifestation of volcanic activity. The chemical nature of most of the volcanic rocks is such that silica is abundant in them. Another rock type, in contrast, has abundant iron and magnesium and a lesser amount of silica and potassium. This would seem to indicate that the molten magmas of the high-silica volcanic rocks originated, at least in part, in the melting of the Precambrian rocks of the crust, rather than from basalt magma originating in the mantle of the Earth's interior.

Mineral origins

The western division of the Rockies, especially where the Basin and Range faulting occurs, is noted for numerous crosscutting intrusions of solidified magma known as stocks. These are generally one to five miles across where exposed and, like the volcanics are silicic. The magma of the stocks likewise appears to have been mobilized from the silicic rocks of the Precambrian crust. Virtually every stock is the focus of numerous mineral deposits, both large and small. Hot solutions were given off by the crystallizing magma in the stocks, and these carried metal compounds, such as the sulfides of copper, lead, zinc, silver, and iron. These sulfides commonly contained a little gold. The hot solutions worked upward through the existing pore spaces in the rocks, and as the temperatures dropped, various metaliferous sulfide minerals were precipitated. Where the intrusive stocks have penetrated limestones, chemical reactions may have occurred between the solutions and the limestones, and in places massive sulfide deposits have resulted. Mining districts such as Bingham, Park City, and Ely have sprung up as a result of the mineralization in and around such stocks.

ENVIRONMENTAL AND ECONOMIC RESOURCES

Water supply. Water for irrigation, industry, and culinary needs is generally in short supply in the Rocky Mountains. To the south, as the climate becomes drier, the water supply factor becomes more critical. New Mexico and Arizona are the states most affected. The U.S. Bureau of Reclamation is the federal agency that has been most concerned in the construction of water-storage reservoirs all over the western United States, and certainly the Rocky Mountains has its share of these projects. Not many more favourable unused dam sites remain, so that within a short time—certainly not more than 50 years—all the annual precipitation in the Rocky Mountains will have been stored and used. It may then be necessary to import water from the Columbia River (*q.v.*) and from western Canada. This process would be fraught with numerous technical, financial, and political difficulties.

National parks, forests, and recreational areas. Many of the nations finest national parks, national monuments, and wilderness areas are in the ranges of the Rocky Mountains and in the Colorado Plateau. To these areas of natural beauty have been added such large recreation facilities as the Glen Canyon National Recreation Area, located on either side of Lake Powell in Utah and Arizona. The parks, monuments, and recreational areas have been withdrawn from mining, oil, and gas drilling and in general from stock grazing and are under federal control and carefully regulated to maintain the natural conditions.

Most of the ranges and mountain groups have been designated as U.S. national forests. As such the principle of multiple use obtains, with lumbering, mining, oil and gas drilling, and grazing permitted under well-regulated federal laws. In the marginal and, as yet, unclaimed basin and arid lands, the federal government still has control of the public domain. In effect, the Rocky Mountain states control less than half of the land; more than half is controlled and regulated by the federal government. Serious problems have arisen as a result of grazing, mining, and oil exploration, although by the start of the 1970s the federal government, with the cooperation of the several states, was attempting to preserve the natural heritage of the Rockies from misuse and pollution. By this time the disturbance of primitive habitats in some areas of the Rockies had generated concern among experts and laymen alike, although the floral and faunal associations had been hurt badly only in limited sections.

Pollution and its control

Nymph Lake in Rocky Mountain National Park.
By courtesy of Union Pacific Railroad

Metallic and nonmetallic resources. Copper is easily the most valuable of the many metallic resources of the Rocky Mountains. Great mines in Montana, Utah, and Arizona produce nearly all of the nation's red metal. Iron ore in Wyoming and Utah support a steelmaking industry. Perhaps the Rockies, however, are most noted for many underground mines for silver, gold, lead, and zinc. Such mines occur in Colorado, Montana, Idaho, Nevada, Utah, New Mexico, and Arizona. The Rockies produce all of the nation's molybdenum and nearly all of the beryllium and uranium.

Great reserves of nonmetallic substances occur in various places in the Rocky Mountains. These are rock phosphate, potash, trona, magnesium and lithium salts, glaubers salt, gypsum, limestone, and dolomite. If necessary, the nation's needs could be supplied for a long time with the stoles already known in the Rockies.

Oil and gas production. The large basins between the uplifts of the Rocky Mountains contain many oil and gas fields. Wyoming, New Mexico, Montana, Colorado, and Utah are all substantial producers, with the Powder River Basin proving one of the leading regions. The western division of the Rockies has yielded very little oil, perhaps because of the extensive folding and faulting.

The Rockies also hold extensive shale deposits containing a solid hydrocarbon material that can be driven off as oil by heat treatment. They occur principally around the Uinta Mountains in Wyoming, Colorado, and Utah. The amounts of potential oil are vast, and, as viewed by certain economists, if the exploitation of the oil shale is not carefully planned and regulated, a national calamity could result. As yet, the extraction of the oil has not been made profitable.

Immobile oil is located in certain sandstones in various places. These deposits are called bituminous, oil, or tar sands. In amounts, they compare to giant oil fields. Like the oil shales, the bituminous sandstones have not yet yielded their oil economically except in Alberta, where a major enterprise costing $300,000,000 remains marginal.

Coal. The Rocky Mountains and the adjacent Great Plains on the east contain the Western Hemisphere's most abundant and usable coal reserves. These are bituminous, subbituminous, and lignitic in character. Although not readily usable for metallurgical purposes, they constitute a tremendous energy source, and it is predicted that in the near future these coals will be used much more extensively for electrical-power generation than at present.

BIBLIOGRAPHY. A.J. EARDLEY, Structural Geology of North America, 2nd ed., ch. 19–27 (1962), a discussion of the divisions of the Rocky Mountains and a summary of the geological characteristics of each, and "Major Structures of Colorado and Utah," UMR Journal, series 1, no. 1, pp. 79–99 (1968), an article presenting the theory that most mountain ranges are blister-like uplifts and not due to horizontal compression of the crust; R.J. ROBERTS, "Tectonic Framework of the Great Basin," ibid., pp. 101–119, an article presenting a review and theory of evolution of the western geosyncline that is concerned in part with the western division of the Rocky Mountains; R.L. ARMSTRONG, "Sevier Orogenic Belt in Nevada and Utah," Bull. *Geol. Soc. Am.*, 79:429–458 (1968), an article similar in theory to Roberts with a map connecting the exposed thrust faults. These three articles represent the most recent thinking of the evolution of the western division of the Rocky Mountains.

More general descriptive works include: H.R. BELYEA, The Story of the *Mountains* in Banfl National Park (1960); H.E. GREGORY, Geological and Geographic Sketches of Zion and Bryce Canyon National Parks (1956); H.S. ZIM, The Rocky Mountains (1964); D.S. LAVENDER, The Rockies (1968); and R.M. PEARL, Seven Keys to the Rocky Mountains (1968).

(A.J.E.)

Rodentia

The rodents, or Rodentia, are the most abundant order of mammals. At present, over a quarter of the families, 35 percent of the genera, and 50 percent of the species of living mammals are rodents. Probably an even higher percentage of individuals are rodents, for they tend to be small animals with dense populations. They are one of the few groups of animals that flourish in close association with men. Some, such as squirrels, live independently but fairly successfully near humans. Others, such as the house mouse (Mus musculus) and black and Norway rats (Rattus *rattus* and R. norvegicus), have adapted themselves to human civilization, and live everywhere that man does. These two rats (and the Polynesian rat, Rattus *exulans*, of Australia and Oceania) have travelled in ships and boats of all sizes, and have populated the entire habitable world, especially near human habitations.

For entries on particular rodent families and species, see also RELATED ENTRIES under RODENTIA in the Ready ***Ref***erence *and* Index.

GENERAL FEATURES

All rodents possess one pair of upper and one of lower incisors, growing throughout life, with the enamel restricted to a band on the front side of the teeth. Behind this is a large gap (diastema) followed by two to five cheek teeth. The jaw articulation is so arranged that when the cheek teeth are in use, the incisors do not meet, and vice versa. The incisors grow continuously, and must be worn off equally fast, or the whole gnawing mechanism is ruined. Because of the necessity to abrade these incisors, rodents spend a considerable amount of time gnawing hard objects.

Generally rodents are small. Some mice and dormice are among the smallest of living mammals, adults being as small as 75 millimetres (three inches) long, including the tail, and weighing as little as 20 grams (0.7 ounce). The largest living rodent is the South American capybara (Hydrochoerus hydrochoeris), reaching over 1.3 metres (four feet) in length and as much as 50 kilograms (about 110 pounds) in weight. A fossil rodent recently described from Uruguay is reported to have had a skull as large as that of a bull and a body bulk as large as that of a wild boar.

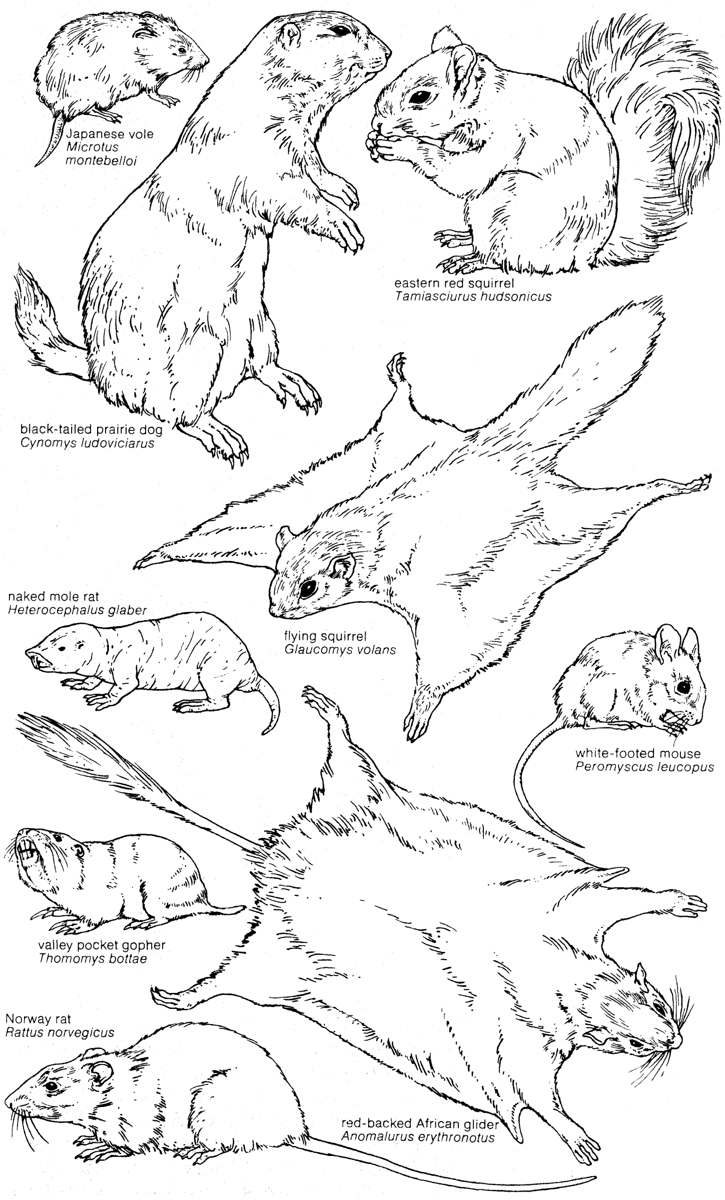


Figure 1: Range of body-plan variation of Rodentia.

Rodents are of major economic importance, primarily as consumers of the grains that are the basic foodstuff for man. It has been estimated that rats and mice destroy up to one-third of grain crops under conditions of heavy infestation. Burrowing rodents may damage root crops. The muskrat (Ondatra *zibethica*) and nutria (Myocastor coypus), introduced into Europe as fur sources, have escaped and spread over much of Europe between the Baltic and the Alps. Their burrows, particularly in canal banks, have been a major source of damage to the drainage system, most especially in The Netherlands. A number of rodents serve as reservoirs for human diseases, such as bubonic plague, tularemia, scrub typhus, and others. The plague that ravaged Europe during the mid-14th century was transmitted by fleas from rats to humans.

Several rodents (beaver, muskrat, chinchilla, nutria, squirrel) produce fur useful to man. All but beaver and squirrel have been domesticated for this purpose. Albino mice and rats, hamsters, and guinea pigs are widely used as laboratory animals for biological and medical pur-

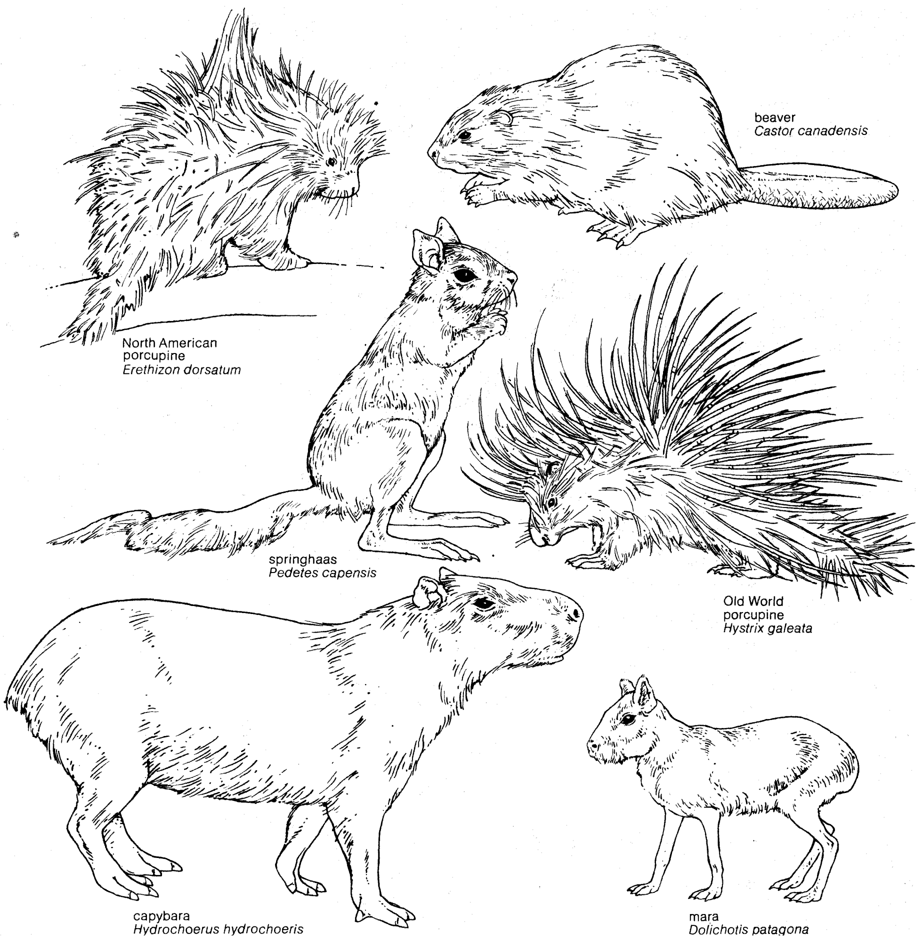


Figure 2: Range of body-plan variation of larger Rodentia.

poses. Guinea pigs were domesticated by the Incas for food; a few kinds of rodent have been raised as pets.

Distribution of rodents

Rodents occur naturally in all parts of the land where there is an adequate food supply and are found in essentially all terrestrial habitats. They range from well above the Arctic Circle to the southern tips of Africa and South America, and were the only terrestrial placental mammals, other than bats, to reach Australia before the arrival of man. Many rodents have successfully adapted to difficult environments such as deserts. Many rodents have broad climatic tolerances, an example being the North American porcupine, which is found from the Arctic Circle to central Mexico and from the Atlantic to the Pacific. Most are quadrupedal scamperers, but they generally have much freedom of use of their forefeet in manipulating food; many are burrowers, spending most of their life underground; some are ricochetal, leaping on their hind legs; flying squirrels use skin membranes to glide from one tree to another; a few (beaver, muskrat, water vole, nutria) have become amphibious in habits, living in freshwater streams and ponds; and a number of South American rodents are cursorial (running) animals.

IMPORTANCE TO MAN

Destruction **of crops and** foodstuffs. The most important rodents, from the point of view of economic damage, are the Norway and black rats, with the house mouse close behind them. It has been frequently estimated that the rat population of the United States is approximately equal to the human population. **A** population of over 1,000 rats per acre on an Iowa farm has been reported. Population explosions of house mice occurred in the Central Valley of California in 1926–27 and 1941–42. During the former, the mouse population was estimated to have reached over 80,000 per acre.

Remarkable population explosions of voles ***(Microtus)*** and wood mice ***(Apodeinus)*** are well known in western Europe, recurring every few years. In France, from 1790 to 1935, there were at least 20 mouse plagues, some lasting several years. Estimates of abundance are highly inaccurate, but there are reports of 8,000 voles per acre and 15 to 20 vole burrows per square metre, in peak conditions. There was a population explosion of voles and wood mice over much of Germany beginning in late summer of 1917 and lasting through 1918. Damage to crops was serious; clover was the favourite food; winter rye and wheat, sugar beets, and potatoes also were destroyed in many areas. During the winter the voles invaded barns and farm buildings, destroying all kinds of stored food. In 1932–35 voles and lemmings were in epidemic proportions in over 40,000 square miles of the southern U.S.S.R. This plague resulted in extensive damage to wheat and other field crops and to orchards.

Rats will eat almost anything that humans eat. Perhaps the most serious damage is to the seeds of grain plants, both before and after harvesting. Grain stored on farms is often not only eaten by rats but also rendered unsuitable for human consumption by being mixed with rat droppings. Food that has reached warehouses in cities is also eaten by rats, and here the excess of damage over the amount actually consumed is even greater. It is estimated that rats damage about twice as much grain as they eat. About 50 pounds of grain are required to support a rat for a year, so the total cost is about 150 pounds per rat per year. Rats sometimes demonstrate a fondness for animal food and have been known to kill several hundred baby chickens in a single night. Eggs are frequently eaten, and even full grown hens, baby pigs, and lambs may be killed. Rats will also follow a farmer in the planting season and dig up newly planted seeds.

Depredation by rats

Rats also do extensive damage in searching for food or in making nests. They gnaw holes to gain entrance to barns, warehouses, or houses, or through walls once they are inside. Clothing, upholstered furniture, and other textiles are gnawed to provide nest-building materials. In areas of cities where extensive quantities of garbage and other refuse are available, the rat populations may exceed the human populations, and rat infestation is one of the commonest complaints of slum dwellers. Such rats are very apt to bite sleeping humans, especially children, and fatal attacks on babies have been known to occur.

The house mouse has food preferences similar to those of the Norway rat but is not as abundant nor as large. Other rodents are much less likely to attack stored supplies of human food, but they do eat considerable quantities before harvesting. This is true of such forms as voles, field mice, and squirrels. The seeds of both wild and domesticated grasses are a major food source for large numbers of wild rodents. Because of the quantities of these that are eaten, such animals are often considered a major threat by grazing interests, but the benefit that rodents give by collecting seeds into underground storehouses, where they may later sprout and grow into new plants, certainly comes close to balancing any damage they cause to man's interest.

Transmission of diseases. Plague. Rodents serve as reservoirs for a number of diseases that may be transmitted to humans by arthropod agents. The most devastating of these is bubonic plague. This disease is fundamentally a disease of rodents, especially rats, transmitted from one rodent to another by an intermediate host, the flea, which also serves to transmit the disease to humans. An epidemic (called a "pandemic" because of the totality of infection in the human population) that seems to have been plague spread over Europe in the 6th century AD. If this epidemic was indeed plague, it must have involved an unusual rodent host because it antedated the arrival of Rattus in Europe. The epidemic known as the Black Death originated in Mesopotamia about the middle of the 11th century, and spread to Europe, particularly in the 14th century, being accompanied by the spread of rats throughout the Continent. It has been estimated that 25,000,000 people died of that pandemic of the plague in Europe. The latest pandemic originated in southwestern China in the late 19th century and was spread all over the world by the rat populations of oceangoing ships. Plague is controlled by controlling rat populations, particularly those on ships, and preventing shipborne rats from reaching land. Although plague epidemics have been brought under control, there still remain numerous foci of infection.

The role of rats in the "Black Death"

The plague bacillus (Pasteurella *pestis*) can infest a variety of other rodents, and foci have been established among native rodents other than Rattus in many parts of the world. This disease is referred to as "sylvatic plague" to distinguish it from the basically urban occurrence of ratborne plague. Over 80 species of ground-living and burrowing rodents are known to be involved in sylvatic plague, including ground squirrels, some cricetids (*e.g.*, voles, lemmings, muskrats), several murids (*e.g.*, Old World mice), and a few others, including guinea pigs.

Tularemia. Tularemia is primarily a disease of lagomorphs (rabbits and hares) and secondarily of rodents. It has been reported from all parts of the United States, and seems to have spread from there to many other parts of the world. Among rodents, it is carried by ground squirrels, tree squirrels, prairie dogs, chipmunks, muskrats, and beavers, and is transmitted largely by ticks.

Rickettsia1 diseases. Three rickettsia1 diseases—murine typhus, Rocky Mountain spotted fever, and tsutsugamushi disease—involve rodents as reservoirs. Murine typhus, which is much less severe than epidemic typhus, is transmitted to man by fleas, primarily from Rattus. It is probably almost universal in tropical and subtropical areas. Rocky Mountain spotted fever, which apparently originated in the northwestern United States, has spread over much of that country, and very similar if not identical diseases are found in Mexico, South America, and Africa. It is transmitted to humans by tick bites, and is presumably endemic in a considerable number of rodents as well as other animals. The cottontail rabbit is believed to be a major reservoir animal. Tsutsugamushi disease, or scrub typhus, is transmitted to humans by the bite of a rat mite. It is found in the East Indies and Southeast Asia. The reservoir seems to be primarily rodents of the genus Rattus. A variety of other diseases of lesser importance are transmitted to humans from a reservoir in rodents, largely Rattus, by rat bites, ticks, or in other ways.

Damage by burrowing and gnawing. Rodents have been accused, in many parts of the world, of damage to agricultural interests because of their burrows. Cattlemen in the western United States supported campaigns to poison prairie dogs (*Cynomys*) because horses or cattle might (and occasionally did) break their legs in prairie-dog holes. But perhaps the most striking case of damage from rodent burrows has been caused by the American muskrat, introduced into European fur farms, from which it escaped as early as 1905, spreading widely over much of Europe, from Great Britain and Brittany to the Ukraine. It burrows in the banks of streams, and has caused damage especially by making burrows in the banks of drainage ditches, canals, and in dikes.

The ever-growing incisors of rodents need to be used regularly on hard substances to wear them off. This need results in the gnawing of lead pipes or telephone transmission cables, or the making of holes in boxes, walls, and stored inedible items.

Rodents as valuable furbearers

Benefits derived from rodents. Trapping for furs. The beaver (Castor canadensis) was extremely important in the development of the western United States and Canada, the trappers of the first two-thirds of the 19th century being primarily interested in beaver skins. Most of the initial exploration of the West was performed by beaver trappers. The beaver skin became the basic unit of currency over much of the western United States. The trapping was so extensive and effective that beavers were exterminated over much of the area, but they are making a slow recovery under current conditions of protection. The fur of the chinchilla (*Chinchilla* laniger) was of such value that these rodents were almost exterminated in Argentina. Muskrat and nutria have also been extensively hunted for their fur.

Rats in the laboratory. Among the best laboratory mammals, for all types of biological, medical, and psychological investigation and for testing of new drugs, are the laboratory rats and mice, normally albino strains of wild species of the Norway rat and the house mouse. Guinea pigs (Cavia cobaya) are also widely used. There are a number of other laboratory rodents of lesser importance, of which the Syrian or golden hamster (*Mesocricetus* auratus) is the most widely used. Some strains of rats are nearly unique among nonhuman animals in being susceptible to dental caries, and most of the experimental work on tooth decay has been performed on them.

ECOLOGY AND NICHE RELATIONSHIPS

The basic scampering rodent

Habitat and locomotion. The most typical members of the order are the small, ground-living rodents, of the type generally called rats or mice, similar to what was probably the ancestral condition for the order. While these generally stay on the ground, most can climb shrubs and bushes with ease and many climb trees; among the smallest rodents, harvest mice climb stalks of wheat to reach the seeds on which they feed. This central generalized adaptive type of rodent, the scamperer with limited burrowing or arboreal adaptations, includes the family Muridae (Old World rats and mice, now worldwide), the family Cricetidae (field mice, wood rats, voles, lemmings, muskrats, to name a few; found in most parts of the world other than Australia), some Sciuridae (the ground squirrels and chipmunks), the nonleaping members of the New World family Heteromyidae (pocket mice); some South American rodents (degus, spiny rats, and chinchilla rats), and the South African rock rats (Petromyidae). Although there are many adaptive variants among the scampering rodents, there is a strong tendency for them to inhabit either open country or limited areas of woodland. Rodents of this type are found over essentially the entire land surface of the world. In most cases individual or family territories are set up, and intruders are successfully kept at bay by a variety of threatening and defiant postures. Normally there is no actual combat. The territories are, of course, of varying size, related to the available food supplies and the size of the animals. Individual ground squirrels have been observed over a territory of 2,500 square metres (about 0.6 acre).

Changes with time in the territories of chipmunks (*Eutamias*) indicate rather striking seasonal changes, with an increase in the area toward the end of the foraging season (perhaps to allow greater accumulation of food), and very little shift of the territory from one year to another. With few exceptions, these terrestrial rodents make burrows near the centre of the territory where they may spend as much as half of the time, where the young are born, and where the adults hibernate, in forms where hibernation occurs. Normally more young are produced than there is room for and as a result there is extensive migration of subadult or young adult individuals. For most rodents, the lengths of these migrations are unknown, but it has been shown that muskrats will migrate as much as 20 miles (over 30 kilometers), including several miles cross-country, over a period of one to two weeks, looking for suitable new stream or swamp sites.

From occasional tree climbing it is a relatively short evolutionary step to an arboreal habitat, such as that occupied by the tree squirrels, or, with an increase in size and decrease in rapidity of movement, by the New World porcupines. Tree squirrels are, essentially, scamperers that have developed the capability of holding onto bark with their claws. They are as much at home on the ground as in trees, and show no special anatomical modifications for tree life. Their nests may be either in hollow trees or other sheltered places (*e.g.*, barns) or built of twigs and leaves among the smaller branches of trees. In Europe and adjacent parts of North Africa and Asia, the dormice (Gliridae) have similar habitat but spend an even larger part of their lives in the trees. Many murids and cricetids (rats and mice, in the broad sense) are about as fully arboreal in their habits as are the tree squirrels.

Gliding rodents

Scampering arboreal animals are faced with the necessity, from time to time, of moving rapidly from one tree to another, and do so by leaping. It is generally agreed that gliding forms have evolved their skin fold (patagium), which assists them in gliding, as a direct adaptation to this type of habitat. Patagia have evolved, probably several times independently, among the sciuridae (squirrels) of both the Old World (*Pteromys* and others) and North America (*Glaucomys*) and in the quite distinct and unrelated African family of scaly-tailed squirrels (Anomaluridae). The patagium of the anomalurids is supported at the anterior end by a long, slender cartilage, which has apparently evolved from the olecranon process of the ulna (the posterior bone of the forearm) at the elbow.

Larger, heavier arboreal forms such as the New World porcupines move slowly along the trunks and main branches of trees. One of these, in South and Central America, has developed a prehensile (grasping) tail that assists in climbing.

A large proportion of the small, ground-living rodents construct underground nests, where the young are born and reared. Often there are adjacent chambers for food storage. In many cases two or more exits provide an opportunity for escape from predators. From ancestors such as this, many lines of rodents have evolved into habitual burrowers that make extensive subterranean tunnels and rarely or never come to the surface, foraging, feeding, mating, and raising their families underground. Among the numerous varieties of burrowing rodents are the prairie dogs, colonies of which were formerly abundant in the western United States. The mouth of the burrow is surrounded by a volcano-shaped mound of dirt, on which a prairie dog frequently sits upright, alert for potential enemies. The animal forages on the ground surface, returning to the mound or to the burrow to eat. Voles (Microtus) may make lengthy runways, two or three inches below the surface, arching up the ground surface over them. The pocket gophers (*Geomys* and Thomomys) are even more effective burrowers, throwing up mounds a foot or more across and spending most of their life beneath the ground. They use the incisors as well as the front legs in digging. During times of heavy snow, they may burrow through snow, feeding on plants above ground. Similar types of burrowing occur in the Old World mole rats (Spalacidae). An extreme form of burrowing is found in the African mole rats or blesmols (Bathyergidae), in which the incisors extend forward, with a fold of skin closing the mouth behind them, permitting their use as the primary digging tools, having replaced the forelimbs in this respect. The burrows of a single blesmol form a crisscross of tunnels, at several levels below ground surface, that may occupy an area a thousand or more square metres. All of these burrowing forms have short, heavy limbs used in digging, short tails, and small eyes. In one genus of blesmol, Heliophobius, the eyelids have grown together so that the eyes do not form images. These animals remain permanently below ground. Many burrowing rodents are solitary, but a number of them live in colonies, unlike other members of the order.

Giant rodents

Relatively few rodents have become large animals. Several of the South American caviomorphs have done so, however, and have evolved as quadrupedal terrestrial cursorial animals. Usually these have the claws modified into or toward hooves, reducing the number of toes on each foot to four or occasionally three, and tending to make each foot symmetrical about the median plane. Among these are the capybara (the largest living rodent), nutria, paca (*Cumiculus* paca), and agouti. The guinea pig is a smaller relative with the same adaptations. Such animals normally escape predators by running away, although both capybara and nutria will take to the water. The Old World porcupines are slow-moving terrestrial quadrupeds that are protected, as are the New World porcupines, by the quills, which are elongate, pointed hairs.

Many rodents have become bipedal jumpers, escaping their enemies with long leaps. These forms include the kangaroo rats (*Dipodomys* and Microdipodops) of western North America, the jumping mice (*Zapus* and *Na*paeozapus) of eastern North America and eastern Asia, the jerboas (Dipodidae) and gerbils (Gerbillus) of North Africa and Southeast Asia, the saltatorial dormouse (*Selevinia*) of Kazakstan, U.S.S.R., the chinchilla of the southern Andes, and the springhaas or Cape jumping hare (Pedetes) of South Africa. All of these are small, except Pedetes, which stands 30 centimetres (12 inches) tall. Most of the saltatorial rodents have very highly enlarged tympanic bullae—the bony covering of the middle ear—which is not true of other rodents. It is probable that this is an adaptation to life in a desert, where an amplification of the noise made by approaching predators may be of great importance. Certainly the large middle-ear cavity greatly increases the amplification of sounds. The fact that the bullae are particularly large in leaping forms suggests further that there may be some relationship between the size of the bulla and the balancing and stabilizing activities of the ears.

Although a number of rodents are at least partially aquatic, special adaptations are rare, being found only in the beavers. Beavers have large hind legs, with strong webs between the toes and an enlarged first digit, as long as the other toes, making very efficient paddles. Swimming is carried out primarily by the feet, but the broad flat tail may be used on occasion for sculling. A number of other rodents that are semiaquatic have webbed feet and soft, dense underfur but few other aquatic adaptations. These semiaquatic forms include muskrats, water voles, capybaras, nutrias, South American fish-eating rats, and the water rats of Australia and New Guinea.

Vegetable foods

Food habits. As a whole rodents are herbivorous, although most of them will eat animal food on occasion. The diet includes a wide variety of plant foods, although seeds are the favourite item. Many are particularly fond of the seeds of the various grasses, both wild and domestic. These may be eaten on the spot, but large quantities are usually carried home in the cheek pouches that are common among rodents. Seeds carried home may be eaten in the safety of the burrow or stored for later use. Such storage is an important item to the plants concerned, as it places the seeds in a favourable location for sprouting, if the rodent either forgets about them or himself becomes food for a predator. Many rodents, such as squirrels, are fond of nuts and can open even the hardest black walnuts to get at the seed within. Acorns, seeds, fruits, berries, young leaves, buds and shoots, green leaves, tubers, and bulbs are eaten by a wide variety of rodents. Burrowing rodents tend to concentrate on the roots, tubers, and bulbs that they encounter in making their tunnels, although some burrowers forage on the surface as well. The larger caviomorphs, such as the capybara, paca, mara (Dolichotis *patagona*), and agouti (Dasyprocta aguti), are essentially grazing animals, eating all kinds of green vegetation, young stems, and fruits. Marmots (Maimota) have rather similar food habits. Old World porcupines are essentially omnivorous, including carrion in their food. New World porcupines eat leaves, twigs, buds, and bark. Beavers cut aspens, willows, and poplars, using leaves, buds, and bark for food; they also eat a variety of aquatic plants. Many desert-living rodents have a diet that, in the dry season, may be almost exclusively seeds, from which they are able to obtain enough water for survival, using the water produced by metabolism. African cane rats eat coarse grasses and shrubs, one of their favourite foods being sugarcane. A number of rodents, especially during hard winters, strip the bark from trees. Because of their large numbers, rodents can seriously damage the natural plant cover.

On the other hand, many rodents will, on occasion, eat animal food. Muskrats eat freshwater mussels and crayfish, as well as a wide variety of dead or dying water animals. Many rodents, will, at least occasionally, eat birds' eggs, nestlings, or insects. Several genera of South American cricetids live mostly on aquatic animals, of which fish make up the largest part. The grasshopper mice (Onychomys) of western United States and adjacent parts of Mexico and Canada are almost exclusively carnivores. Their food consists primarily of insects and earthworms, but may include birds and even other rodents that they kill. In the Arctic portion of their range, ground squirrels eat any carrion available, including their own dead relatives or even walrus or whale meat. They particularly like meat with a high fat content, presumably for the energy value.

Longevity. Rodents are generally considered to have a very short life span. This is certainly true in nature, where rodents are one of the primary food sources of carnivorous birds and mammals. Particularly among the small, mouselike rodents, the average life expectation at birth must be only a few weeks or months; relatively few individuals live much more than a year; and an animal two years old has reached a ripe old age. It is clear that this limitation is largely imposed on the animals by the activities of predators. Cricetids, for example, which normally live less than two years in the wild, have been kept up to 5½ years in captivity, chipmunks and ground squirrels up to seven or eight years, gerbils and dormice over five years, guinea pigs about eight years, chinchillas, and North American porcupines about ten years, and woodchucks, pacas, and the Old World porcupine even longer.

Limits of
life span

Predation on rodents. Because rodents are one of the most important items in their food supply, carnivorous birds (especially hawks and owls) and mammals provide a very strong selective force acting on small rodents. Many raptorial birds use their eyes for hunting and are able to detect any rodent slightly less well concealed than the average. It has been shown experimentally that owls selectively capture light-coloured mice over dark ones on a dark background and dark over light on a light background down to a light level of 10^{-7} of a footcandle. As a result of such selective pressure, rodents have evolved fur colours that closely match the environmental background where they live. For example, a late Pleistocene flow of dark lava in New Mexico is now inhabited by melanistic (dark) rats and mice, in strong contrast to the normally coloured rodents surrounding it and to the almost white rodents found in the nearby White Sands area.

Among mammalian carnivores, all but the very largest normally eat rodents, at least as an important part of their diet, and many will start teaching their young how to hunt at the expense of small rodents. The abundance of rodents, moreover, affects the numbers of their predators in at least some instances. The English ecologist Charles Elton has shown that there are striking parallels between cyclic abundance of lemmings and snowy owls and foxes in northern Labrador, the lemming maxima and minima preceding those of the predators by about six months to a year.

BEHAVIOUR

Gnawing. One of the most general habits of rodents is gnawing, much of which serves merely to wear down the incisors. These teeth grow continuously, at rates that have been found to range upward from two millimetres per week in the few species in which rates have been measured. Gnawing is performed by all rodents at frequent intervals. It involves the fore-and-aft movement of the lower jaw, the upper incisors holding a hard object and the lower incisors cutting against it. The same movements may be used with nothing held in the teeth, the lower incisors merely alternating between cutting against the rear side of the upper incisors and shifting forward so that the uppers cut against the rear of the lowers. Because the hard enamel is limited to the anterior side of the teeth, such use wears away the softer dentine at a faster rate, leaving the enamel to form a sharp chisel edge. Rodents prefer to use their incisors on hard objects, which provide part of the abrasion. These may be nuts, bark, tree trunks (as in beavers), often bones (which may also be gnawed for a supply of calcium), or even human possessions such as boards in houses and barns or even metal telephone cables.

Function
of gnawing

Nesting and denning. Food storage. A large proportion of rodents build underground homes, with a central nest chamber, in which they sleep, raise their young, and, often, hibernate. Many others have similar nests in grass or trees. Special food storage areas often are associated with such living quarters; for burrowing forms, these food areas are special chambers off the main passageways. Ground squirrels, kangaroo rats, and other rodents bring grass seeds back to such chambers in their cheek pouches (on the inside of the mouth of ground squirrels; outside the mouth and fur lined in the Heteromyidae and Geomyidae), and will often store several times as many seeds as they will eat during the season when they rely on their stores. Excavation of one kangaroo rat den, from which a single rat weighing 149 grams (about five ounces) was removed, showed nine underground storage chambers, containing from one to over eight quarts of seeds, with a total of almost 35 quarts. Kangaroo rats, in addition, dig storage pits, about an inch in diameter and an inch deep, near their burrows, which they fill with seeds. In one case, 875 such caches were located in an area of 5.1 square metres (55 square feet) adjacent to a single den. It has been estimated that seed collection in this manner may result in the loss of as much as a quarter of the grain crop in some parts of the world, and at times of rodent plagues, the destruction locally may approach totality. On the other hand, such underground stores are an important source of germinating seeds for wild grasses in steppe regions such as western North America and Central Asia.

The habit of squirrels of carrying acorns and nuts to hollow trees or barns, or digging holes in the ground in which the nuts are placed, is well known. The animal's memory of the location of these stores is poor, but if enough nuts or acorns are hidden, the animal will be able to find a sufficient number to keep him alive during the winter.

The North American pack rat or trade rat (*Neotoma* cinerea) is attracted by bright and shining objects, which it picks up to carry home to its nest, a jumble of sticks, twigs, grass, and assorted collectors items. A popular superstition is that this animal is a fair businessman, who, on seeing something he wants, always leaves a replacement that is, in his opinion at least, of equal value. The fact is that, while carrying one trophy, the rat may see another that is more attractive and so puts down the first to pick up the second, leading the human who has lost a possession to believe that the rat was carrying out a "business deal."

Beaver dams and lodges. Among the best known activities of rodents are those of beavers, which go to great lengths to store food, cutting alder and other food trees into 0.5 to 2.5 metre (two to eight foot) lengths and floating the logs in their ponds or embedding one end of these sticks in the mud of the ponds formed by their dams.

Constructions by
beavers

These sticks make an available food supply during the winter when the ponds are frozen. In order to provide an aquatic habitat, safe from their normal predators, beavers build dams that often run several hundred feet in length, and as much as two metres (more than six feet) high on the downstream side. The largest dam reported, from near Three Forks, Montana, was 622 metres (2,140 feet) long, 4.3 metres (14 feet) high at the highest point, and seven metres (23 feet) thick at the base. The dams are made of twigs, branches, and logs from which the bark has been eaten, piled on the ground with alternating layers of gravel or mud. When finished, the upper surface is plastered with mud, usually dug from the bottom of the pond above the dam, to make it watertight. The pond behind the dam serves as the refuge and first defense of the beaver. Consequently, beavers work continually to maintain the dams, by repairing, raising, and lengthening them, a process that may extend over several generations. Abandoned beaver ponds gradually silt up, forming very characteristic meadows.

Beavers also build lodges in the ponds behind their dams or burrows in the banks. Lodges are built up in the same manner as dams, and they are enlarged along with the rising water level of the ponds as the dam is raised. For ventilation, however, the upper part of the lodge is much more loosely constructed than the dam. The floor is kept above the water level, and there may be two levels, the lower for feeding, the upper for sleeping. The entrance to the lodge is through a tunnel, the outer end of which is below water level. Bank lodges are burrows in the bank, with the inside arranged like that of lodges. Again, access is gained through a tunnel with the mouth below water level. Burrows may extend more than nine metres into the bank.

Beaver canals are less well known but highly remarkable. They are dug to permit the transport of birch or aspen logs for food supplies from the sources where they are cut to the ponds. The logs are floated in the water, where the beaver is much safer than on land. Such canals are dug whenever they can be made with little or no damming. They vary from 0.3 to 1.2 metres (one to four feet) in width, up to 0.6 metres (two feet) deep, and canals as long as 223 metres (732 feet) have been reported.

Beavers live on the bark of trees, that of aspen being their favourite, but birch, cottonwood, willow, and alder also are used. When these are in short supply other plants are eaten, but conifers are rarely cut except for building materials. In one night a single adult beaver can fell a seven- to ten-centimetre (three- to four-inch) poplar, cut it into sections one to three metres long, and drag the logs to the water. Normally, beavers do not cut up logs over 15 centimetres in diameter, although they will fell trees considerably larger. The largest tree on record cut by beavers was 115 centimetres (46 inches) in diameter. The cutting of trees is apparently done at random, with all gnawing likely to be on one side, unless the tree has a large diameter, in which case it may be gnawed all the way around. Because trees normally fall downhill, hence toward a pond, this is probably the origin of the legend that beavers can make trees fall as they wish.

Types of dormancy

Hibernation. Many rodents use their nests or burrows for shelter during unfavourable seasons, especially winter. There is considerable variation, even among closely related animals, in the amount of activity that occurs at such times and how often animals wake and eat some of the stored food. In a considerable number, including chipmunks, pocket mice (Perognathus), some jerboas, and some field mice, there is extensive torpor in wintertime, the animals sleeping extensively, but waking two or three times a day to eat. Hamsters (Cricetus and *Meso*cricetus) are deep hibernators, going into a sound sleep from which they awaken only periodically to eat. The most pronounced type of hibernation involves the accumulation of extensive amounts of body fat during the late summer and early fall, and deep hibernation with few or no awakenings during the winter. This condition occurs in marmots, some ground squirrels, a large proportion of dormice (whose English name is derived from the French *dormir* meaning "to sleep"), the jerboa (Allactaga), and most or all zapodids (jumping mice and birch mice). Estivation, or summer sleep, is rare in rodents but occurs in woodchucks and some ground squirrels. No suggestions of hibernation occur in murids, caviomorphs, or phiomorphs, any other tropical rodents, or rodents with a recent tropical or subtropical ancestry.

Population **movements.** Several types of rodents, subject to periodic fluctuations in numbers, reduce the amount of overpopulation by migration. Generally, the migration is simply the movement of juveniles away from a home area that has become fully populated. Migration of such animals may be quite extensive, but, because it is usually an individual phenomenon, it is often not noticed. Such animals are subject to very high mortality from predators because they are moving in unaccustomed territory and without the shelter of a permanent home. Muskrats and beavers are examples that are more readily noted, because they move from one suitable stream to another and may travel as much as several miles across country looking for a suitable home. There are many reports of major migrations. In August 1946, gray and fox squirrels migrated out of an area in northwest Wisconsin, apparently due to a shortage of acorns. When they reached streams and rivers, they crossed them by swimming, many drowning in the wider rivers. In 1935 a mass movement of gray squirrels was noted from the area east of the Hudson River, to the west. Hundreds (one report says 2,000) of drowned squirrels were found along the west bank of the Hudson River.

Lemming migrations

The most famous migrations of rodents, however, are those of lemmings, especially in Scandinavia. There tend to be two migrations per year, one in spring and one in autumn. The exact causes of the movement are not known, but a change of habitat with season seems to be the most important single factor. It has been believed that the migrations are caused by overpopulation and a shortage of food, or by claustrophobia (a feeling of mental anguish from the crowded conditions), but these probably are not valid causes.

The lemming migrations are individual activities rather than unified movements, as is popularly supposed. Careful observations of migrations show that each individual acts alone, but that there may be groups of several individuals within a few minutes, followed by a gap of ten minutes or more. In autumn, at least, the migrations normally occur at night. Individual lemmings generally move fairly rapidly in a generally constant direction, although they will occasionally stop to feed or groom. Rates of migration have been calculated as about one metre per second (2.25 miles per hour), but such speeds are not sustained, and it is probable that distances of eight–ten kilometres (five–six miles) per day are normal. The migrations tend to be oriented by roads or paths made by humans, trails made by reindeer, or other similar routes, although the general direction seems to be outward in all directions from the general source area and to be determined before the movement begins. Contrary to the older reports and popular belief, lemmings do not rush forward without hesitation into water when reaching a shoreline. They apparently try to avoid swimming, if possible, running up and down the shore looking for a narrow crossing or an area where they can cross on ice. Apparently, they enter water willingly only if they can see the silhouette of the far shore. This is in marked contrast to activities of their relative, the water vole (Arvicola), which always heads for the water when threatened. When swimming lemmings use their hind legs almost exclusively but swim very high out of the water. Lemmings become exhausted and drown after swimming 15 to 25 minutes in water with waves about 15 centimetres (six inches) high. Observations along the Swedish-Finnish border show that the lemmings had no problem crossing a strait 200 metres (650 feet) wide on a calm night, but that considerable numbers drowned on a windy night. A lemming was observed swimming at approximately the middle of a lake two kilometres (1.2 miles) wide. Reports of lemmings swimming out into the North Sea and being found alive 16 kilometres (ten miles) or more from shore must be viewed skeptically. Recent studies in Scandinavia indicate that the migrations are a normal part of the activity of the lemmings, serving not only as a means of dispersal of the species but also to permit the animals to occupy different habitats at different seasons.

Rodents are often thought of as defenseless animals, but the incisors are very sharp and powerful, are activated by powerful jaw muscles, and can inflict a deep wound. Among themselves, rodents fight for territories, the possessor of a territory usually being able to drive interlopers away. They also fight very effectively in self-defense, as indicated by the popular expression, "fighting like a cornered rat." A full-grown wharf rat is capable of defending itself against a determined tomcat. Even animals as small as lemmings are able to discourage attack by weasels.

REPRODUCTION

Flexibility in breeding habits

The reproductive habits of rodents are exceedingly varied in nature and are capable of being even further modified in domesticated or partially domesticated forms. Many, especially the larger types, reproduce once a year. Others produce several litters during a single season. Some have only one or two young at a time; others, large numbers. Some are capable of reproduction at very early ages, others (particularly members of the Caviomorpha) not until after a considerable period of growth. Most are born naked and helpless, and are cared for in nests; a few are able to run and keep up with their mothers almost at once. Most rodents are polygamous; some mate for the duration of a single breeding season; and a few (beavers, for example) have permanent mates.

Beavers do not breed until the second January or February after birth, producing a litter of two or four young after about **12** weeks of gestation. The capybara has a litter that is, on the average, slightly larger, after **15** to **18** weeks gestation; the viscacha, a smaller animal, has two young in the spring about **22** weeks after breeding. Nutrias breed twice a year after reaching an age of one year, if the winter temperature does not fall below **60**" **F** for lengthy periods, and produce litters averaging about five young. The European porcupine (Hystrix cristata) breeds in the spring, with one to four young being born about **16** weeks later, and some of its tropical relatives (other members of the Hystricidae) produce two litters a year. The North American porcupine (Erethizon dorsatum) of the caviomorph family Erethizontidae, has a single young after **30** weeks gestation. Most of the smaller members of the order, however, have considerably higher reproductive rates. In many cases there may be more than one litter per year, with high numbers per litter, examples being shown in the Table.

Reproductive Rates of Certain Rodents

animals	litters/year	young/litter
North American gray squirrel (*Sciurus carolinensis*)	*2*	*2–3*
Eurasian gray squirrel (*S. vulgaris*)	*2*	*5–7*
Red squirrels (*Tamiasciurus*)	2	*4–6*
Pocket gophers (*Geomys*)	*1–2*	1–4
Pocket mice (*Perognathus*)	1–2	*2–8*
Kangaroo rats (*Dipodomys*)	1–3	*2 4*
Deer mice (*Peromyscus*)	**up to** *4*	*2–7*
Hamsters (*Cricetus*)	**several**	*6–12*
Lemmings (*Lemmus*)	**several**	*3–9*
Muskrats (*Ondatra*)	*2–5*	*5–7*
Voles (*Microtus*)	**up to** *13*	*4–8*

Source: E.P. **Walker,** ***Mammals of the World (1964)*** **and** S.A. Asdell, ***Patterns of Mammalian Reproduction (1964).***

This rather high rate of breeding is intensified by the fact that, in many of the smaller rodents, sexual maturity is reached at an early age, normally earlier in the females than in the males. The females breed when less than a year old in many squirrels, some pocket gophers, and the pack rat. In most of the Cricetidae, however, the young appear to reach sexual maturity considerably earlier—harvest mice (*Reithrodontomys*) in five weeks; the hamster (Cricetus) in six weeks; and voles (Microtus) in six to seven weeks.

Reproductive potential

In the United States, wild populations of the house mouse reproduce throughout the year, with an average of **5.5** litters and **31** young per female per year in buildings, and **10.2** litters and **57** young per year on farms. In the laboratory mouse, the mean litter size has been reported as ranging from **4.5** to **7.4,** depending on the strain. The second litter is the largest, after which there is a steady decrease. Litter size has been reported to have a range of two to **12,** and as many as **19** healthy embryos have been removed just before term from a single female. There may be five or more litters per year. The gestation period is about **20** days, and the first mating usually occurs at seven to ten weeks of age, though it has been reported that, in one strain of albino mice, the first estrus occurs, on the average, at **39** days and results in about **50** percent pregnancies. The breeding span of most mice in the laboratory is **12** to **18** months. Lactating females may become pregnant, but the gestation will be lengthened by one to two weeks.

Wild Norway rats breed throughout the year, taking advantage of the sheltered environment available from living in association with humans. The size of the litter is correlated with the size of the mother. In the laboratory female rats reach sexual maturity in **33** to **120** days. The number of young varies with the strain, two typical ones having litters averaging **6.5** and **8.9.** Normally, a female's second or third litter is the largest, and the litter size decreases rapidly after the tenth. Over a number of generations of laboratory life in a colony of gray rats, however, the total number of young per female grew from an average of **23** to an average of **63.**

The laboratory guinea pig has a gestation period of about **68** days, a very long period for such a small animal but characteristic of caviomorphs. There are usually three or four young per litter. The age of puberty varies from **55** to **70** days.

Although most rodents seem to increase their reproductive potential under domestication, this is not true of chinchillas. In the wild, they produce litters one to six after about **15** weeks pregnancy; domesticated ones seem normally to have only one offspring at a time.

FORM AND FUNCTION

In most aspects rodents are relatively primitive, not highly specialized mammals. The skeleton is usually that of a quadrupedal, scampering mammal, not far from the basic placental stock. The digits are clawed and usually five in number, with little or no opposability of the first digit. The limbs are not far from the same length, although the hind legs are always somewhat longer. The animals are adapted to a broadly herbivorous diet. The main specialization of the digestive system is the possession of a large cecum (a blind pouch) at the junction of the large and small intestines.

Specializations for gnawing. As opposed to this basic primitiveness, the rodents have developed, as their basic innovation, a gnawing mechanism of an efficiency that is not approached by that of any other mammals. The gle-

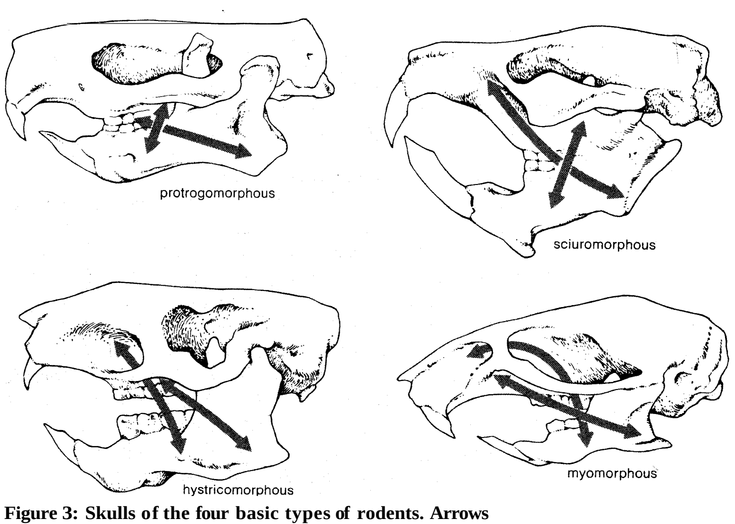


Figure 3: Skulls of the four basic types of rodents. Arrows show positions of branches of the masseter muscle.

noid fossa, the concavity where the lower jaw articulates with the skull, slopes downward, from rear to front, and there are no processes (as there are in all other mammals) at the anterior and posterior limits of the fossa that restrict the anteroposterior movement of the jaw. The jaw, therefore, is capable of functioning when pulled to the rear, separating the incisors from each other, and permitting chewing or grinding by the cheek teeth. The jaw can also be pulled forward and downward, separating the cheek teeth and bringing the tips of the incisors together for gnawing.

Evolutionary shifts in jaw muscle attachments

These shifts in the position of the jaw, as well as the functioning of the jaw in either position, are brought about by a combination of the actions of the various jaw muscles, of which the temporal, pterygoid, and masseter are the most important. The forward and backward (anteroposterior) movements, including those of gnawing, were caused, primitively, by all three muscles, all of which have a slight anteroposterior component. But a shift of the anterior muscle, the masseter, began to occur during the Eocene Epoch (about 50,000,000 years ago). Originally, in a condition called protrogomorphous, the masseter ran from the zygomatic arch (cheekbone) to the lower jaw. Parts of the muscle, retaining the original points of insertion on the lower jaw, shifted their origins forward onto the snout, doubling the length of the muscle and greatly changing its direction of pull. In some rodents it was the lateral portion of the masseter that shifted forward onto the face, compressing the infraorbital foramen (an opening in the bone), through which pass nerves and blood vessels, between the muscle and the snout (the sciuromorphous condition); in others it was the deeper portion of the masseter that shifted forward, through the infraorbital foramen (presumably following an initial enlargement of the foramen), onto the face, resulting in further enlargement of the foramen until it may now be larger than the eye socket (the hystricomorphous condition); and in still others, a combination of the two shifts has occurred (the myomorphous condition). Although these changes are of importance in determining rodent taxonomic relationships, they are probably not as controlling as was once thought, as each structural condition may have originated more than once.

Continuous growth of incisors

But the main structural change from the primitive mammalian condition has occurred in the incisors. In all known rodents, beginning with the very earliest, the incisors are reduced to a single pair in each of the upper and lower jaws. These teeth grow throughout life, the enamel cap being restricted to the anterior 30 to 60 percent of the perimeter of the tooth. The rate of incisor growth has been measured in several rodents, and varies from about two to three millimetres per week in nonburrowing forms to five millimetres per week in pocket gophers, in which the incisors are used for digging as well as for gnawing. In hibernating rodents, the incisors continue to grow, but at a greatly reduced rate. If the incisors do not meet, due to deformity or breakage, the teeth will continue to grow indefinitely. Since the lower teeth are arcs of large circles, the lower incisors simply curve forward and upward, becoming completely nonfunctional. But the upper incisors are large arcs of much smaller circles. With growth, they may spiral outward, like the spirals on a handle-bar moustache, or they may grow around and upward through the throat between the lower jaws, growing through the snout and eventually locking the jaws closed. Either condition is fatal, but rodents have survived in nature for a long time with such malformed incisors.

Because of the continual growth of their incisors, rodents have an unusually heavy need for calcium, which, together with a need for abrasion of the incisors, explains the frequency with which rodents gnaw bones.

Adaptations for specialized locomotion. Many rodents have shifted their locomotion from the characteristic scampering type. In burrowing rodents, the limbs are short and massive, and the hands, particularly, are widened and strengthened with heavy claws. Many, but not all, burrowing rodents use their incisors to help in digging. In such cases, a fold of skin is likely to close the mouth behind the incisors, and the radius of curvature of both upper and lower incisors is increased, so that they extend forward from the mouth, cutting the soil an appreciable distance in front of the animal. Eyes and ears are generally reduced in size in these forms.

There is little difference, structurally, between the terrestrial scampering rodents and the arboreal scamperers, such as squirrels. The claws of the latter are sharper, but locomotion is basically the same. A few rodents, such as the New World porcupine, actually climb (as would a human) rather than running up and down the bark, as does a squirrel. Only one rodent, the tree porcupine (Coendou) of Central America and northern South America, has a truly prehensile tail (although several rodents have semiprehensile tails that they can wrap around branches for extra support). A number of gliding rodents, "flying squirrels" of one sort or another, some of which are not closely related to the true squirrels (Sciuridae), are supported in the air by a fur-covered membrane (patagium) extending between the fore and hind limbs and usually between the hind limbs and the tail.

A rather common development among rodents has been the reduction of the forelimbs and an increase in the size of the hind limbs and tail for locomotion by means of a series of kangaroo-like leaps, which in several forms can exceed 15 feet.

Adaptations for water conservation. With their wide geographic distribution, rodents have invaded desert regions in all parts of the world. Many of these have developed the ability to function on limited supplies of water. This has been achieved in several ways. Some are nocturnal, remaining asleep during the heat of the day in burrows deep enough to reach ground moisture, thus reducing evaporation. Others seek out succulent desert plants for food. But it has been shown that a number of desert rodents (kangaroo rats and jerboas, for example) have the ability to live exclusively on seeds, with no external water source, and to obtain all of their needed water from their metabolism. Such forms also produce a greatly reduced quantity of extremely concentrated urine and a small amount of feces. Two species of spiny mouse *(Acomys)* were investigated in Israel. Both inhabit the same desert areas, but one is diurnal and the other is nocturnal. Both produce urine the urea content of which is extremely high. Neither species can survive on a diet of barley without water to drink, although gerbils and jerboas from the same area showed no ill effects after four weeks without water. The diurnal spiny rats, however, were able to survive and grow with no water except seawater, whereas the nocturnal ones could not tolerate such a salt concentration.

EVOLUTION AND PALEONTOLOGY

Rodents are relatively poorly represented in collections of fossils, in spite of their great abundance at the present time. They have often been overlooked because of their small size, but modern intensive exploration for fossils usually results in a much more abundant representation of small mammals, especially rodents.

Paleocene. The earliest known rodents come from the late Paleocene (about 57,000,000 years ago) of North America, by which time they had already acquired all of the diagnostic features of the order. The ancestral family, the Paramyidae, was also present in Europe, where it first appeared in the earliest Eocene. There was very rapid diversification of the order during the Eocene, initially involving the Paramyidae, but with other families soon appearing. Most of the Eocene rodents were protrogomorphous, the masseter muscle restricted to its primitive origin on the cheekbone. By the middle Eocene of Europe, rodents with advanced types of masseters were present, and others occurred in the late Eocene of North America. The skeleton of the paramyids was basically that of a generalized scampering animal, approximating a rat in its method of locomotion. Before the end of the Eocene it is probable that both leaping and burrowing variations had arisen, although there is some uncertainty on this point.

Paramyidae, the earliest rodents

Oligocene. A major gap in the knowledge of rodent evolution occurs at the Eocene-Oligocene boundary (about 38,000,000 years ago). A number of modem families with advanced types of jaw muscles appeared at about the same time in North America, Europe, or Asia, including the Cricetidae, Heteromyidae, Geomyidae, Castoridae, and Ctenodactylidae, with no good indications at present as to their geographic or ancestral sources. At about the same time, rodent groups appeared suddenly in South America and Africa (Caviomorpha and Phiomorpha, respectively). Both groups are hystricomorphous and hystricognathous and have often been considered to have been related, though others consider them to have acquired their similarities independently.

Miocene. The majority of the living families of rodents had appeared by the Oligocene, and most of the remaining ones occur in the Miocene (around 26,000,000 years ago), so that the later part of the evolution of the rodents was largely a diversification of the various groups. The ancestors of a wide variety of specialized modem forms can be recognized from teeth, but most of them are not known from skeletons, and the evolutionary status of the modern locomotor conditions is not well known. Highly evolved burrowing rodents were certainly present in the Oligocene in Mongolia, however, and less specialized ones occurred in North America. Ancestral kangaroo rats had attained a leaping ability by the Miocene of North America. Gliding and burrowing forms are known from the Miocene of Africa. In South America, the differentiation of the descendants of the invaders present in the early Oligocene proceeded rapidly, and the Miocene forms occupied nearly all potential rodent niches on that continent.

Richard Keane

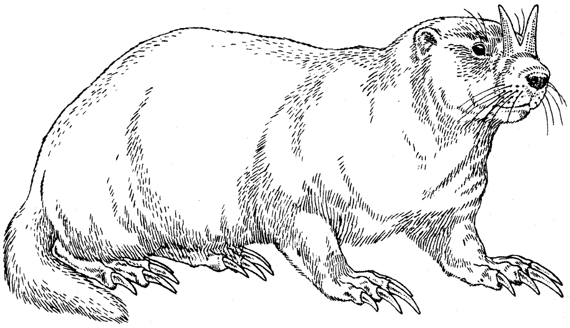


Figure 4: Reconstruction of *Epigaulus*, a primitive Pliocene rodent.

Later evolution of rodents. The latest rodents to differentiate apparently were the members of the family Muridae. This group appears suddenly in early Pliocene deposits of Europe, probably having invaded that continent from Asia, the southeastern part of which, together with the adjacent East Indies, is now considered the evolutionary centre of the family. The Muridae expanded rapidly all over the Old World. An isolated incisor is known from the Pliocene (approximately 7,000,000 years ago) of New Guinea, where considerable local evolution has occurred. The murids reached Australia late in the Cenozoic Era (they are unknown there before the Pleistocene), and underwent a rapid evolution, developing arboreal, burrowing, and leaping forms, paralleling the evolution of the entire order in the rest of the world.

Appearance of Muridae

When North and South America were united in the beginning of the Pleistocene (about 2,500,000 years ago), there was a major invasion of South America by cricetids, which expanded rapidly and became highly diverse in that area.

An interesting aspect of rodent evolution are the very rapid diversifications (evolutionary explosions) that occurred whenever highly specialized members of the order reached areas previously uninhabited by gnawing mammals (South America and Africa in the Oligocene, New Guinea in the Pliocene, Australia in the Pleistocene). These, apparently, resulted in a rapid spread into most of the major available ecologic niches in a matter of, at most, 2,000,000 or 3,000,000 years. The differentiation of the South American cricetines since the beginning of the Pleistocene has not been quite as fast, though still explosive, presumably because there already were numerous rodents on that continent. The initial diversification of the rodents during the Eocene seems to have been considerably slower, probably because they had not yet become as thoroughly adapted for gnawing as was the case later.

Because of their small size and great abundance, rodent fossils are becoming progressively more important as guide fossils that enable accurate separation of successive geologic horizons.

CLASSIFICATION

Distinguishing taxonomic features. Many features have been used to determine the relationships of the subgroups of the rodents. The most generally used are the differentiations of the jaw muscles or the modifications of the skull and jaws resulting from muscular changes. The masseter muscle may have its primitive position, arising from the cheekbone (protrogomorphous); its lateral division may have shifted forward onto the snout, compressing the infraorbital foramen against the snout (sciuromorphous); the medial division may have moved forward through the infraorbital foramen, onto the snout, greatly enlarging the foramen (hystricomorphous); or both branches of the masseter may have shifted (myomorphous). In most rodents the angular process at the rear of the lower jaw extends downward from the ventral side of the alveolus of the incisor (sciurognathous); in some, all of which are also hystricomorphous, the angle arises from the side of the alveolus (hystricognathous). Other features that have been given important weight in delimiting the major subgroups include crown patterns of the cheek teeth (premolars and molars); histologic structure of incisor enamel, which may have one layer (uniserial), a few (pauciserial), or many (multiserial); structure of the penis or of the baculum (os penis); structure of the male genital tract, including particularly the presence or absence of an outgrowth, the sacculus urethralis; fusion or lack thereof of two of the ear ossicles, the malleus and incus; and the pattern of development of the extra-embryonic fetal membranes.

ANNOTATED CLASSIFICATION

The classification presented here is based upon that of American paleontologist A.E. Wood. The rodents are one of the outstanding examples of a group in which closely similar structures have evolved numerous times independently, a phenomenon known as parallelism, with the result that there is very strong disagreement among taxonomists working with the group as to how the order should be subdivided.

Groups indicated by a dagger (†) are known only from fossil remains. Dental formulas indicate the number of pairs of teeth present in the upper jaw (above the line) and lower jaw (below), the figures representing, respectively, pairs of incisors, canines, premolars, and molars. When a partial formula is given the type of teeth is indicated by I, C, P, or M.

ORDER RODENTIA

Gnawing mammals with ever-growing incisors, reduced to a single pair in both upper and lower jaws and with the enamel limited to the anterior face of the incisors, so that wear maintains a sharp chisel edge; glenoid fossa slanted, with neither preglenoid nor postglenoid process, masseter the principal jaw muscle; dental formula reduced, never exceeding $\frac{1.0.2.3}{1.0.1.3} = 22$. About 350 living genera and 2,400 living species; over 400 extinct genera have been described.

Suborder Sciuromorpha

Masseter primitively either limited to zygoma, or rarely with only a slight forward displacement, but in the squirrels shifted forward onto face (sciuromorphous). Always sciurognathous. No sacculus urethralis. Malleus and incus bones of ear separate. Incisor enamel pauciserial or uniserial. Paleocene to Recent.

†Family Paramyidae

Paleocene to early Miocene; Northern Hemisphere. Cheek teeth cuspidate, clearly derived from basic mammalian tribosphenic type. Tympanic bullae became co-ossified with the skull several times independently, within the family. Incisor enamel pauciserial, with one possible exception. Presumably ancestral to the rest of the order. One genus incipiently hystricognathous. From size of a mouse to as large as a beaver. Locomotion, when known, scampering, with possible incipient leaping in one form.

†Family Sciuravidae

Eocene of North America; one genus from Eocene of Central Asia. Cheek teeth formed of 4 transverse crests rather than of separate cusps; locomotion probably scampering; mouse-size to rat-size. Incisor enamel pauciserial. Perhaps ancestral to several mouse- or ratlike groups.

†Family Ischyromyidae

Oligocene of North America. Masseter muscle has begun to move forward onto the snout in some forms, in primitive position in others. Cheek teeth 4-crested. Incisor enamel uniserial. Locomotion scampering. Size of a gray squirrel or somewhat larger.

†Family Cylindrodontidae

Middle Eocene to Oligocene of North America, Oligocene of Central Asia. Burrowing rodents with high-crowned to evergrowing cheek teeth, based on a pattern of 4 transverse crests. The most specialized forms used the incisors for digging, as shown by forward extension of these teeth. Incisor enamel uniserial. Size from that of a chipmunk to that of a marmot. Presumably derived from North American Sciuravidae.

†Family Protoptychidae

Late Eocene of North America. Skull with highly inflated auditory bullae, suggesting, by analogy with other rodents, that these were leaping animals. Cheek teeth high crowned, with pattern of 4 crests. Size of a gray squirrel.

Family Aplodontidae (mountain beaver, or sewelled, and fossil relatives)

Late Eocene to Recent of North America, Oligocene and Miocene of Europe, Pliocene of Asia. Single living species restricted to wet areas from British Columbia south to San Francisco Bay. Extreme hypsodont cheek teeth, based on cuspidate rather than crested pattern. Incisor enamel uniserial. Burrowers. Head and body length of living species (***Aplodontia rufa***) 300 to 460 mm, with a short tail; weight 900 to 1,800 g.

†Family Mylagaulidae

Miocene to Pliocene of North America. Strongly hypsodont cheek teeth, with greatly enlarged premolars, whose continual growth forced 1 or more of the anterior molars out of the jaws. Incisor enamel uniserial. Some individuals had paired horns on the snout. Whether these are taxonomic or sexual characters has not yet been demonstrated. Powerful burrowers. About the same size as ***Aplodontia.***

Family Sciuridae (squirrels, chipmunks, and marmots)

Oligocene to Recent of Northern Hemisphere, Miocene to Recent of Africa, Recent of South America. Highly developed sciuromorphous pattern, with masseter muscle extending forward along the side of the snout and compressing the infraorbital foramen. Postorbital process of frontal separates the eye from the temporal muscle. Cheek teeth typically very similar to those of the Paramyidae, but some forms have rather complicated tooth patterns. Incisor enamel uniserial. About **70** living genera.

Suborder Myomorpha

Lateral branch of masseter muscle usually displaced forward alongside of snout, forcing infraorbital foramen against side of snout; in most forms, deep masseter penetrates a variable distance through upper part of infraorbital foramen. Sciurognathous. Incisor enamel uniserial. Locomotion scampering, jumping, arboreal scampering, or fossorial. A few forms are partially adapted to an aquatic life. Malleus and incus never fused. No sacculus urethralis. Cheek teeth usually reduced in number, with only a single fossil genus known that retains the primitive rodent dental formula; except for the Geomyoidea, the premolars are either lost or greatly reduced; the last molars are occasionally lost as well. Most members of the group are small, from the size of a mouse to that of a rat (head and body length ranges from 50 to 300 mm, except in the Rhizomyidae, which may be somewhat larger).

Family Cricetidae (field mice, deer mice, voles, lemmings, muskrats)

Early Oligocene to Recent of Europe and North America, middle Oligocene to Recent of Asia, late Pliocene to Recent of South America, Pleistocene of Madagascar, Recent of Africa. Deep masseter expanded through upper part of infraorbital foramen to face; no premolars, molars $\frac{3}{3}$; cheek tooth pattern based on cusps arranged into 5 transverse crests in both upper and lower teeth; teeth low crowned to very high crowned. Scampering, fossorial, arboreal scampering, jumping, or partly aquatic.

Family Muridae (Old World rats and mice)

Early Pliocene to Recent of Europe and Asia. late Pliocene to Recent of East Indies, Pleistocene to Recent of Africa and Australia. Introduced throughout the world by man. Deep masseter as in Cricetidae; cheek teeth normally the 3 molars, but one subfamily, the Hydromurinae of Australia and New Guinea, has lost the third molars, reducing the cheek teeth to $M\frac{2}{2}$, tooth pattern based on rounded cusps, arranged in transverse rows, but derivable by modification of a pattern like that of primitive cricetids; low to medium high crowned teeth. Scampering, arboreal scampering, occasionally fossorial or semiaquatic.

Famiy Heteromyidae (pocket mice, kangaroo rats and mice)

Oligocene to Recent of North America, Recent of northern South America. These and the next two families are sciuromorphous, with no penetration of the infraorbital foramen by the masseter, and seem to be closely related. Fur-lined cheek pouches, opening beside the mouth and reaching back to the shoulders, are used for the transportation of food to underground storage areas near the nests. Central stock of family scampering; specialized ones leap with their hind legs. Generally small, mouselike; length of head and body from 55 to 180 mm. Cheek teeth, which include P) $\mathbf{M\frac{3}{3}}$, are low crowned to very high crowned. Pattern of teeth based on **2** transverse rows of 3 cusps each. Derivable from late Eocene members of the Eomyidae.

Family Geomyidae (pocket gophers)

Early Miocene to Recent of North America. Like heteromyids in jaw muscles, dental formula, and tooth pattern, and in possession of fur-lined cheek pouches. Highly adapted burrowing animals. Cheek teeth have already become high crowned in the Miocene, and those of the living forms grow throughout the animal's life, with the enamel reduced to a plate on the anterior side of the upper teeth and the posterior side of the lowers.

†Family Eomyidae

Late Eocene to late Pliocene of North America; late Eocene to Pleistocene of Europe. Sciuromorphous jaw muscles; cranial anatomy and jaw structure very similar to heteromyids. Cheek teeth usually low crowned, with pattern of 5 crests, rather similar to that of the cricetids, although some members of the family developed teeth suggesting that they might be ancestral to heteromyids. Cheek teeth usually P) $\mathbf{M\frac{3}{3}}$, but **1** genus had $\mathbf{P\frac{2}{1}}$. Skeleton and habits unknown.

Family Zapodidae (birch and jumping mice)

Late Oligocene to Recent of Eurasia, early Miocene to Recent of North America. **A** late Eocene genus from California is often (perhaps incorrectly) placed here. The Old World birch mice are arboreal scampering forms, as were probably all the earlier fossil members of the family. The jumping mice are as saltatorial as the kangaroo rats. lnfraorbital foramen of medium size, with the deep masseter passing through it but with the superficial masseter remaining on the zygomatic arch. Cheek teeth $\mathbf{P\frac{1}{0}\,M\frac{3}{3}}$, with pattern very similar to that of cricetids, with which the fossils have frequently been confused.

Family Dipodidae (jerboas)

Late Oligocene to Recent of Asia, late Miocene to Recent of Europe, Pleistocene to Recent of North Africa. Masseter muscle like that of zapodids but with larger deep division. Highly saltatorial rodents of the steppes and deserts of the Old World, with extremely inflated auditory bullae. Cheek teeth $\mathbf{P\frac{1-0}{0}\ M\frac{3}{3}}$, high crowned in all living and most fossil forms. Tooth pattern suggestive of that of zapodids, from which the family is probably descended. General tendency to elongate the hind foot and to fuse the 3 median metatarsals to form a cannon bone, as part of the leaping adaptation.

Family Spalacidae (mole rats)

Early Pliocene to Recent of Europe, Pleistocene to Recent of Western Asia and North Africa. Highly specialized burrowing rodents with short, powerful legs and no external tail. Eyelids permanently closed. Masseter muscle a short distance forward on face. Cheek teeth reduced to $\mathbf{M\frac{3}{3}}$ (or possibly $\mathbf{P\frac{1}{1}\ M\frac{2}{2}}$). Ancestry unknown.

Family Rhizomyidae (African mole rats, bamboo rats)

Oligocene of Europe, early Miocene to Recent of Asia, and Pleistocene to Recent of Africa. Myomorphous, with considerable forward extension of the masseter. Burrowing animals with powerful legs and short tails, superficially resembling pocket gophers (Geomyidae). Cheek teeth high crowned to ever-growing, consisting of $\mathbf{P}\frac{1-0}{1-0}$ $\mathbf{M}\frac{3}{3}$. The European Oligocene *Rhizospalax* has features suggesting that this family and the Spalacidae are related. Otherwise, relationships unknown.

Suborder Caviomorpha

Deep branch of masseter has shifted its origin forward onto the face, passing through the very large infraorbital foramen. Angle of jaw hystricognathous. Malleus and incus fused; usually a sacculus urethalis in male genital tract. Incisor enamel multiserial. Cheek teeth usually P) $\mathbf{M}\frac{3}{3}$, but occasionally the milk premolar is retained throughout life, the formula being $\mathbf{dP}\frac{1}{1}$, $\mathbf{M}\frac{3}{3}$. Early Oligocene to Recent of South America, Pleistocene to Recent of North America and West Indies.

Familv Octodontidae (octodonts, degus)

Early Oligocene to Recent of South America, Pleistocene to Recent of the West Indies. Small for caviomorphs, generally somewhat ratlike in appearance. Scampering to fossorial in habits. Occur from sea level to over 3,000 m elevation in the southern half of South America. The teeth range from low crowned to ever-growing, the enamel of the grinding surface arranged in the shape of a kidney or figure 8. The burrowing forms use the incisors in digging. The genus Platypittamys from the early Oligocene of Patagonia is very close to being a common ancestor of all the Caviomorpha.

Familv Echimvidae (spiny rats)

Oligocene to Recent of South America. Pleistocene to Recent of West Indies and Central America. Ratlike in appearance. Usually with spiny fur, although there are a few exceptions. Cheek teeth rooted, with a pattern formed by transverse folds of enamel. In all but the earliest known member of the family, the milk or deciduous premolars are retained throughout life, and permanent premolars never make an appearance. Inhabit moist forested regions; climbing or scampering; one genus is burrowing.

Familv *Ctenomvidae* (tuco-tucos)

Pliocene to Recent of South America. The single living genus inhabits all of the southern half of South America, from sea level to elevations over 4,000 m. They are fossorial and presumably derived from octodontids. Body form and size are very similar to those of the North American pocket gophers, with powerful digging muscles in the forelimbs and long, powerful claws.

Family Abrocomidae (chinchilla rats or abrocomes)

The common name comes from the soft underfur and ratlike appearance. Pliocene to Recent of South America. The living species inhabit mountainous areas of Peru, Bolivia, Argentina, and Chile. Presumably evolved from octodontids. The cheek teeth grow throughout life.

Family Chinchillidae (chinchillas and viscachas)

Eary Oligocene to Recent of South America. One living genus (Lagostomus) lives in lowlands of Argentina; the other two (Lagidium and Chinchilla) live at elevations of about 800 to 6,500 m in the southern half of the Andes. The fur of Chinchilla is very soft and dense—that of the other genera less so. All gregarious, but especially the lowland viscachas (*La*gostomus), which occupy colonies resembling those of prairie dogs. The lowland viscachas are large, head and body length being 470 to 660 mm; the mountain genera are 230 to 320 mm (Lagidium) and 225 to 380 mm (Chinchilla). Cheek teeth ever-growing, but details of pattern lost early in life.

*Family Caprom*yidae (hutias, coypus)

Middle Pliocene to Recent of South America, Pleistocene to Recent of West Indies, introduced in southern United States and parts of Europe. Differ from the Chinchillidae, from which they are probably derived, in having high-crowned to ever-growing cheek teeth, but with the details of the pattern persistent. The hutias of the West Indies resemble large rats. The coypus of South America are considerably larger, looking rather like a large muskrat. The fur is excellent, and the flesh is widely eaten in South America.

Family Dasyproctidae (pacas and agoutis)

Early Oligocene to Recent of South America, Pleistocene to Recent of West Indies, Recent of Central America. Large rodents (head and body length 320 to 800 mm), with weight up to 10 kg. Limbs modified for cursorial locomotion, with the lateral toes, especially on the hind foot, reduced in size. Flesh is very palatable. Pacas in moist forested areas from Mexico to southern Brazil; agoutis in the same regions (plus the Lesser Antilles), but not restricted to forested areas.

Family Dinomyidae (pacaranas)

Early Miocene to Recent of South America; Pleistocene of West Indies. There is only a single living genus inhabiting the lower elevations of the northern half of the Andes The animal is among the largest of living rodents. A considerable number of fossil forms, once thought to represent a separate family (Heptaxodontidae) probably belong here.

†*Family Elasmodontomyidae*

Pleistocene to Recent of Puerto Rico and northern Lesser Antilles. These forms are all extinct, but survived until about the time the area was colonized by the Spaniards. Medium-sired ground-living rodents, with ever-growing cheek teeth formed of a series of enamel plates inclined at about 45° to the long axis of the jaws.

†*Family* Eocardiidae

Early Oligocene to middle Miocene of South America. Cheek teeth medium to high crowned. Size about that of a guinea pig.

Family Caviidae (guinea pigs, cavies, and maras)

Late Miocene to Recent of South America. Cheek teeth ever-growing, with a simplified crown pattern of alternating V's. Digits reduced to 4 on the front and 3 on the hind foot. Cavies with short legs, ears and tails; maras superficially resembling hares.

Family Hydrochoeridae (capybaras)

Early Pliocene to Recent of South America, Pleistocene to Recent of Central America, Pleistocene of West Indies and southern United States. Most members of this family are extinct, there being but a single living genus, with 1 species in eastern Panama and the other in South America east of the Andes and north of the Rio Paraná. Forest dwellers, capable of escape by running, but tend to retreat to water when closely pursued. Derived from Caviidae and have the same reduction of digits. Largest living rodents, head and body length reaching 1.2 m; they may weigh over 45 kg.

Family Erethizontidae (New World porcupines)

Early Oligocene to Recent of South America, Pleistocene to Recent of North America, recent of Central America. Large, slow-moving, heavy-bodied rodents, with some hairs modified into sharp, barbed spines that are easily detached when they make contact with an enemy. Cheek teeth rooted, with a simple pattern of reentrant folds, essentially unchanged since the Oligocene. One South American genus has a prehensile tail.

Suborder Phiomorpha

A group of families of African rodents the ancestry of which can be traced back to the early Oligocene of Egypt. Most of them are hystricomorphous and hystricognathous, and they are often considered very closely related to the Caviomorpha. Multiserial incisor enamel; malleus and incus fused and sacculus urethralis present in living forms. Cheek teeth), consisting (in all but a few of the earliest types) of the last deciduous premolar, retained throughout life, and $\mathbf{M}\frac{3}{3}$.

†*Family* Phiomyidae

Oligocene to Miocene of Africa. These rodents reached Africa when it was isolated from the rest of the world and differentiated to become highly diverse, including scampering, burrowing, and perhaps arboreal and leaping forms. In the Miocene other rodents reached Africa, and only a few lines of phiomyids were able to survive the resulting competition. Several Oligocene genera retained permanent premolars, although in 1, at least, they apparently never were functional.

Family Petromuridae (rock rats or dassie rats)

Pleistocene to Recent of southern South Africa. Ratlike in body form, the single species inhabits rocky hills, living in narrow crevices in rocks.

Family Thryonomyidae (cane rats)

Miocene to Recent of Africa and early Pliocene of India. Body form and habitat generally similar to a muskrat, inhabiting marshes and the borders of streams and lakes. There is only a single living genus, spread over Africa south of the Sahara. Head and body 350 to 610 mm.

Family Bathyergidae (blesmols or African mole rats)

Miocene to Recent of Africa. The most completely fossorial of all rodents, with short, powerful limbs and heavy claws. The incisors are highly procumbent, and in some forms are used in burrowing. In some genera the growing base of the upper incisors has shifted backward to a level behind the rear of the upper cheek teeth. Only slight penetration of infraorbital foramen by masseter (probably secondary reduction); angle hystricognathous but peculiar. One genus (Heterocephalus) has the pelage reduced to scattered short hairs. External ears are greatly reduced; eyes are reduced and the eyelids usually kept closed. The number of cheek teeth reaches $\frac{6}{6}$; it is by no means clear which teeth are involved. These animals spend essentially all their lives in their burrows; some are colonial.

Families of uncertain relationships

The remaining 10 families of rodents are of completely uncertain relationships, and are best treated as individual families.

Family Hystricidae (Old World porcupines)

Pliocene to Recent of Asia, Africa, and Europe, Pleistocene to Recent of East Indies. Spined forms, resembling the New World porcupines in appearance, but less arboreal in habits; quills without barbs. The ease with which even a slight touch loosens the quills, leaving them inserted in the attacker, is the origin of the belief that these animals are capable of shooting their quills like arrows, as reported, for example, by Marco Polo. Hystricognathous and hystricomorphous; malleus and incus fused; incisor enamel multiserial; sacculus urethalis present; cheek teeth consist of P+, $\mathbf{M}\frac{3}{3}$, all exhibiting a crown pattern broken up into numerous small cuspules. The origin and relationships of the group are not clear, but they may have originated in southern Asia.

Family Castoridae (beavers)

Early Oligocene to Recent of Europe and North America, late Oligocene to Recent of Asia. Sciuromorphous and sciurognathous rodents, with cheek teeth ($P\frac{(2-1)}{1}$, $\mathbf{M}\frac{3}{3}$) progressively becoming high crowned and ever-growing. Restricted at present to a single genus ***(Castor)***,formerly widespread in Eurasia and North America, but greatly reduced as a result of very intensive trappping for the very fine, soft fur. Aquatic animals, with webbed feet and broad, flattened tail. Numerous fossils from the Tertiary do not seem to have been aquatic but rather to have been fossorial. In western Nebraska an early Miocene beaver seems to have constructed spiral burrows ("Devils' corkscrews").

†***Family Eutypomyidae***

Rodents from the Oligocene of North America sometimes (but probably erroneously thought to have been related to the beavers, but with very complexly folded enamel on the cheek teeth, and with a peculiar foot structure of uncertain function (very slender inner digits and heavy outer ones).

Family Anomaluridae (scaly-tailed "squirrels")

Miocene to Recent of Africa. Sciurognathous but hystricomorphous rodents; arboreal, and most living forms with a membrane extending from front to hind legs to tail, used in gliding from one tree to another. A long cartilaginous support, derived from the ulna in the region of the elbow, supports the anterior end of the membrane. Head and body length about 60 to 430 mm.

Family Ctenodactylidae (gundis)

Oligocene of Mongolia, Miocene of India, and Miocene to Recent of Africa. Hystricomorphous and sciurognathous rodents; malleus and incus fused; perhaps a sacculus urethalis; incisor enamel multiserial. Quite diverse in Oligocene of Mongolia, but the fossils do not suggest close relationship to any possible ancestors. Scamperers, with body form similar to that of a guinea pig; 4 toes on each foot. Head and body length 160 to 240 mm.

Family Pedetidae (springhaas or Cape jumping hare)

Miocene to Recent of Africa. Hystricomorphous and sciurognathous; malleus and incus not fused; incisor enamel multiserial. The single living genus lives in eastern and southern Africa. **A** large rodent, with elongate hind limbs, jumping like a kangaroo. The cheek teeth ($P\frac{1}{1}$, $\mathbf{M}\frac{3}{3}$) are ever-growing, with the pattern preserved only in completely unworn teeth. Head and body length about 350 to 430 mm.

Family Gliridae (dormice)

Middle Eocene to Recent of Europe, Recent of Asia, Miocene to Recent of northern Africa. Sciuromorphous and sciurognathous; uniserial incisor enamel. Small arboreal rodents, similar in appearance to the smaller squirrels. Generally have partial hibernation. It has recently been demonstrated that they can be traced back to an ancestry among small European paramyids of early to middle Eocene; as a result, although they are similar to sciurids in being sciuromorphous, they are clearly of independent origin. Head and body length about 60 to 190 mm.

Family Seleveniidae (Selevin's mice)

Probably related to dormice; a single genus recently discovered in the deserts of Central Asia. Elongate hind legs and a long tail. Head and body length 72 to 96 mm.

†***Family Pseudosciuridae***

European, middle Eocene to Oligocene. Hystricomorphous and sciurognathous; incisor enamel pauciserial. Cheek teeth low crowned. They can be derived from European early Eocene paramyids, and gave rise to the Theridomyidae. The gradual development of the hystricomorphous condition can be traced in this group.

†***Family Theridomyidae***

Abundant European late Eocene and Oligocene rodents, with low-crowned to ever-growing cheek teeth. Incisor enamel pauciserial or uniserial. They have had a central position in many theories of the origin and relationships of the various hystricomorphous groups (Caviomorpha, Phiomorpha, Hystricidae, Anomaluridae, Ctenodactylidae, and Pedetidae), but there is no evidence to show that they were actually related to any of these.

Critical appraisal. The rodents are one of the most clearly demarcated of all mammalian orders. No animals are known, either living or fossil, where there is question as to whether or not they belong to this order. Within the order, however, there is great disagreement as to the interrelationships of various families or groups of families. There is no question that the Caviomorpha, as listed above, are related, and there seems to be little question that the Phiomorpha are a natural group. There is disagreement as to whether the two suborders should be united with each other, or with the Hystricidae. If the three are united, the combination should be called the suborder Hystricomorpha (the oldest name for such a group). The Anomaluridae, Ctenodactylidae, Pedetidae, Pseudosciuridae, and Theridomyidae are included in the Hystricomorpha by some authors. The relationships of all five families are very uncertain (except that pseudosciurids clearly gave rise to theridomyids). Much uncertainty exists as to whether the families here included in the Myomorpha are a natural unit. There is considerable diversity within the group. Many authors unite the Muridae and Cricetidae into a single family; still others separate the microtines from the cricetids as a third family. The two families are recognized here largely as a matter of convenience — about 100 genera each of murids and cricetids are known; combining them makes an unwieldy family including over half the order. The microtines were quite obviously derived in late Pliocene times from cricetids; if they are recognized as a distinct family, it is the only family of either animals or plants known to have originated so recently.

Currently, attempts are still being made to find satisfactory bases on which to subdivide the order. The incompleteness of knowledge of fossil rodents is a major handicap that will certainly be overcome in the future. The classification presented here is an attempt to give what seems to be a reasonable interpretation of the present knowledge of the subject.

BIBLIOGRAPHY. J.R. ELLERMAN, ***The Families and Genera of Living Rodents,*** 2 vol. (1940, reprinted 1966), a scientific catalog of existing rodents; C.S. ELTON, ***Voles, Mice and Lemmings*** (1942, reprinted 1965), a popular, historical record of plagues of various rodents, followed by an analysis of the lemming population fluctuations in northern Labrador; P.P. GRASSE and P.L DEKEYSER, "Ordre des Rongeurs," in *Traité de zoologie*, vol. 17, pt. 2, pp. 1321–1525 (1955), a section from a zoological reference work dealing with the anatomy, ecology, habits, and classification of rodents, with brief descriptions down to the generic level (in French); T.G. HULL, ***Diseases Transmitted from Animals to Man,*** 5th ed. (1963), descriptions of diseases, with discussion of causative factors and carriers, many of which are rodents; D. MacCLINTOCK, ***Squirrels of North America*** (1970), an account of the ecology, habits, and relationships of North American squirrels; H.G.Q. ROWETT, ***The Rat as a Small Mammal,*** 2nd ed. (1965), a student laboratory manual with detailed notes on the dissection of the rat; M. SHORTEN, ***Squirrels*** (1954), a study of the red and gray squirrels in Britain for the general reader; G.D. SNELL (ed.), ***Biology of the Laboratory Mouse,*** 2nd ed. (1966), a standard reference work on all of the aspects of mouse raising; B.S. VINOGRADOV and A.I. ARGIROPULO, ***Key to Rodents*** (1968; orig. pub. in Russian, 1941), a handbook on the rodents in the U.S.S.R. and their economic importance; E.P. WALKER ***et al., Mammals of the World,*** 3 vol. (1964), descriptions at the generic level of all living mammals, including every recognized genus of rodent, usually with pictures, and information given on habits, distribution, and morphologic peculiarities; L. WILSSON, *Bäver* (1964; Eng. trans., ***My Beaver Colony,*** 1968), a popular account of beaver behaviour.

(A.E.W.)

Rodeo

A rodeo is a series of contests and exhibitions derived from riding, roping, and related skills developed by cowboys during the era of the range cattle industry in northern Mexico and the western United States.

Five standard rodeo events

Rodeo events. The five standard rodeo events are calf roping, bull riding, steer wrestling (bulldogging), saddle bronc riding and bareback bronc riding (a bronc [bronco, broncho, or bucking bronco] is an unbroken range horse picked for its resistence to training and its tendency to buck, or throw, its rider). Two other events are recognized for championships: single-steer roping and team roping. There is no ban on additional contests, and there are usually contract acts—professional specialty performances such as trick riding, fancy roping, and other exhibitions. The barrel race, a saddle horse race around a series of barrels, is a popular contest for cowgirls. Steer decorating is seen in junior contests. The chuck-wagon race is featured at the Calgary (Canada) Stampede. Prize money may be offered in a wild horse race, wild cow milking, trick and fancy riding, or a contest for cutting horses (horses trained to separate cattle from a herd).

Participants pay entry fees, and the prize money won is their only compensation. More than half of all rodeos are independent of state and county fairs, livestock shows, or other attractions, and many, large and small, are held in arenas devoted to the purpose. The equipment, however, is simple and may be improvised. Most rodeos are sponsored locally by chambers of commerce or other civic organizations, which offer prize money and employ contract acts. Among necessary employees are the clowns who, in addition to their entertainment value, have the duty of distracting animals that might injure contestants.

Origins and early shows. Rodeo events are much older than the North American cattle industry and may go back to the beginnings of the domestication of cattle and horses. Rodeo, however, developed as a peculiarly American sport and is confined mainly to Mexico, the United States, and Canada. The South American gaucho, often featured in Wild West shows, left little impress on rodeo. Australia has made only minor contributions, despite the long popularity of rodeo there, where it is often known as Bushmen's carnival. One notable importation from overseas is the Brahman bull, used in most bull-riding contests.

In the days of open range, when cattle were grazed on unfenced public lands, it was customary to round up and separate the cattle of various owners twice yearly: in spring to brand calves and in fall to select cattle for sale. The roundup brought together cowboys from many ranches. During off hours they found time for horse racing and for betting on experts in such essential skills as calf roping and bronco riding. There is record of such contests as early as 1847.

"Cowtowns" often celebrated the 4th of July with similar events. In 1869 at Deer Trail, Colorado, Emilne Gardenshire was declared champion bronco buster. Wild steer riding was staged at Cheyenne in 1872. An "Old Glory Blow-Out," sponsored by William F. ("Buffalo Bill") Cody at North Platte, Nebraska, in 1882, attracted 1,000 contestants. Its success inspired Buffalo Bill's Wild West show, which opened in Omaha in 1883. Its feature, called "Cowboy Fun," was essentially rodeo. Between 1883 and 1938, some 120 travelling Wild West shows popularized the cowboy and his skills over the United States and throughout the world.

Denver had a cowboy tournament in 1887, followed by several Mountain and Plains festivals. The Pendleton (Oregon) Round-Up started in 1910, and the Calgary (Alberta) Stampede in 1912. The latter has been an annual event since 1919. But the oldest annual show of all is Cheyenne Frontier Days, which has been presented each year since 1897.

Outstanding performers. In 1903 Bill Pickett, a Negro cowboy from Texas, leaped for the horns of a steer to save his horse from being gored and wrestled the steer to the ground, biting its upper lip in a bulldog grip. He found he could repeat the act, which became known as bulldogging—or, more politely, steer wrestling, after rodeo rules eliminated the lip biting. Pickett was hired by Joseph C., Zack T., and George L. Miller for their 101 Ranch near Ponca, Oklahoma. In 1907 they organized the Miller Bros. 101 Ranch Real Wild West show, which also employed such notables as Lucille Mulhall, called the first cowgirl and world's lady champion in roping and tying wild steers; Tom Mix, silent movie cowboy actor; and Guy Weadick, who organized the first Calgary Stampede.

Organization of the sport. In 1929 the Rodeo Association of America, an organization of rodeo managers and producers, was formed to regularize the sport. Contestants took a hand in 1936 after a strike in Boston Garden and organized the Cowboy Turtles Association—"turtles" because they had been slow to act. This group was renamed the Rodeo Cowboys Association (RCA) in 1945, and its rules now are accepted by most rodeos. Amateur rodeo has continued to grow in popularity, and the National Intercollegiate Rodeo Association, formed in 1948, has 80 member schools. Some 500 secondary school, 4-H Club, Future Farmers of America, and other junior rodeos are held annually.

Judging of events. The calf for roping and the steer for wrestling are released from chutes, an innovation since early rodeos. These are timed events. The calf must be roped and thrown, and three feet tied together. In the steer wrestling contest, a hazer helps to keep the steer moving straight forward. The wrestler must throw the steer with head and all feet in line. In calf roping, championships have been won in 16 seconds, but it has been done in well under 15 seconds. In steer wrestling, 11 seconds is championship time, but less than 10 seconds is of record.

In riding events the contestant is mounted before the chute gates are opened. The rider must stay on the animal for eight seconds. Judging, on a point system, is based on the performance of the animal as well as that of the contestant. Broncos are not trained to buck, and RCA rules ban cruelty. In all riding events the contestant is disqualified if he touches the animal or its rigging with his free hand.

All-around championships and championships in each of the five standard events are determined each year on the basis of a point-award system established by the RCA.

BIBLIOGRAPHY. R.W. HOWARD and O. ARNOLD, ***Rodeo: Last Frontier of the Old West*** (1961), general view of the sport, its several events, and history, with glossary of terms; C.P. WESTERMEIER, ***Man, Beast, Dust: The Story of Rodeo*** (1947), and (ed.), ***Trailing the Cowboy*** (1955), well-researched histories of rodeo and of the cowboy; D. RUSSELL, ***The Wild West: A History of the Wild West Shows*** (1970), relationship of rodeo to Wild West shows, circus, and other outdoor entertainment; R.D. HANESWORTH, ***Daddy of 'Em All: The Story of Cheyenne Frontier Days*** (1967), history of the oldest and most famous rodeo, listing all winners; F. CLANCY, ***My Fifty Years of Rodeo*** (1952), an interesting personal reminiscence; E.F. O'BRIEN, ***The First Bulldogger*** (1961), *a* somewhat fictionalized biography of Bill Pickett; A.S. GILLESPIE and R.H. BURNS, ***Steamboat: Symbol of Wyoming Spirit*** (1952); S. SAVITT, ***Midnight: Champion Bucking Horse*** (1957), biographies of rodeo's two most famous buckers.

(D.R.)

Rodin, Auguste

At the beginning of the 20th century, the French sculptor Auguste Rodin was famous throughout the world and long had been revered as a modern-day Michelangelo, a titan of sculpture, an incarnation of the power of inspired genius. Even his prodigious sensuality was excused as a symbol of his Olympian stature. Three-quarters of a century later, however, criticism has become less uniform, pointing to the elements in his work that belie his early life as a decorative sculptor and the concomitant lack of formal discipline. Nonetheless, he exerted an immense influence on sculpture, and his numerous students from many countries helped to spread his style. His example was particularly fruitful for such later French sculptors as Charles Despiau, Aristide Maillol, and Émile Bourdelle. Most major museums own copies of his works, and museums in Paris, Philadelphia, and Tokyo are dedicated to him. Rodin's prime contribution was in bringing Western sculpture back to what always had been its essential strength, a knowledge and sumptuous rendering of the human body. He is often considered the greatest portraitist in the history of sculpture, as if he required these commissions to harness his fertile imagination and unleash his genius. His evocations of great men, such as his "Balzac," are uniformly brilliant.

By courtesy of the Art Institute of Chicago; permission S.P.A.D.E.M. 1971 by French Reproduction Rights, Inc.

Rodin with his sculptures "Victor Hugo" and "The Thinker," gum print by Edward Steichen, 1902. In the Art Institute of Chicago.

Early life and work. Rodin was born in Paris on November 12, 1840, into a poor family. In 1854 he entered a drawing school, where he learned drawing and modelling. At 17 he attempted to enter the École des Beaux-Arts, but he failed the competitive examinations three times. The following year (1858) he was earning his living by doing decorative stonework. Traumatized by the death of his sister Marie in 1862, he considered entering the church; but in 1864 the young sculptor met Rose Beuret, a seamstress, who became his life companion, although he did not marry her until a few weeks before her death in February 1917.

Rodin had begun to work with the sculptor A.-E. Carrier-Belleuse when, in 1864, his first submission to the official Salon exhibition, “L’Homme au nez casst" ("The Man with the Broken Nose"), was rejected. His early independent work included also several portrait studies of Rose. In 1871 he went with Carrier-Belleuse to work on decorations for public monuments in Brussels. Dismissed by Carrier-Belleuse, he collaborated on the execution of decorative bronzes, and Rose joined him in Brussels.

Italian influences

In 1875, at the age of 35, Rodin had yet to develop a personally expressive style because of the pressures of the decorative work. Italy gave him the shock that stimulated his genius. He visited Genoa, Florence, Rome, Naples, and Venice before returning to Brussels. The inspiration of Michelangelo and Donatello rescued him from the academicism of his working experience. Under those influences, he molded the bronze "Le Vaincu" (“The Vanquished"), his first original work, the painful expression of a vanquished energy aspiring to rebirth. It provoked scandals in the artistic circles of Brussels and again at the Paris Salon, where it was exhibited in 1877 as “L’Âge d’ Airain" ("The Age of Bronze"). The realism of the work contrasted so greatly with the statues of Rodin’s contemporaries that he was accused of having formed its mold upon a living person.

In 1877, Rodin returned to Paris, and in 1879 his former master Carrier-Belleuse, now director of the Sèvres porcelain factory, asked him for designs. He was rejected in various competitions for monuments to be erected in London and Paris, but finally he received a commission to execute a statue for the Hôtel de Ville (City Hall) in Paris. Meanwhile, he explored his personal style in "St. Jean-Baptiste prêchant" (1878; "St. John the Baptist Preaching"). Its success and that of "The Age of Bronze" at the salons of Paris and Brussels in 1880 established his reputation as a sculptor at the age of 40.

Toward the achievement of his art. At an age when most artists already had completed a large body of work, Rodin was just beginning to affirm his personal art. He received a state commission to create a bronze door for the future Musée des Art Décoratifs, a grant that provided him with two workshops and whose advance payments made him financially secure.

That bronze door was to be the great effort of Rodin’s life. Although it was commissioned for delivery in 1884, it was left unfinished at his death in 1917. The theme of its scenes was borrowed from Dante's Divine *Comedy,* and eventually it came to be called "La Porte de l’Enfer” ("The Gates of Hell"). His original conception was similar to that of the 15th-century Italian sculptor Lorenzo Ghiberti in his "Gates of Paradise" door for the Baptistery in Florence. His plans were profoundly altered, however, by his visit to London in 1881 at the invitation of the painter Alphonse Legros. There Rodin saw the many Pre-Raphaelite paintings and drawings inspired by Dante, above all the hallucinatory works of William Blake. He transformed his plans for "The Gates" to ones that would reveal a universe of convulsed forms tormented by love, pain, and death. This unachieved monument was the framework out of which he created independent sculptural figures and groups, among them his famous "Le Penseur" (1880; "The Thinker"), originally conceived as a seated portait of Dante for the upper part of the door.

Monument at Calais

In 1884, Rodin was commissioned to create a monument for the town of Calais to commemorate the sacrifice of the burghers who gave themselves as hostages to King Edward III of England in 1347 to raise the year-long siege of the famine-ravaged city. Rodin completed work on “Les Bourgeois de Calais" (“The Burghers of Calais") within two years, but the monument was not dedicated until 1895. In 1913 a bronze casting of the Calais group was installed in the gardens of Parliament in London to commemorate the intervention of the English queen who had compelled her husband, King Edward, to show clemency to the heroes.

While the artist's glory continued to increase, his private life was troubled by the numerous liaisons into which his unbridled sensuality plunged him. In about 1885 he became the lover of one of his students, Camille Claudel, the gifted sister of the poet Paul Claudel. It proved a stormy romance beset by numerous quarrels, but it persisted until Camille's madness brought it to a finish in 1898. Their attachment was deep and was pursued throughout the country. During the years of passion Rodin executed sculptures of numerous couples in the throes of desire. The most sensuous of these groups was "Le Baiser" (1886; "The Kiss"), sometimes considered his masterpiece. Originally conceived as the figures of Paolo and Francesca for "The Gates of Hell," it exposed him to numerous scandals.

Discords and triumphs. In spite of his success, Rodin was often in conflict with L’Institut de France, the national art academy, with the public, and even with the parliament. He devoted a decade to executing four monuments honouring the landscape painter Claude Lorrain, Pres. Dorningo Sarmiento of Argentina, and the writers Victor Hugo and Honor6 de Balzac; and each of the four monuments was challenged. In Nancy, France, the Claude statue and, in Buenos Aires, the "President Sarmiento” caused riots. The conflicts over the "Victor Hugo" and the "Balzac" were even more serious.

Bronze of Victor Hugo

In 1886 he received the order for the monument to Hugo for the Panthéon, France's hall of its great men. The nudity depicted in the work caused such shock that he had to abandon the project. It was 1909 before another "Victor Hugo," also nude but seated, was installed at the gallery of the Palais-Royal, although it had been intended for the Luxembourg Gardens. In 1891, Rodin was commissioned to portray Balzac for the Société des Gens de Lettres (Society of Men of Letters). He gave himself over completely to massive research designed to translate the several Balzac portraits into sculpture. He obtained the exact measurements of the novelist's body by finding his former tailor. After much conjecture and experimentation to find an appropriate posture for the statue, he finally conceived of the writer as partly draped. The concisely designed model resembled a menhir, or upright prehistoric altar stone, foreshadowing the simplicity of modern art. The artist's delays and his design for the statue brought on a legal dispute with the Société; and, when the model was shown at the Salon de la Société Nationale des Beaux-Arts in 1898, it generated a violent debate in which the sculptor was defended by Georges Clemenceau, the future premier of France. Finally Rodin reimbursed the Société and took back the model. The statue, cast in bronze, was not erected until 1939, in the crossroads of the Montmartre section of Paris.

The Exposition Universelle of 1900 in Paris featured a pavilion in which 150 of Rodin's sculptures and numerous drawings were displayed, testifying to the international scope of his fame. After it closed, he had his works transported to a property that he had bought at Meudon in 1896. His residence there became a vast workshop where he employed a legion of assistants amid an endless stream of "favourites" who passed as his students. He was by then less a sculptor than an entrepreneur of sculpture. He himself executed only models, of which he made many, while searching for the form that suited him. Casting in bronze was the domain of specialists; but he also delegated the hewing of marble to others, to be executed under his direction but not by him. He was assisted in this "industrial" enterprise by a series of secretaries, including for a brief period the Austrian poet Rainer Maria Rilke.

International success and honors

After 1900, Rodin's worldwide success attracted abundant orders for portrait busts from the United States, Germany, Austria, England, and France. He enjoyed great renown in England, where he had numerous friends and which he often visited. In 1902 he was carried in triumph by students at a banquet in his honour in London. In 1907 he went to London for the inauguration of his monument to the poet William Henley at Westminster Abbey, and he—along with the French composer Camille Saint-Saens and the U.S. writer Mark Twain—was made a doctor *honoris causa* at Oxford University. In May 1908, King Edward VII of England visited him at his workshop in Meudon.

In the same month Rodin also rented a floor in one of the most beautiful 18th-century Parisian hotels, the Hôtel Biron, which was surrounded by an immense garden. Eventually he occupied the entire premises under an agreement by which the French state agreed to acquire and preserve the hotel as a Rodin museum in return for his donation to the state of all his works. These negotiations were endangered, however, by the self-serving intrigues of the last of his great favourites, an American who became duchess of Choiseul. They were furthered by Judith Cladel, who became his chronicler and who worked to see that the negotiations were successful, and by his last secretary, Marcelle Tirel, who defended him from the covetousness of women who tried to coax away his legacy. The purchase of the hotel and the donation of Rodin's goods was finally completed in 1916. The museum is constituted as an autonomous organization maintained by sales of castings from plaster casts that he left. Rodin died at Meudon on November 17,1917, and on the day of his burial a solemn service was celebrated in his honour at Westminster Abbey in London.

To his sculpture, Rodin added, during his lifetime, book illustrations, dry-point etchings, and innumerable drawings of nudes, principally female. He also had literary pretensions and produced several writings with the help of friends. He was enamoured of the art of the Middle Ages, and among his major efforts was the book *Les Cathédrales de France* (1914; *Cathedrals of France,* 1965).

MAJOR WORKS

SCULPTURE: All bronzes (in multiple editions) unless otherwise noted. "L'Homme au nez cassé" ("The Man with the Broken Nose"; 1864; Muste Rodin, Paris); "Young Girl with Flowers in Her Hair" (plaster, 1865–70; Musée Rodin); "Young Mother and Child" (1865–70; Muste Rodin); "The Age of Bronze" ("L'Âge d'Airain"; 1876; Minneapolis Institute of Arts, Minnesota); "St. John the Baptist Preaching" ("St. Jean-Baptiste prêchant"; 1878; Museum of Modern Art, New York); "The Gates of Hell" ("La Porte de l'Enfer"; 1880-1917; Rodin Museum, Philadelphia); "Le Penseur" ("The Thinker"; 1880; Musée Rodin); "Adam" (1880; Rodin Museum, Philadelphia); "Eve" (1881; Toledo Museum of Art, Ohio); "Le Petit Homme au nez cassé" ("The Little Man with the Broken Nose"; 1882 Muste Rodin); "The Earth" (1884; MusCe Rodin); "The Burghers of Calais" ("Les Bourgeois de Calais"; 1884–86; Joseph H. Hirshhorn Collection, New York); "The Martyr" (1885; Rodin Museum, Philadelphia); "The Kiss" ("Le Baiser"; 1886; Tate Gallery, London); "Balzac, Nude" (1893; Musée Rodin); "Balzac" (1897; Rodin Museum, Philadelphia); "Victor Hugo" (1897; Muste Rodin); "Fugit amor" (1897; Musée Rodin); "The Bather" or "Beside the Sea" (marble, 1905; Metropolitan Museum of Art, New York); "The Walking Man" (1905; MusCe Rodin); "George Bernard Shaw" (1906; Rodin Museum, Philadelphia); "Nijinsky" (1912; MusCe Rodin).

WRITINGS: A *la Vénus de Milo* (1910; *To the Venus de Milo,* 1912); *L'Art* (1911 and 1946; *On Art and Artists,* 1957), conversations collected by Paul Gsell; *Les Cathédrales de France* (1914; *Cathedrals of France,* 1965).

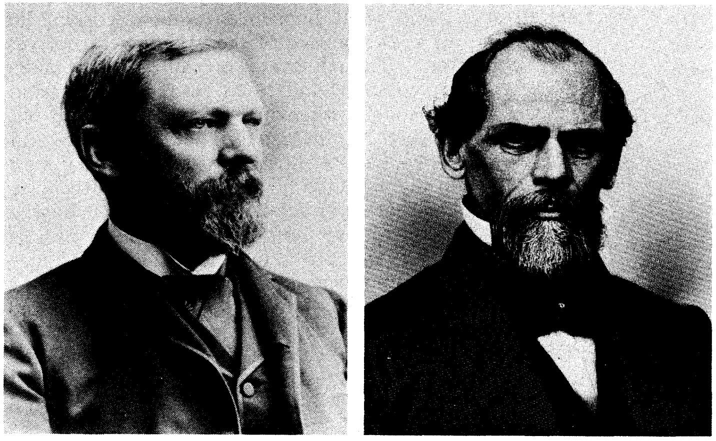
BIBLIOGRAPHY. ROBERT DESCHARNES and J.F. CHABRUN, *Auguste Rodin* (1967), is the most complete monograph on the life and work of the artist. Other important works are: C.J. BURCKHARDT, *Rodin und das plastische problem* (1921), a morphological study; JUDITH CLADEL, *Auguste Rodin: l'oeuvre et l'homme* (1908; Eng. trans., *Rodin: The Man and His Art,* 1917); *Rodin sa vie glorieuse, sa vie inconnue* (1936; Eng. trans., *Rodin*, 1938), studies by a direct witness; H. CHARLES-ESTIENNE DUJARDIN-BEAUMETZ, *Entretiens avec Rodin* (1913); E. HERRIOT, A *la gloire de Rodin* (1927), an expression of admiration from a great man of politics; G. GRAPPE, *Rodin* (1955), by the conservator of the MusCe Rodin in Paris; R. MARIA RILKE, who was the artist's secretary for a short time, *Auguste Rodin* (1902; 2nd ed., 1913; Eng. trans., 1945); and *Lettres à Rodin* (1928–34); DENYS SUTTON, *Triumphant Satyr: The World of Auguste Rodin* (1966), an aesthetic and biographical study; MARCELLE TIREL, *Rodin intime* (1923; Eng. trans., *The Last Years of Rodin*, 1925), by Rodin's secretary; and L. WEINBERG, *The Art of Rodin* (1918).

(G.R.M.B.)

Roebling, John Augustus and Washington Augustus

John Augustus Roebling, a pioneer in the design of steel suspension bridges, whose best known work is the Brooklyn Bridge, was among the first to recognize the resilience of steel wire and to weave it into massive cables of great strength and flexibility.

By courtesy of (left) Rutgers University Library, Nevi Brunswick, New Jersey, (right) the Smithsonian institution. Washington. D.C.



(Left) Washington Roebling, 1870. (Right) John Roebling, engraving by William Murray.

He was born on June 12, 1806, in Miihlhausen, Prussia. After graduating from the polytechnic school in Berlin, he worked for the Prussian government for three years and at the age of 25 emigrated to America. He settled with others from his hometown in a small colony that was later called Saxonburg, near Pittsburgh, in the hills of western Pennsylvania. He married the daughter of another Miihlhausen emigrant, and they had nine children—the eldest of whom, born on May 26, 1837, was Washington Augustus Roebling. After a few years of unsuccessful farming, John Roebling went to the state capital in Harrisburg and applied for employment as a civil engineer.

Development of method of stranding cables

He had often watched canalboats being hauled over hills from one watershed to another, and he persuaded the canal commissioners to let him replace the hempen hawsers with wire cables. He developed his own method for stranding and weaving wire cables, which proved to be as strong and durable as he had predicted. The demand for such cable soon became so great that he established a factory to manufacture it in Trenton, New Jersey. This was the beginning of an industrial complex that finally was capable of producing everything from chicken wire to enormous 36-inch cables. It remained a family-owned business, carried on by three generations of Roeblings.

Roebling was less a businessman than an engineer, and with the growth of his reputation as a designer and builder of long-span suspension bridges he spent less and less time at the Trenton factory. His eldest son Washington, after graduating from Rensselaer Polytechnic Institute, joined him in his work. In the 1850s and 1860s the Roeblings built four suspension bridges: two at Pittsburgh, one at Niagara Falls, and another across the Ohio between Cincinnati and Covington, Kentucky, with a main span of 1,051 feet (320 metres). New York state accepted Roebling's design for a bridge connecting Brooklyn and Manhattan with a span of 1,595 feet (486 metres) and appointed him chief engineer. He sent his son Washington, who had been in charge of the construction of the enormous masonry towers that supported the Cincinnati–Covington cables, to Europe to study new methods for the sinking of the foundations on which the granite towers of the Brooklyn Bridge were to stand.

The building of Brooklyn Bridge

According to legend, "every bridge demands a life." In this instance the life was that of the designer himself. Roebling was taking final compass readings while standing on some pilings at a ferry slip and did not notice that a boat was docking. As it banged into the slip, one of his feet was caught between the pilings. He was rushed to his son's house in Brooklyn Heights, where the doctors amputated his injured toes. Three weeks later, he died of tetanus at the age of 63.

Washington Augustus Roebling had been a colonel in the Union Army, and the title remained with him the rest of his life. On his father's death he was asked to serve as chief engineer and immediately began work on the foundations for the two towers. The use of pneumatic caissons (watertight chambers) was still in a somewhat experimental stage, and what happened to men working in compressed air at the bottom of the caisson was not yet fully understood. Though every precaution was taken there were more than a hundred cases of decompression sickness (the "bends") when men were brought up too rapidly. There were also the usual difficulties of fires and breakdowns, as well as so-called blow-outs that shot mud and water into the air. Like his father, Colonel Roebling felt he had to inspect every detail of the work. One day he remained 12 consecutive hours in the compressed-air chamber, finally being carried out unconscious. In those days they did not understand the amount of time needed for slow decompression. The nitrogen bubbles in the bloodstream could paralyze a man for life.

His health was permanently affected, and though he lived to be almost 89, the Brooklyn Bridge was his last undertaking. From his bedroom window he watched the cables being spun, and his wife carried orders to the engineers and foremen. The bridge took 13 years to complete. On a spring day in May 1883, Pres. Chester A. Arthur and his Cabinet, along with the Governor of New York and a host of distinguished visitors, marched across the East River from Manhattan to Brooklyn. The design of the bridge has been an inspiration to painters and poets, and after 90 years it is still one of New York's important traffic arteries, carrying thousands of vehicles a day.

BIBLIOGRAPHY. The definitive biography of the Roeblings is HAMILTON SCHUYLER, ***The Roeblings: A Century of Engineers, Bridge-Builders and Industrialists*** **(1931).** See also WILLIAM C. CONANT and MONTGOMERY SCHUYLER, ***The Brooklyn Bridge—with an Account of rhe Opening Exercises (1883);*** and D.B. STEINMAN, ***The Builders of the Bridge: The Story of John Roebling and His Son*** **(1945).**

(H.H.B.)

Roger II of Sicily

Creator of the Norman Kingdom of Sicily, Roger II ruled one of the best-governed states in 12th century Europe, where Italians, Greeks, Muslims, and Jews lived together in a cosmopolitan society. He was the son of the Great Count Roger I of Sicily and his third wife, Adelaide of Savona. Born on December 22, 1095, he succeeded his elder brother Simon on September 28, 1105, at the age of nine. Little is known of his childhood; the tradition that he was baptized by St. Bruno, founder of the Carthusian Order, is undocumented. These years, during which his mother acted as regent, he probably spent between Mileto in Calabria, the family castle in northeast Sicily, and Messina; but it was at Palermo in 1112 that he was knighted and assumed the reins of government, and there his Sicilian capital was henceforth established.

Alinari

Roger II, mosaic depicting his coronation by Christ, 12th century. In the Church of Martorana, Palerrno, Sicily.

Character and personality

Though the island that Roger I and his brother Robert Guiscard had conquered was populated predominantly by Arabs—with a strong admixture of Greeks—the Great Count had always remained essentially a Norman knight. His son, by contrast, was a man of the Mediterranean. Deprived of paternal influence from the age of five, brought up in a cosmopolitan, multiconfessional world of Greek and Muslim tutors and secretaries, of studies pursued and state affairs conducted in four languages, Roger soon revealed an exotic strain in his nature that cannot wholly be ascribed to his mother's Italian blood. The latter was obvious enough in his complexion and in the darkness of his eyes and hair, but his contemporaries soon learned to their cost that he was not only a southerner—he was also an Oriental. He was a ruler for whom diplomacy, however tortuous, was a more natural weapon than the sword, and gold, however corrupting, a more effective currency than blood.

Two qualities, however, he had inherited from his Norman forebears: his energy and his ambition. It was these, combined with a gift for imaginative statesmanship all his own, that enabled him to profit from the fecklessness of his cousins—the son and grandson of Robert Guiscard—and to acquire, in return for military aid against a rebellious baronage, more and more of their mainland territories. By 1122 all Calabria was his; and in 1127, when

Duke William of Apulia died without issue, Roger laid claim to the duchy as his rightful heir. Opposition was considerable; the barons had always resented the domination of the Hautevilles, whom they looked upon as upstarts no better than themselves, and the papacy had no wish to see too powerful a state established on its southem frontier. But they were no match for Roger's particular technique of armed diplomacy, and in 1128 Pope Honorius II invested Roger as duke of Apulia, Calabria, and Sicily.

Thus, at 32, the young duke found himself one of the most influential princes in Europe. Only one thing more was necessary before he could weld his triple duchy into a single nation and treat with his fellow rulers on equal terms: a royal crown. Two years later he got it. Honorius' death early in 1130 led to a dispute over the papal succession. One of the two candidates, Innocent II, thanks to the energetic advocacy of St. Bernard of Clairvaux, soon had almost the whole continent behind him. His rival, the antipope Anacletus II, turned to Roger, who promised full support in return for coronation.

Enthronement as King of Sicily

The first king of Sicily was crowned on Christmas Day 1130 in the cathedral at Palermo. The antipope Anacletus died in 1138 and in the following year, after routing a papal army at Galluccio and taking the Pope captive, Roger forced Innocent to confirm him in the kingdom of Sicily, with the overlordship of all Italy south of the Garigliano. After this he was quickly able to pacify his mainland realm, where his vassals—abetted by the German emperor Lothair II who led a large, though unsuccessful, expedition to South Italy in 1136–37—had kept up an almost permanent insurrection. In Sicily itself, where the ban on large fiefs had left little opposition to Roger's rule, the new kingdom steadily grew more prosperous. The King himself, more than any other ruler of his day, was an intellectual who had thought deeply about the science of government, and although he cherished no love for the empire of the East—which, like that of the West, maintained its claim to its former South Italian possessions—his whole upbringing inclined him toward the Byzantine concept of monarchy: a mystically tinged absolutism in which the sovereign, as God's viceroy, lived remote and elevated from his subjects in a magnificence that reflected his intermediate position between Earth and heaven. It is no coincidence that in one of the only two portraits of Roger with any claim to authenticity—the mosaic in the Church of the Martorana at Palermo—he is depicted in Byzantine robes being symbolically crowned by Christ.

But splendour did not mean empty extravagance. A contemporary chronicler notes that Roger would personally go through his exchequer accounts, recording even the smallest expenditure, and that he was as scrupulous in the payment of debts as in their collection. Still less did it mean idleness. In the words of his court geographer, the King "accomplished more in his sleep than others did in their waking day." Building on the foundations his father had laid, he created a civil service, based eclectically on Norman, Greek, and Arabic models, that was the wonder and envy of Europe. He entrusted finance to his Arab subjects, who also supplied him with the spearhead of his army. The navy, by contrast, was predominantly Greek; its chief, known by the Arabic title emir of emirs—from which the word admiral derives—served also as head of the government, ranking second after the king himself.

Roger's navy

It was on this navy above all that Sicily's security and prosperity depended, and Roger's use of it was not overscrupulous. Under the greatest of its admirals, George of Antioch, it subdued much of what is now Tunisia to form a profitable, if short-lived, North African empire; it captured Corfu; it harassed the Greek coast, abducting the best of the Theban silk workers to found the court workshop at Palermo; and in 1149 it sailed up the Bosporus to fire a few impudent arrows into the gardens of the imperial palace. Significantly, however, it played no part in the Second Crusade of 1147. Roger had hated the Frankish rulers of Jerusalem ever since his mother's disastrous remamage to King Baldwin I of Jerusalem 34 years earlier. Besides, most of his Sicilian subjects were Muslims, and toleration was the cornerstone of his kingdom.

This policy even showed itself in his church buildings. Roger's first great building, the cathedral at Cefalu, shows little Saracenic influence; but the Palatine Chapel in Palermo, conceived on a Latin plan and aglow with Byzantine mosaics, is topped by a stalactite roof of pure Arab workmanship. Oriental inspiration is equally evident in the five vermilion cupolas of the Church of S. Giovanni degli Eremiti, built in 1142 for the Benedictines.

The Assizes of Ariano

After the pacification of South Italy, the King promulgated in 1140 at the so-called Assizes of Ariano a corpus of law covering every aspect of his rule. He then returned to Palermo, which he seldom left again. There he spent his last 15 years in the most intellectual court of Europe, surrounded by the leading thinkers of the time. Sicily was already the only land where scholars could study both Greek and Arabic—then the scientific language par *excellence.* Through Roger's enthusiasm, Sicily became a cultural clearinghouse where, for the first time, Western and Oriental scholars could meet on an equal footing.

Roger II was mamed three times. He outlived his first wife, Elvira, daughter of Alfonso VI of Castile, and his second, Sibyl of Burgundy. His third wife, Beatrice of Rethel, whom he mamed in his last year, bore him a daughter, Constance, after his death. Constance married the future emperor Henry VI, bringing Sicily under the control of the Hohenstaufens. Roger died, aged 58, on February 26, 1154, and was succeeded by his fourth but oldest surviving son, William. Despite his repeatedly expressed wish to rest in Cefalu, the King was buried in the cathedral at Palermo, having created, in a Europe rent by schism and exhausted by the Crusades, not just a kingdom but a political and religious climate in which all races, creeds, and cultures were equally encouraged and equally favoured.

BIBLIOGRAPHY. A full biography of Roger II in English is JOHN JULIUS NORWICH, *The Normans in the South* (U.S. title, *The Other Conquest,* 1967), which takes the story of the Norman conquest of South Italy and Sicily up to Roger's coronation in 1130; and its sequel, *The Kingdom in the Sun 1130–1194* (1970). Both volumes contain comprehensive bibliographies. E. CURTIS, *Roger of Sicily and the Normans in Lower Italy, 1106–1154* (1912), is inaccurate in several important respects. See also F. CHALANWN, *Histoire de la Domination normande en Italie et en Sicile,* 2 vol. (1907, reprinted 1960); and E.L.E. CASPAR, *Roger II. (1101–1154) und die Gründung der normannisch-sicilischen Monarchie* (1904).

(N.)

Roman Catholicism

Roman Catholicism, a Christian church characterized by its uniform, highly developed doctrinal and organizational structure, traces its history to the college of Apostles in the 1st century AD. As with Eastern Orthodoxy and Protestantism, it is a major branch of Christianity.

This article is divided into the following sections:

VI. Practices
Missions, education, and eleemosynary activities
Sociopolitical views and practices
VII. Roman Catholicism following the second Vatican Council

Other articles that deal with the nature and history of Roman Catholicism are CHRISTIANITY; CHRISTIANITY BEFORE THE SCHISM OF 1054; ROMAN CATHOLICISM, HISTORY OF; EASTERN CHRISTIANITY, INDEPENDENT CHURCHES OF; PROTESTANTISM, HISTORY OF; EASTERN RITE CHURCHES; PAPACY; PROTESTANTISM; EASTERN ORTHODOXY; and ECUMENISM.

I. Nature and significance

The name Roman Catholicism designates that Christian church that has the largest number of communicants and the oldest single identity. "Catholicism" is from the Greek word *katholikos,* meaning "universal," and can be traced to the early period of Christianity, when there was only one Christian church. This early unity concealed various diversities that led to ecclesiastical schisms from the 4th century on, but the largest church retained the designation *katholikos.* Although the word Roman appears to make the name self-contradictory, Roman Catholics believe that this contradiction is only apparent and not real; they continue the identification of the *katholikē* with the church that acknowledges the primacy of the bishop of Rome, the pope.

HISTORICAL AND CULTURAL IMPORTANCE

Influence on the development of Europe

Roman Catholicism has been one of the major factors in the historical and cultural development of western Europe and in the extension of European culture to other continents beginning in the 15th century. This statement is quite independent of any evaluation of the influence of Roman Catholicism, an evaluation that has long been hotly disputed between apologists of Roman Catholicism and its critics and that has to be a mixed judgment if it is to be at all fair. The effects of Roman Catholicism, not always evident on the surface in modern times, can be seen in the political theory and structure of Europe, the mind-set of European thought (as opposed to Muslim, Indian, and Chinese, for example), European education, the Roman Catholic patronage of literature and the arts, and the very theory of the universality of European culture. For many historians, the worldwide expansion of European culture as the one culture for all is a secular reflection of the Roman Catholic belief in one church for all.

On the other hand, Roman Catholicism has received—and taken—very little from any other history and culture. Its Romanism has in modern times taken away something from its Europeanism; but it is the single most "European" religious and cultural phenomenon in the world today, even if much of its Europeanism is archaic. The history of Roman Catholicism since the colonial expansion of Europe has been normally a refusal to adapt itself to new cultures, which were evaluated by the churchmen at much the same level as pagan religions.

Although the Roman Catholic Church by its original documents and traditions is not a political body or agent, in fact it has been deeply involved in the politics of Europe. When the Papal States existed (754–1870), these were as genuinely political as the kingdoms of France and Spain. Theoretically, this should not have involved the Roman Catholic Church, for, even though it was the church of all Europe for most of this period, its members were not subject to the sovereignty of the Roman pontiff as head of the Papal States. Practically, however, the popes were unable to draw a real distinction between their spiritual sovereignty and their temporal sovereignty, and without scruple, it often seemed, they used the Papal States to support the authority of the church or—more frequently—the church to support the power of the Papal States. Many of the manoeuvres can now be seen to have been abuses of power, and many churchmen saw them as abuses of power when they happened.

This has meant that the structure of Roman Catholic power over the centuries took on the features of the political sovereignty normal in the Middle Ages and in early modern times. It also meant that the political sovereignties of Europe assumed some of the features of the spiritual authority of the church. The divine right of kings is the most obvious example of this; it survives to this day in the idea of absolute sovereignty of the nation, which is accountable neither to God nor its citizens nor another nation.

CHARACTERISTICS AND MEMBERSHIP FIGURES

Centralization and uniformity

Among the Christian churches Roman Catholicism has been characterized by its rigorously centralized structure; its uniformity of doctrine; its uniformity of ritual; its dependence on tradition and precedent, indeed a tendency to create or to fortify tradition when there is a weak connection between existing belief or practice and the past; a complex gradation of hierarchy and clergy; and a clear division between clerics and laymen. These characteristics are truly Roman, both in the sense that they developed within the genius of the historic Roman Catholic Church and were not borrowed from other religious bodies and also in the sense that they distinguish the Roman Catholic Church from other Christian churches. It is, for example, opposed to the Eastern Orthodox churches in its centralization (and its understanding and use) of authority, its uniformity of doctrine and ritual, and its esteem of tradition and precedent. The Orthodox and Protestant churches find Rome rigid and authoritarian; Rome finds these churches too loosely structured and vacillating.

Exact figures for the number of Roman Catholics in the world are not available. Estimates of 1980 are listed in the table.

Roman Catholic Populations	
Europe	251,256,000
United States, Canada	78,647,000
Latin America	328,896,000
Asia, Australia, Oceania	73,569,000
Africa	76,789,000
World total	809,157,000

These figures show that the strength of the Roman Catholic Church lies in Europe; being merely for continents, they do not show that the church's strength outside Europe lies in those countries that are most Europeanized, especially in the Americas. Roman Catholicism was not strong in Russia under the tsars, and it has lost no ground under Soviet Socialism. Its success in the "missionary continents" of Asia, Africa, and the Americas results largely from the fact that the earliest colonization of these continents was carried on by the Catholic powers: Spain, Portugal, and France. It must also be said in all candour that Roman Catholicism traditionally exhibited a missionary zeal that the Protestant churches had to learn from the Roman Church.

II. History

THE EARLY CHURCH (TO AD 313)

Roman Catholicism claims continuity with the church of the New Testament. What church historians call "early Catholicism" appears in some of the later books of the New Testament; these include the Pastoral Letters (I and II Timothy and Titus) and the letters of Peter. Early Catholicism means the first appearance of characteristic church structures and a normative scheme of belief. The major elements of what appeared later as Catholic structure and Catholic belief cannot be clearly perceived before the 2nd century. Historical sources up to AD 200 are not extensive. The Christian Church of the early period was an underground movement, whatever its numbers may have been, in the sense that by far the majority of its membership was drawn from the lower classes, including slaves. It attracted very few of the ruling classes, the wealthy, or the intellectuals. The government of the Roman Empire tolerated Christianity more frequently than not, although it did not have legal standing (such as Judaism possessed). The traditional "persecutions" of the Christians were isolated local incidents, except for the persecutions by the emperors Decius, Valerian, and Diocletian in the 3rd century. It appears that the numbers of Christians had by this time grown sufficiently to alarm the government, and these three emperors and their officers seemed to recognize that Christianity would destroy the Roman Empire if it destroyed paganism. They were correct in their evaluation of the Christian ethos, but they did not foresee that Christianity would reach a compromise with the empire, that it would become "Roman."

During this period the collection of the New Testament was formed. Theological literature made its appearance in the apologists, who intended to present Christianity as a belief acceptable within the Roman policy and society, and in the writings of the first theologians, most of whom were bishops; these writings were elicited by doctrinal controversies. Doctrine became sufficiently stabilized for a canon of belief to arise that was accepted throughout the entire Christian world; communications between the churches of different regions were remarkably close and frequent. Questions of doctrine and discipline were often dealt with by meetings of the bishops of a particular region. The most important theological event of the period was the establishment of the school of Alexandria, illuminated by Clement and Origen, which initiated a celebrated school of Platonists.

THE MIDDLE AGES (313–1517)

Christianity as the state religion

The Edict of Milan, a cluster of documents issued by Constantine the Great in 313, made Christianity a lawful religion; by the end of the 4th century it had become the state religion. Historians call the church of this period the Constantinian church, designating the privileged position of the church within the state. This position is characteristic of the entire Middle Ages. It did not mean that there were no quarrels between church and state; the privileged position appears in the fact that the church could contend with the state on equal terms and sometimes win the conflict.

During this period Roman Catholicism developed the hierarchical structure that endured largely unchallenged until the Reformation. The fall of the Roman Empire and the barbarian conquests meant that the Roman Church was the cultural link between the old and the new; and the Europe that arose out of the barbarian conquests was a community identified with Roman Catholicism. This identity was sharpened by the Muslim conquests after the 7th century, which forced Christian Europe to affirm its identity against the Muslim peril; and it was sharpened further by the schism of the Greek churches, which became permanent in 1054. By the 10th century the religious and cultural community that is called Christendom had come into being. In every European state the religion of the state was Roman Catholicism. Christendom fought back against Islām in the Crusades, which failed to repossess the lost territories but strengthened the unity of Christendom and rendered it conscious of its power.

The Middle Ages saw the rise of the universities and of a "Catholic" learning, sparked, oddly enough, by the transmission of Aristotle through Arab scholars. Scholasticism, the highly formalized philosophical and theological systems developed by the medieval masters, dominated Roman Catholic thought into the 20th century and contributed to the formation of the European mind-set. With the rise of the universities, the threefold level of the ruling classes of Christendom was established: *imperium* (political authority), *sacerdotium* (ecclesiastical authority), and ***studium*** (intellectual authority). The principle that each of these three was independent of the other two within its sphere of authority had enduring consequences in Europe.

Rise of monasticism

The same period saw the growth of monasticism. One may see in this withdrawal from the world a response to the essential conflict between Christianity and Roman civilization; those who refused to accept the prevailing compromise between the religious and secular spheres could find no place in the world of the early Middle Ages. Perhaps the most remarkable feature of monasticism was that this withdrawal did not take the form of heresy or schism. Monasticism found a way of refusing the compromise without departing from the church that had made the compromise.

The early Middle Ages, especially from the 4th to the 7th century, were the ages of the great ecumenical, or general, councils. It was no doubt the new "Roman" character of the Constantinian church that led Christians to look for an expression of the voice of the whole church, the *katholikē*, that would be the voice of the whole world, the *oikoumenē*, just as the emperor was the voice of secular authority for the whole world.

This period also revealed the possibilities of corruption within the Roman Catholic Church. The moral collapse of the papacy in the 10th and 15th centuries, a similar collapse of much of the episcopacy during the 13th to the 16th century, and even the collapse of religious orders during the 15th century demonstrated the effects of the compromise between the church and the world. By seeking wealth and power the church lost much of its moral authority to proclaim even a diluted version of the gospel. This, more than anything else, led to the religious revolution that concluded the Middle Ages of the Roman Catholic Church and initiated modern times.

THE REFORMATION TO THE FIRST VATICAN COUNCIL (1517–1870)

The Protestant Reformation is conventionally dated from Martin Luther's publishing of his Ninety-five Theses in Wittenberg on October 31, 1517. By the time Luther died in 1546, much of Europe had renounced allegiance to the Roman pontiff; France had been internally rent by religious strife; and the seeds had been planted that came to maturity in the horrible religious struggle of the Thirty Years' War (1618–48). The Reformation must still be viewed as the greatest revolution within the Christian world community. It attacked just those institutions of Christendom that have been mentioned as established during the Middle Ages and destroyed or profoundly altered them, even in countries that remained Roman Catholic. Although there were doctrinal issues that separated Rome and the Reformers, these differences became greatly exaggerated. More importantly, there was no longer a unified Christendom, and Europe was no longer identified with the Roman Catholic Church. The political power of the papacy shrank to minimal dimensions with the loss of religious power. Moreover, the Roman Catholic Church lost its cultural leadership: the new learning—in particular, the beginnings of modern science—proceeded largely without the blessing of the Roman Catholic Church. The educational system of Europe was often Protestantized or secularized. The old Catholic world view gave way before new systems in philosophy and science. The rise of the nation states was both helped by Protestantism and a help to Protestantism; the result was a theory and practice of politics that was secular in the sense that religion, and in particular the Roman Catholic Church, was a negligible factor.

The Council of Trent

The Roman Catholic Church rallied to meet what appeared to some observers to be a fatal threat by the movement called the Counter-Reformation, which took form in the Council of Trent (1545–63). The council set firm lines of doctrine against Reformed theology and took firm action against the more notable forms of moral corruption that had lent credibility to the charges of the Reformers. It closed the ranks of discipline by centralizing more responsibility in the papacy. Unfortunately, the Council of Trent occurred during a period of deep and rapid change. It responded to this change by doing the best it could to crystallize the past; and, although there is no doubt that it was the major factor in preserving Roman Catholicism from disintegration, it did this by making the Catholic Church almost impervious to change in the centuries that followed.

Although the Roman Catholic Church lost many members in Europe during the 16th century, the discoveries of new lands, mostly under the leadership of stoutly Catholic Spain and Portugal, gave the church new members to offset these losses; and Roman Catholicism became a force in colonialism. The 17th and 18th centuries, however, were a period of decline in Roman Catholicism in almost every area of activity. In spite of this growing archaism, the Roman Catholic Church survived the French Revolution, which again many observers thought would be its death blow. But survival is the best word; the 19th century was another period of decline, one in which the church adopted a firm position against almost every intellectual, social, and political development that is called modern. Pius IX, whose long pontificate (1846–78) enabled him to dominate Roman Catholicism during the 19th century, saw the strength of the Roman Catholic Church in the supremacy of the Roman pontiff. In line with the Pope's view, the first Vatican Council (1869–70) defined both the primacy and the infallibility (assurance against error when teaching the whole church on matters of faith or morals) of the Roman pontiff only three months before Victor Emmanuel II destroyed the Papal States and rendered the papacy politically impotent.

FROM THE FIRST VATICAN COUNCIL TO THE PRESENT

The 100 years following the first Vatican Council was a period of frequent attacks on the Roman Catholic Church by anticlerical governments in Europe, which attempted to reduce Roman Catholicism to utter impotence. The losses in Europe were balanced by the fantastic growth of Roman Catholicism in the United States (the result of European immigration), and North America became the most prosperous area of the Catholic Church. Much of the first half of the 20th century witnessed the struggles of Roman Catholicism against Fascism in Italy and National Socialism in Germany and the adoption of a firm position against Communism, based not so much on the Socialism of Communism as on its explicit atheism. Within the church the struggle against Modernist theology seemed to be ended with its condemnation by Pius X (1907). In fact, the Modernist attempt to update Roman Catholic theological thought did not end and reached fruition in the sessions of the second Vatican Council (1962–65). This council reversed the Council of Trent to the extent that it abandoned the attempt to crystallize the past; but in opening Roman Catholicism to the present the council created much uncertainty among Roman Catholic communicants. In the early 1970s it was.asserted by many that the second Vatican Council had initiated a revolution within Roman Catholicism perhaps as profound as the Protestant Reformation, yet without the same shattering effect.

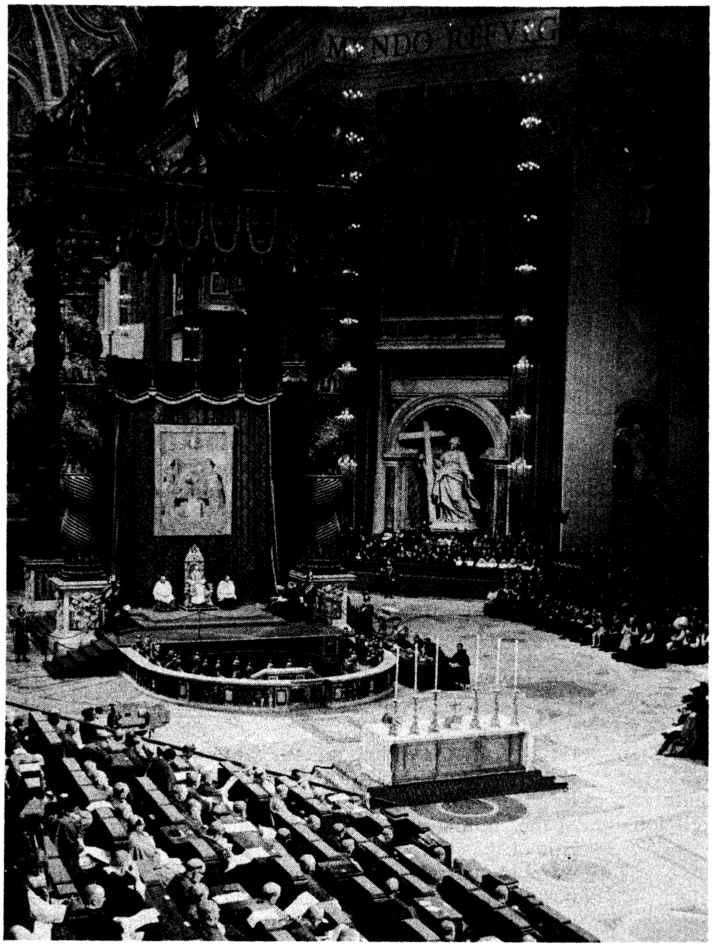
The second Vatican Council

III. Nature and structure of the church in Roman Catholic teaching

DOCTRINAL BASIS OF THE CHURCH STRUCTURE

The nature of the church. In 1965 M.-J. le Guillou, a Roman Catholic theologian, defined the church in these terms: "The Church is recognized as a society of fellowship with God, the sacrament of salvation, the people of God established as the body of Christ and the temple of the Holy Spirit." The progress of Roman Catholic theology can be seen in the contrast between this statement and the definition still current as late as 1960, substantially the definition formulated by Robert Bellarmine, a Jesuit controversialist, in 1621: "the society of Christian believers united in the profession of the one Christian faith and the participation in the one sacramental system under the government of the Roman Pontiff." The older definition, created in response to the Protestant claims, defines the church in external and juridical terms. The more recent definition is an attempt to describe the church in terms of its inner and spiritual reality.

From the earliest heresies, the church has thought of itself as the one and only worshipping group that traced itself back to the group established by Jesus Christ. Those who withdrew from this group were religiously no different from those who had never belonged to it. The ancient adage, "There is no salvation outside the church," was understood as applying to membership in this group. When this adage was combined with the notions contained in Bellarmine's definition of the church, lines were clearly drawn. These lines were maintained in the breakup of Western Christendom in the Reformation.



General assembly of church fathers with Pope Paul VI presiding during the second session of the second Vatican Council, December 1963.
NC Photos/KNA

There were, however, other factors determining the idea of the one true church. The Roman Catholic Church had never excluded the Orthodox Church, which had seceded from the Roman Church in 1054, from the community of Christian believers. Furthermore, the juridical definition of the church did not include such traditional themes as the communion of the saints and the body of Christ. Both of these themes look beyond the visible, juridically constituted church. The communion of saints views the church as a whole that includes both the living (the church militant) and the dead (the church suffering in purgatory—a state for those who must be cleansed from lesser sins—and the church triumphant in heaven). The idea of "communion" appears in early church literature to indicate the mutual recognition of union in the one church and the notion of mutual services.

The mystical body of Christ

The theme of the body of Christ appears in the letters of Paul (Rom. 12; I Cor. 12; Eph. 4 and 5; Col. 1). In modern Roman Catholic theology the term mystical has been added to "body," doubtless with the intention of distinguishing the church as body from the juridical society. Pius XII, in the encyclical *Mystici Corporis* (1943; "The Mystical Body"), identified the mystical body with the Roman Catholic Church. Most Roman Catholic theologians and the second Vatican Council have taken a less rigorous view, trying to find some way of affirming membership in the body for those who are not members of the Roman Catholic Church. The documents of the council described the church as the "People of God" and as a "Pilgrim Church," but no generally accepted statement of membership in this church has yet emerged. The second Vatican Council also departed from established Roman Catholic theology since the Reformation by using the word church in connection with the Protestant churches. This use has caused some confusion, but the trend is now rather to think of one church divided than of one true church and other false churches.

Apostolic succession. The claim of the Roman Catholic Church to be the one legitimate continuation of the community established by Jesus Christ is based on apostolic succession. This does not mean that there are apostles, nor does it mean that individual Apostles transmitted some or all of their commission to others. The officers of the church, the bishops, are a college that continues the college of the Apostles, and the individual bishop is a successor of the Apostles only through his membership in the college.

The idea of apostolic succession appears in the writings of Irenaeus, a Church Father who died about 202. Against the Gnostics (dualistic sects that maintained that salvation is not from faith but from some esoteric knowledge) Irenaeus urged that the Catholic teaching was verified because a continuous succession of teachers, beginning with the Apostles, could be demonstrated. In the 3rd and 4th centuries, problems of schism within churches were resolved by appealing to the power of orders (*i.e.*, the powers a person has by reason of his ordination either as deacon, priest, or bishop) transmitted by the imposition of hands through a chain from the Apostles. Orders in turn empowered the subject to receive the power of jurisdiction (*i.e.*, the powers an ordained person has by reason of his office). In disputes between Rome and the Eastern churches, the idea of apostolic succession was centred in the Roman pontiff, the successor of Peter; it will be observed that this goes beyond the idea of collegial succession. Apostolic authority is defined as the power to teach, to administer the sacraments, and to rule the church. Apostolic succession in the Roman Catholic understanding is validated only by the recognition of the Roman pontiff; and the Roman Catholic Church understands the designation "apostolic" in the creed as referring to this threefold power under the primacy of the Roman pontiff.

Apostolic succession in other churches

The Roman Catholic Church has not entirely denied apostolic succession to non-Roman churches. Rome recognizes the validity of orders in the Orthodox churches; this means that it recognizes the sacramental power of the priesthood but does not recognize the government of these churches as legitimate. The orders of the Anglican and the Swedish Lutheran churches, on the contrary, are not recognized by Rome, and the entire threefold quality of apostolic succession is denied them. Oriental churches in union with Rome (Eastern Catholics) are recognized as in full apostolic succession. Luther and Calvin saw clearly that their position could not be maintained if apostolic succession were necessary; they therefore affirmed that apostolic succession had been lost in the Roman Church by doctrinal and moral corruption and that the true church was found only where the gospel was rightly preached and the sacraments were rightly administered. Thus, Protestant churches generally have not accepted the necessity of apostolic succession.

STRUCTURE OF THE CHURCH

The papacy. In Roman Catholic belief the Roman pontiff is the successor of Peter, upon whom Jesus conferred the primacy of the college of the Apostles. This belief is based on some rather involved biblical arguments (for details see PAPACY). There is no clear assertion of this belief before the conversion of Constantine in 312. The fullness of Roman Catholic teaching on the papacy was defined as dogma in the first Vatican Council; it was stated that the pope enjoys absolute supreme jurisdiction (meaning legislative and judicial power) within the church. He has the same absolute supreme power to declare Catholic doctrine, and he enjoys, even when he speaks alone as the head of the church, the same infallibility that the church as a whole enjoys. Most of these claims did not depart radically from the traditional teaching of the Roman Catholic Church, although many at the time thought they were excessive or that the statement was imprudent. The declaration, however, did not include a statement on the jurisdiction and teaching power of the bishops, and this meant that for almost 100 years Roman Catholicism lived with what was recognized as a one-sided statement of papal power, completed only in the second Vatican Council.

The papacy is an institution as well as a person, and the term Holy See officially includes the Roman offices (the Roman Curia) through which the pope governs the Roman Catholic Church.

The College of Cardinals. Cardinals are designated as cardinal bishops, cardinal priests, or cardinal deacons; these designations have nothing to do with the orders they possess, and since John XXIII all cardinals are ordained bishops. The cardinal bishops represent the early seven suburban sees, or dioceses, of Rome and, since 1965, the Eastern Catholic patriarchal sees. The cardinal priests and deacons represent the clergy of the diocese of Rome, and each of them is attached to a Roman church, his "titular" church.

Electors and advisers of the pope

Election of the pope has been the prerogative of the cardinal bishops since Nicholas II (1059) and of the entire College of Cardinals since Alexander III (1159–81). As the college became the electoral body, it also became the chief advisory body of the pope, and these two functions continue to the present day. In neither of these functions is the college a representative body because it has always had a majority of Italian members. Since the Reformation there have been demands for a wider representation, but these have never been fully met. The number of cardinals was set by Sixtus V (1586) at 70; this was not changed until John XXIII and Paul VI began increasing the number.

Cardinals are selected by the personal choice of the pope, in consultation with the cardinals in Rome at the time, in a consistory, or solemn meeting, which is secret. The pope's choice is limited somewhat by the fact that certain sees in large (or once large) cities are traditionally held by cardinals. Cardinals are either residential bishops, who live in their own sees, or resident members of the Roman Curia, who form the highest rank of papal advisers and officers of the Curia. Insofar as they are the personal staff of the pope, it is no more than proper that he should select them. Insofar as they are representatives of the whole church and electors of the pope, this method is not apt. The Roman Catholic Church is aware of this problem but seems unable to depart enough from its traditions to solve it. One solution that has been proposed is not to find a different way of selecting cardinals but to divorce the staff from the electors and representatives.

The college of bishops. It has been noted that in Roman Catholicism the college of bishops is the successor to the college of the Apostles. This is said in spite of certain differences between the two offices. The Apostles in the New Testament were a college (except for Paul, not one of the Twelve); the bishops are individual officers, and their collegial function has not been operative in recent centuries. The Apostles had a power that was not defined locally; every Roman Catholic bishop is a bishop of a place, either a proper area, a jurisdiction, of which he is the ordinary (as he is called in church law), or a fictitious place, a see no longer existing, of which he is named titular bishop. Such a monarchical officer does not appear in the New Testament. Nevertheless, Ignatius of Antioch, whose letters (written about 107) provide an early description of the Christian community, was clearly a monarchical bishop, and he did not think himself the only one of his kind; thus, the institution must have arisen in apostolic or early post-apostolic times.

Monarchical officers

The bishops in Roman Catholic belief succeed to the apostolic power, which is understood as the power to teach Catholic doctrine, to sanctify the church through the administration of the sacraments, and to govern the church. The residential bishop is supreme in his territory in this threefold function, having no superior other than the Roman pontiff. An archbishop governs a metropolitan see, usually the largest or oldest see in a region of several dioceses called a province. The metropolitan archbishop convokes and presides at provincial synods, or meetings, and has certain rights of visitation; but he has no jurisdiction in the suffragan, or subordinate, sees. The power of the bishop in governing is only over his own diocese; even there, however, it is not absolute, because church law provides the bishop with certain advisory bodies.

Until the second Vatican Council, the Roman Catholic Church had not dealt with the ambiguity of two concurrent jurisdictions, pontifical and episcopal. The pope cannot define or limit the powers of a bishop; the powers are "ordinary," inhering in the office itself. The second Vatican Council accepted the emphasis that recent theologians have laid on the collegial character of episcopacy, and the supremacy of the pope is understood as supremacy in the college; the pope needs the college of which he is head, although the first Vatican Council declared that he needs neither its consultation nor its approval. It is now understood that such solitary action should be the emergency rather than the rule; and the synod of bishops, established after the second Vatican Council, was a step toward involving the body of bishops in the policy of the entire church, hitherto formulated exclusively by the Roman see.

The qualifications for a bishop as defined in church law, which is known as canon law, are so general as to suit candidates for any office in the church, major or minor. Because bishops have been chosen by the pope since the 11th century, the qualifications are not made public and depend on a number of factors that are difficult to assess. In modern practice most bishops have been career administrators in the church, rarely pastors or scholars. Election is a much older tradition, and there have been many calls for a restoration of election of bishops, but, because the selection of bishops is a basis of Rome's power over the whole Catholic Church, Rome has been generally unsympathetic to these calls.

Modern duties of a bishop

Bishops in modern times are more visible as managers of the business of the diocese than as pastors and teachers. The responsibilities of the office are great and demand leadership and the ability to delegate business to a competent staff. A common criticism from certain quarters within the Catholic Church in the mid-20th century was that the episcopacy has been conceived more in terms of power than in terms of leadership. One of the reasons that episcopal conferences of nations and regions were set up following the second Vatican Council was to promote leadership by giving bishops the strength that lies in community. No authentic "Catholic" activity is conducted in a diocese without at least the tacit approval of the bishop; his disposal of funds and persons makes it evident that the activity will flourish much more vigorously if it enjoys his active support and encouragement. His power to discourage or forbid activities, which he is free to use according to his own sole judgment, is both a strength and a weakness of the Roman Catholic structure.

The bishop is assisted in governing the diocese by a staff called, like the staff of the pope, a curia. The structure of the staff is to some extent determined by canon law—*e.g.*, vicar general, chancellor, and official, or head, of the diocesan tribunals. Otherwise, the bishop at his discretion may appoint a staff according to the needs of the diocese. In modern times a great amount of diocesan business has settled upon the bishop. This has become a constant strain on the structure of the Roman Catholic episcopacy, and it cannot be a strength when so much of the time of the supreme officer must be consumed by purely routine business.

The power of the bishop over his clergy has been absolute, with almost no effective restraint except the human kindness of the bishop. This appeared to be changing, following the second Vatican Council, with the institution of senates of priests. The council strongly recommended that bishops introduce priests into the decisions of the diocese, but this was left to the discretion of the bishop.

Ecumenical councils. Regional councils of bishops to settle doctrinal and disciplinary questions appeared in the 2nd century. The first general council representing the bishops of the whole world (the Greek *oikumenē* referred to the inhabited world) occurred at Nicaea in Asia Minor in 325. The council was convoked not by an ecclesiastical authority but by Constantine, who wished to have a final decision on the Arian controversy. (According to Arius, the Son of God was a creature of like but not the same substance as God the Father.) The representatives of Constantine's bishop, the bishop of Rome, presided over the council. The Roman Catholic Church has held 21 such assemblies. The chronological distribution of the last three (Trent, 1545–63; first Vatican, 1869–70; second Vatican, 1962–65) shows that in modern times the ecumenical council has been convened less frequently.

The settling of doctrinal and disciplinary questions

Canon law defines an ecumenical council and its procedure; actually, the law represents the procedure followed in the convocation of the first Vatican Council. There is no real criterion for an ecumenical council, and one can say only that those councils are ecumenical that the Roman Catholic Church regards as ecumenical. The Orthodox churches recognize the first eight only.

The ecumenical council is recognized by the Roman Catholic Church as the supreme authority. With the pope this makes two supreme authorities; the Roman Church reconciles this logical dilemma by asserting that the ecumenical council, acting with the pope, is supreme. Only the pope can convoke an ecumenical council, and he or his legates must preside. There are no limits to the competence of an ecumenical council, but its decrees must be approved by the pope for validity.

The scandals of the Great Western Schism (1378–1417), which at its worst saw three men claiming the papacy, and the corruption of the papal court during the 15th century led to the movement of "conciliarism," according to which the ecumenical council was the means of saving the church from scandal and corruption. Much of the policy of the Roman see since that time has been devoted to the suppression of any feeling like conciliarism. This has naturally led to questions about the value of ecumenical councils, which are cumbersome and expensive, when an omnicompetent office such as the papacy is prepared to handle the business of the Roman Catholic Church. Both the first and second Vatican councils illustrated the values of the ecumenical council. Apart from the public and psychological impact produced by a consensus so broad, the council not only makes available for the church a fund of worldwide wisdom and experience not available to the Roman Curia but also seems to generate a state of mind that raises the members of the council above their normal level of thought and action.

The priesthood. The title of priest (Greek *hiereus*) is given to no church officer in the New Testament. Nevertheless, the office appears in the 2nd century, no doubt with the development of the monarchical episcopate; the bishop needed assistance in his threefold task of teaching, sanctifying, and governing, and the priest exercises this power as an officer of the bishop. A priest is either a member of a diocese or of a religious community; but in the exercise of the threefold ministry every priest is subject to the bishop of the diocese where the ministry is conducted.

Assistant to the bishop and cultic officer

The priest is by definition a cultic officer, and the title designates the second element of the office, the work of sanctification. Certain ambiguities in the Roman Catholic clerical hierarchical system appear clearly in the priesthood. Ordination empowers the priest to administer the sacraments, but he cannot use this power except by receiving "faculties" (proper permission or license) from a bishop. Teaching and preaching are not powers conferred by ordination, but they are subject to the same "faculties." The priest is lowest in the system of government and actually does not govern unless he is a pastor. Governing is not exercised by curates (priests who assist the pastor) or by the large number of priests engaged in specialized works that can hardly be called ministries: administration, teaching, scholarship, journalism, and other activities.

The pastor of the parish is the ideal and the model priest; in spite of the fact that in large parishes the pastor may be primarily an administrator, the Catholic experiences his church directly through his parochial clergy.

The Catholic hears sermons, worships and receives the sacraments, and looks for religious counsel and direction in his parish. Many Catholics, particularly in the U.S., have their children educated in a school run by the parish. The parish is also the centre of activities ranging from merely recreational to adult education and genuinely social works, all under the direction of the clergy. In Roman Catholicism the parochial clergy are authentic pastors; the pastoral office has often been diminished in the bishop and barely visible in the pope. The strength of the Roman Catholic Church historically has been rooted in its priests, especially in its parochial clergy.

Roman Catholicism for centuries has fostered a distinct clerical identity, symbolized by clerical garb, which sets the priest as a class apart not only from non-Catholics but from Catholics. The most striking feature of this caste is celibacy. There is in the modern church considerable dissatisfaction with this clerical separation and a feeling that it interferes with the ministry. Critics point out that neither in the New Testament nor in the pre-Constantinian church was there a clerical caste; the whole church was a people set apart with a mission to the unbelieving world. Together with this dissatisfaction and related to it has been an unprecedented number of departures from the priesthood and an equally alarming fall in the number of candidates.

Religious communities. Religious communities in the Roman Catholic Church are groups of men or women who live a common life and pronounce vows of poverty, chastity, and obedience (the evangelical counsels). Life in these conditions is traditionally regarded as a state aimed at the achieving of Christian perfection (theologically defined as perfect love) and thus a state that is an option only for a minority of the members of the church. Roman Catholic theology has never quite rationalized the elitism implicit in this idea nor escaped the implicit denigration of the lay state; but up to modern times both religious and seculars have overcome the need for rationalization by mutual respect and services. Religious are distinguished from seculars; the clergy is distinguished from the laity. Religious can be laity, and clergy can be secular (priests ordained for the service of a diocese).

Hermits and monks

The origins of the religious life are seen in the anchorites, or hermits, of the 2nd and 3rd centuries, who escaped sin and temptation by flight from the world, mostly in the deserts of Syria, Egypt, and Palestine. Flight from the world became the rule of the cloister, forbidding either free entrance of "externs" into the enclosure or free egress of religious from the enclosure and imposing supervision in all dealings with seculars. The evangelical counsels meant a life of solitude and destitution and an effort to attain union with God by prolonged, almost constant contemplation. Where large numbers of hermits assembled in the same place, cenobitisni (common life) emerged, and the hermits or monks (Greek *monachos,* "solitary") elected one of their members abbot (Aramaic *abba,* "father"). Eastern monasticism produced the rules of Pachomius and Basil in the 4th century, and travellers (most notably John Cassian) introduced monasticism into the Latin Church. Eastern monasticism, principally because of a lack of discipline, dissipated much of its energy and had no further influence on the West. Western monasticism was dominated by the rule of Benedict of Nursia in Italy, who founded his communities in the 5th century (see MONASTICISM).

The Benedictine Rule emphasized less austerity and contemplation and more common life and common work in charity and harmony. It has many offshoots and variations, and it has proved itself sturdy; it is the longest continuous religious community in the Roman Catholic Church, and it has survived many near collapses and reforms. The monk did not join an "order" but a monastery. Benedictine monasteries were almost always located in remote areas, but the labour of the monks transformed these areas into food-producing areas so that they became centres of settlement. The monks who fled the world found that the world sought them out for services, which they gladly rendered. They conducted what charitable works there were and were the only people who did anything to preserve the learning of antiquity. They supported church reform and furnished many reforming popes and bishops. Benedict did not put contemplation into his rule; prayer was fulfilled by the chanting of the divine office (a set form of liturgical prayer), celebrated at specific times during the day.

Mendicant friars and clerks regular

The 13th century saw the rise of the mendicant friars (Franciscans, Dominicans, Carmelites, Augustinians). The friary was like a monastery, with common life and the divine office in choir; but the friars made excursions, sometimes at great length both in time and distance, for apostolic works, mostly preaching. All of the mendicant orders had apostolic work in mind in their foundation, and they desired a mobility that was had neither by the monks nor the diocesan clergy. They were thus at the ready disposal of the pope, and the principle of clerical exemption (exemption from the jurisdiction of the bishop) became much more important than it had been for the monks. Originally, the friars did not need even the approval of the bishop to preach in his diocese, although this freedom has been restricted in modern times. Preaching became almost the specialty of the mendicant friars in the Middle Ages, and they were important in the foundation of the universities of the Middle Ages.

The 16th century saw the third major form of religious life, the clerks regular. These communities were formally and frankly directed to the active ministry. Even the friary, with the divine office in choir and other monastic restrictions, was dropped; they wore no distinctive religious habit. According to Ignatius of Loyola, founder of the Society of Jesus (Jesuits), the best known example of clerks regular, their life imitated the manner of living of devout secular priests. The Jesuits, almost by accident, had no particular ministry and placed themselves at the disposition of the pope. The clerks regular had even greater mobility than the friars and had the resources to undertake specialized works. Since the 16th century the works of religious communities have been education, foreign missions, preaching, and theological scholarship. Orders founded since the 16th century have adopted the manner of life of the clerks regular.

Nuns and brothers

Religious communities of women until the 17th century were entirely contemplative and subject to rigid cloister, although from the 16th century they had begun to admit girls into the convent not as novices (those admitted to probationary membership in the community) but for the education of a gentlewoman. The modern communities of women all stem from the type of community instituted in France in the mid-17th century by Vincent de Paul under the name of the Daughters of Charity. At first these women were not religious and deliberately so; Vincent did not wish cloister. The group was founded to help the poor and sick and to educate their children in the rudiments and in their religion. These have remained the major works of the communities of women.

Religious communities are orders if the members (or some of them) pronounce solemn vows; they are congregations if the members pronounce simple vows. Solemn vows are perpetual; simple vows may be perpetual or temporary. The difference is subtle; solemn vows were meant to be a more permanent and durable consecration, but they are dispensable. Men who make religious profession but who do not receive the sacrament of holy orders are "brothers."

Secular institutes have arisen since World War II. They are not religious (and therefore do not pronounce the three vows), have little or no common life in a common residence, have no superior but rather a manager of the few common affairs, and intend to bear Christian witness in the world in any type of secular employment.

The laity. The laity as a class do not appear in the New Testament; there could only be a laity when there had become a clergy. When the laity appear, they are the passive element of the church. If the office of the clergy is conceived as teaching, sanctifying, and governing, then the laity are the taught, the sanctified, and the governed. Misleading identification of the church with the clergy (and, within the clergy, with the hierarchy) results.

The modern term Catholic Action (especially under

Pius X and Pius XI) meant in general the assistance of the laity in the mission of the church. Yet, as it was more closely defined, the mission of the church was still entirely clerical, and lay action was accessory to the mission proper. The laity was merely the arm of the hierarchy. Furthermore, lay action lay under close direction and supervision of the hierarchy and clergy. It is not surprising that an action so vaguely defined, so patronized, and so uninspiring aroused relatively little response.

Anti-clericalism and clericalism

Much of the 19th and 20th centuries saw the Roman Catholic Church engaged with anticlericalism in the "Catholic" countries of Europe; this seems to be a peculiarly Roman Catholic phenomenon. Actually, anticlericalism is a rejection of the medieval belief in the power of the clergy to direct all the decisions of the laymen that they thought themselves entitled to direct. Reaction in an exaggerated form nearly excluded the clergy from any activity except public worship in some countries.

The second Vatican Council definitely rejected clericalism. It called "secular" all non-ecclesiastical activity and declared that the secular is the proper area of the layman. This means that the layman is the judge of how to realize his Christian destiny in the secular life. Proper does not mean exclusive, but the statement implies that the clergy can offer only principles and general directions, not concrete decisions. The Roman Catholic Church intended to make the laity the channel of its relevance in the world.

The council also took steps against the passive role of the layman in ecclesiastical life. It recommended the establishment of lay councils in each diocese and in each parish. This has moved slowly, because Roman Catholics are not accustomed to the idea and are uncertain about how it should be implemented. As the secular is the proper but not the exclusive area of the layman, so the ecclesiastical is the proper but not the exclusive area of the hierarchy and the clergy.

CANON LAW

The earliest individual church law was called a canon (Greek *kanōn*, "rule, measure, standard"); the canons were finally called Canon Law. Church laws appear almost as soon as church authority, and some passages of the New Testament reflect early rules; whether they should be called law at this primitive stage is doubtful. Laws of dioceses or of regions appear even before Constantine; they were formed by diocesan synods or regional councils. Laws for the whole church appear with the earliest ecumenical councils. Canon Law remained scattered pieces of papal, conciliar, and diocesan legislation until the 12th century. The first collection and synthesis of Canon Law was made by Gratian in 1142, the *Decretum* Gratiani. To this collection in the next 400 years were added the decretals (papal decrees on points of law) produced in the reigns of Gregory IX (1234), Boniface VIII (1298), and John XXII (1317) and two collections known as Extravagantes (1500). These formed the Corpus *Juris* Canonici ("Body of Canon Law"); no further collection was made of laws later than the Corpus. Effectively, although not formally, Canon Law included the opinions of canonists interpreting the Corpus.

This unsatisfactory and cumbersome collection led to calls for codification. No doubt the desire was influenced by the production of the Code Napoléon, which became the basic law of most of the nations of western Europe. The codification was begun by a document of Pius X (1904) and was completed, directed by Cardinal Pietro Gasparri throughout, under Benedict XV (1917); it became law in 1918. This code has remained the basic law of the Roman Catholic Church, although, following the second Vatican Council, several different attempts were being made to revise it.

The history and structure of church law are treated more fully under CANON LAW.

IV. Beliefs

FAITH

Notions of faith. The idea of faith shared by all Christian churches is rooted in the New Testament. But the New Testament idea of faith is not simple, and it permits a breadth of meaning that has led to variations even within a single Christian communion. Most modem interpreters of the New Testament would agree to a description of New Testament faith as a total commitment of the self to God revealing himself in Christ. Yet, it is doubtful whether the post-Reformation theology of any Christian church has presented faith simply in these terms.

Faith as an intellectual assent

Even before the Reformation, faith in Roman Catholicism had developed an emphasis that is not rooted in the New Testament but can be traced back to the Alexandrian school of theology and Augustine. Faith appeared primarily as acceptance of revelation, and revelation appeared as a revelation of doctrine rather than as revelation of a person. This emphasis ultimately was formulated in the 13th century by Thomas Aquinas in a definition of faith—canonized by the Council of Trent and the first Vatican Council—as an intellectual assent given to revealed truth by the command of the will inspired by grace and motivated by the authority of God revealing.

The Reformers, with Martin Luther as the leader, rejected this idea of faith as nonbiblical and exclusively doctrinal; it seemed to place the teaching authority of the Roman Catholic Church between man and God not as a means of communication but as a replacement of God. Luther saw faith as confidence in the saving power of grace. This, Luther believed, was a return to the New Testament faith. Roman Catholicism rejected this as a mere sentiment; and these positions were crystallized up to the 20th century. At the risk of oversimplification, it is possible to say that both represented exaggerations of the New Testament. New Testament faith is more than either trust in the saving power and will of God or assent to revealed truth, although neither element can be entirely excluded. Controversy was wasted in trying to prove the adversaries wrong rather than in trying to understand the New Testament. The documents of the second Vatican Council reflect a shift in Roman Catholic theology from emphasis solely on faith as intellectual assent to recognition of faith as a loyal adherence to a personal God.

Roman Catholic theology, having chosen the option of faith as assent, was faced with the problems of showing that it was a rational assent rather than an irrational assent and of maintaining that faith was a deliberate and free meritorious act under the inspiration of grace. At first glance the two problems seem to cancel each other out; one can maintain one affirmation only by denying the other.

Preambles and motivation of faith. The study of the problems connected with faith involves the investigation of what are called the preambles of faith and also of the motivation of faith. The preambles of faith include those processes by which the believer reaches the conclusion that it is reasonable to believe—*e.g.*, the proof of the existence of God by the use of one's own reason. The freedom of faith is respected by affirming that this conclusion is as far as the preambles can take one. This process as proposed is a theoretical construction that actually occurs in no one, but the analysis can be of value in uncovering the psychological processes that occur without reflection. The preambles include the study of the scientific and historical difficulties raised against the Christian fact (*i.e.*, the incarnation, Resurrection, Ascension, and glorification of Jesus Christ) itself or against the Roman Catholic interpretation and proclamation of the Christian fact or against the Roman Catholic claim to be the exclusive custodian of revealed doctrine and the means of salvation. These studies were efforts to show what cannot be shown by scientific and critical methods, but in the exaggerated claims of their defenders they showed that faith was a necessary conclusion of a valid rational process. Such a faith could be neither free nor the result of grace.

The evidence for the act of faith

The study of the motivation of faith attempted to meet this difficulty. Some earlier analyses candidly presented faith as resting on evidence and clumsily postulated a movement of grace necessary to assent to this particular evidence. Normally, one "wills" to believe something because the evidence is not compelling; thus, one chooses to believe that the candidate of his choice has the qualities desired for the office, although the evidence is less than overwhelming. The Roman Catholic thinks this is an assent to the probably rather than the certainly true and yet insists that the certainty of faith is the highest of all certainties. Ultimately, the Roman Catholic analysis must say that the evidence that belief is reasonable can never be so clear and convincing that it compels the radical deviation from worldly patterns that assent implies. At this point, the will inspired by grace chooses to accept revelation for other reasons than the evidence.

The motive of faith that has been presented by Catholic theologians is "the authority of God revealing." It is held that the preambles of faith show beyond reasonable doubt that God exists and that he has revealed himself. This evidence and an acceptance of the notion that, if God reveals himself, he does so authoritatively motivate a person to make the act of faith. The problem with such an analysis has been how the authority of the revealer is manifest to the believer. It seems that the authority of God revealing must be an object of faith rather than a motive, because the conjunction of this authority with the fact of revelation cannot be the object of historical experience. In the mid-20th century, this dilemma caused an increasing number of Catholic theologians to move closer to a view that emphasized faith as a personal commitment to God rather than as an assent to revealed truth.

Heresy. Heresy is the denial by a professed, baptized Christian of a revealed truth or that which the Roman Catholic Church has proposed as a revealed truth. The unbaptized person is incapable of heresy, and the baptized person is not guilty of "formal" but only of "material" heresy if he does not know that he denies a revealed truth. The seriousness with which Roman Catholicism regarded heresy is shown by the ancient penalty of excommunication. Civil penalties, including the supreme penalty, did not appear until the Constantinian age. Lesser civil disabilities continued in force, although the law was often ignored, into the 20th century. Protestant governments often borrowed some of this severity from Roman Catholic governments.

Roman Catholic theologians often deal with heresy, paradoxically, as a necessary step in the development of dogma. In order to save themselves from an extremely crass and even cruel rationalization, they point out that the questions raised by heresy were legitimate but that heretics too quickly assumed a one-sided and exclusive view of doctrine that they wished to impose on the entire church. Modern studies have sometimes been less kind to such champions of orthodoxy as Athanasius and Cyril of Alexandria, who were not themselves free of one-sided views and who showed themselves unwilling to listen to their adversaries with sympathy and understanding. In recent times most of the theses of Modernism (a movement to change the Catholic Church by means of radical renovation), which were condemned vigorously by Pius X in 1907, have found their way into Catholic theology. This may have something to do with the absence of the words heresy and heretics from the acts of the second Vatican Council. Like the use of the word church for Protestant churches, this indicates a substantial change of attitude toward a genuinely ecumenical position.

REVELATION

Notion of revelation. Although other religions have an idea of revelation, none of these ideas bears a close resemblance to the idea of revelation found in the Old and New Testaments and in Christianity. Roman Catholic theologians distinguish between revelation in a broad sense, which means knowledge about God deduced from nature and man (and therefore actually philosophy), and revelation in the strict formal sense, by which they mean the utterance of God. This latter idea, of course, can only be conceived by analogy with the utterance of man, and its precise definition involves difficulties.

The earliest idea of revelation is the idea found in the Old Testament in which the speech of God is addressed to Moses and the prophets. They in turn are described as quoting the words of God rather than interpreting them. Jesus, the fulfillment of the prophets, does not speak the word of God; he is the word of God. This phrase, which occurs only in the opening verse of both the Gospel and the First Letter of John, has become a technical term in theology; Jesus is the Incarnate Word. As such he is both the revealer and the revealed. He reveals the Father both by what he says and by what he is. Thus, the earliest Gospel (literally "good news") is the account of the life, death, and Resurrection of Jesus. The Gospel as the recital of his words appears in a later phase of development.

Jesus as revealer and revealed

It has been noted that the Roman Catholic Church has regarded revelation primarily as the revelation of propositions rather than the revelation of a person. Thus, it has thought even of Jesus more as a spokesman who tells of God than as a reality who himself in his being and actions manifests God. Though this latter aspect is found to some extent in the documents of the second Vatican Council, it has normally been considered only in the miracles of Jesus, which have been regarded in Roman Catholic apologetics as works of divine power that assure the credibility of the words of Jesus. These words, which were spoken in a particular historical context, have been preserved in a twofold way. They are written in the Gospels, which together with the Old Testament form a book of revelation that is distinct from the spoken words; but, because the Bible itself is written under divine inspiration, it has the same authority of revelation as the spoken words of Jesus. The Roman Catholic Church also preserves the words of Jesus, independently of the Bible, in its traditional teaching; but it does not utter the very words spoken by Jesus, and thus its words have a lower formal quality of revelation than the words of the Bible, although they are of equal authority. The idea of a book of revelation was taken by the early Christian Church from Judaism when it accepted the sacred books of the Jews as its own, just as it accepted the God of Judaism as the Father whom Jesus claimed for his own.

Mystery and the supernatural

The content of revelation. The proper content of revelation is designated in Roman Catholic teaching as mystery; this theme was important in the documents of the first Vatican Council. The development of the theme of mystery responded to those intellectual movements of the 18th and 19th centuries that are called by such titles as the Enlightenment, Rationalism, scientism, and historicism. To the Roman Catholic Church these movements were threats to the idea of a sacred revelation; they appeared to claim that human reason had no frontiers or that human reason had demonstrated that revelation was historically false or unfounded or that the content of revelation was irrational. The affirmation of mystery meant that the reality of God was unattainable to unaided human reason; theologians had long used the word incomprehensible, which says more than modern theologians wished to say. Mystery refers both to the divine reality and to the divine operations of the world. These operations can be observed only in their effects; the operation itself is not seen, nor is its motivation seen. The plan of God, which is realized in history, is mysterious. The first Vatican Council insisted that the existence of God and of a moral order is attainable to reason, and some of the fathers of the council wished to state that these truths were imposed upon reason by the evidence, a step that the council did not choose to take. Mystery does not mean the incomprehensible or the unintelligible; it means, in popular language, that man cannot know who God is or what God is doing or why God is doing it unless God tells him. Mystery also means that, even when the revelation is made, the reality of God and his works escapes human comprehension.

The term supernatural has been used in Roman Catholic theology since the 17th century to designate not only revelation but other aspects of the divine work in the world. The term has an inescapable ambiguity that has led many modern theologians to avoid its use. The "natural" that the supernatural presupposes is the world of human experience; the quality of this experience is not altered by technological and social changes as long as these are fulfillments of the potentialities of nature. Indeed, it is the spectacular growth in the knowledge of these potentialities in modern times that leads to doubt as to whether there can be a supernatural at all. The supernatural reality is identified with God in his reality and in his operations. This is a reality that man cannot create or control. The supernatural in cognition is this reality as it is perceptible to man; it is, for man, simply unknown as far as unaided reason can move. The first Vatican Council affirmed that without revelation human reason historically has not reached anything but a distorted idea of the divine and an imperfect idea of the moral order. This means also that man is unaware of his destiny, either individually or collectively, without revelation and that he is unable to achieve it without the entrance of the supernatural into the world of history and experience.

Contemporary theologians of revelation are aware of the problems raised by historical and literary criticism that render it impossible to cherish the primitive idea of revelation as the direct utterance of God to man. Roman Catholic theologians have not found a satisfactory way of describing revelation, but they do not see that the destruction of a naïve idea of revelation destroys the whole idea. Theologians also recognize that the older idea of revelation of propositions as a collection of timeless and changeless verities, almost like a string of pearls, is no longer tenable. Every utterance that is called revelation was formed in a definite time and place and bears the marks of its history. There is no revealed proposition that cannot be restated in another cultural situation. Contemporary theologians are aware that these propositions must be restated if the Roman Catholic Church is to speak meaningfully in the modern world. Roman Catholicism does not accept the possibility of a new revelation; it believes that reason can never completely penetrate the "mystery" and that it must continue the exploration of the mystery that has already been revealed.

TRADITION AND SCRIPTURE

Tradition as channel and content

In Roman Catholic theology tradition is understood both as channel and as content. As channel it is identical with the living teaching authority of the Catholic Church. As content it is "the deposit of faith," revealed truth concerning faith and morals. In Roman Catholic belief, revelation ends with the death of the Apostles; the deposit was transmitted to the college of bishops, which succeeds the Apostles.

The Reformers contended that the Roman Catholic Church had imposed teachings that were not contained in the Scriptures, and this Protestant objection has been maintained in modern times. The objection was raised more intensely when the Immaculate Conception of Mary, the mother of Jesus (Pius IX, 1854), and her Assumption (Pius XII, 1950) were defined as dogmas. For neither of these is there any biblical evidence; more significantly, there is no evidence in tradition for either before the 6th century.

The Roman Catholic Church recognizes that the Bible is the word of God and that tradition is the word of the church. In one sense, therefore, tradition yields to the Scriptures in dignity and authority. But against the Protestant slogan of *sola* Scriptura ("Scripture alone"), itself subject to misinterpretation, the Roman Catholic Church recognized that the church existed before the New Testament. In fact, the church both produced and authenticated the New Testament as the word of God. For this belief, at least, tradition is the exclusive source; and this furnished a warrant for the Catholic affirmation of the body of truth transmitted to the church through the college of bishops preserved by oral tradition (meaning that it was not written in the Scriptures). The Roman Church therefore affirmed its right to find out what it believed by consulting its own beliefs as well as the Scriptures. The Council of Trent affirmed that the deposit of faith was preserved in the Scriptures and in unwritten (not in the Bible) traditions and that the Catholic Church accepts these two with equal devotion and reverence. The council studiously avoided the statement that they meant these "two" as two sources of the deposit, but most Catholic theologians after the council understood the statement as meaning two sources. Protestants thought it meant the Roman Catholic Church had written a second Bible.

The question of the two sources of revelation

Only in contemporary Catholic theology has the question been raised again, and a number of theologians believe that Scripture and tradition must be viewed as one source. They are, however, faced with the problem of nonbiblical articles of faith. To this problem several remarks are pertinent. The first is that no Protestant church preaches "pure" gospel; they have all developed dogmatic traditions, concerning which they have differed vigorously. It is true, on the other hand, that they do not treat these dogmatic traditions "with equal devotion and reverence" with the Bible. The second is that the early Christian Church through the first eight ecumenical councils (before the Eastern Schism in 1054) arrived at nonbiblical formulas to profess its faith. Protestants respond that this is at least a matter of degree and that the consubstantiality of the Son (*i.e.*, that he is of the same substance as the Father), defined by the Council of Nicaea, is more faithful to the Scriptures than the Assumption of Mary.

Roman Catholics and Protestants should be able to reach some consensus, not yet formulated, that tradition and Scripture mean the reading of the Bible in the church. Protestants never claimed that a man and his Bible made a self-sufficient Christian church. The New Testament itself demands that the word be proclaimed and heard in a church, and the community is formed on a common understanding of the word proclaimed. This suggests a way to a Christian consensus on the necessity and function of tradition. No church pretends to treat its own history of belief as nonexistent or unimportant. By reading the Scriptures in the light of its own beliefs it is able to address itself to new problems of faith and morals that did not exist in earlier times or to which the church for some reason did not attend.

Development of dogma

Catholic theologians of the 19th century dealt with the problem under the heading of development of dogma. To a certain extent the question can be reduced to epistemology (*i.e.*, theory of knowledge) : is a new understanding of an ancient truth a "new" truth? The problem does not arise out of faith; Sir Isaac Newton's observations of falling bodies saw nothing that men had not seen for thousands of years. Yet the effects of Newton's insights and calculations altered man's understanding of the universe and his actions within the universe. The problem is important in theology because of the necessity of basing belief on the historical event of the revelation of God in Christ. Unless the link is maintained, the church is teaching philosophy and science, not dogma. Hence, the Roman Catholic theological teaching has tended to say that dogma develops through new understanding, not through new discoveries.

THE TEACHING AUTHORITY OF THE CHURCH (THE MAGISTERIUM)

Notion of teaching authority. The Roman Catholic Church claims for itself a teaching authority that is unparalleled in the christian community. The Reformation was primarily a rebellion against the teaching authority, and the Reformers did not claim for their own churches the authority they rejected in the Roman Church.

To teach with authority means that the teacher is able to impose his doctrine upon the listener under a religious and moral obligation. This moral obligation does not flow from the nature of teaching, which of itself imposes no obligation upon the learner; the learner is morally obliged only to assent to manifest truth. The Roman Church derives its teaching authority from the commission given by Jesus to the Apostles as contained in the New Testament ("He who hears you hears me"). The response of the hearers of the Apostles was faith; the response of the Roman Catholic is expected to go beyond faith. The Apostles were presumed to speak to those who had not yet believed; the Roman Catholic Church imposes its teaching authority only upon its members. The definition of the teaching authority must show that these modifications do not exceed the limits of legitimate doctrinal development.

The hierarchy

Organs of teaching authority. The teaching authority is not vested in the whole church but in certain well-defined organs. These organs are the hierarchy—the pope and the bishops. The Roman Catholic Church traditionally has divided the church into "the teaching church" and "the listening church." Clergy below the hierarchical level are included in "the listening church," even though they are the assistants of the bishops in the teaching office. The hierarchy alone teaches what the Roman Catholic Church calls "authentic" doctrine. There is an unresolved antithesis between this idea and the traditional belief that "the consent of the faithful" is a source of authentic doctrine; the conventional resolution that defines the consent as formed under the direction of the pastors of the faithful resolves the problem by depriving the consent of the faithful of any meaning.

The Roman pontiff is vested with the entire teaching authority of the Roman Catholic Church; this was solemnly declared in the first Vatican Council. This means that he is the only spokesman for the entire Roman Church; the papacy carries in itself the power to act as supreme pastor. It is expected that he will assure himself that he expresses the existing consensus of the church, but in fact the documents of the first Vatican Council are open to the understanding that the pope may form the consensus by his utterance. The second Vatican Council clarified this ambiguity in the idea of the spokesman of the church by its emphasis on the collegial character of the primacy of the pope. The pope, however, does not always speak as the supreme pastor and head of the Roman Church, and he is expected to make this clear in his utterance.

The bishops are authentic teachers within their dioceses. Thus, the same implicit conflict exists in regard to teaching as was noted in connection with governing. The conflict is resolved by collegiality; that the authentic teacher teaches orthodox doctrine is recognized by comparing his doctrine with that of his episcopal colleagues. In this way, doctrinal disputes were resolved in the pre-Constantinian church, and a regional council was called if necessary. Since the Reformation, the Roman see has never admitted publicly that a bishop has fallen into doctrinal error; the united front of authentic doctrine is preserved, and the matter is dealt with by subtle means. What is taught by all the bishops is authentic doctrine; it is understood that they teach in communion with the Roman pontiff, and a conflict of doctrine on this level is simply not regarded as a possibility. This consensus of the bishops is known as "the ordinary teaching." "The extraordinary teaching" signifies the solemn declaration of an ecumenical council, which is the assembly of the bishops, or the most solemn type of papal declaration, known as a definition of doctrine ***ex cathedra*** ("from the throne"), a term that signifies that the declaration exhibits the marks of the teaching of the supreme pastor addressed to the universal church.

Object and response. The object of authentic teaching is defined as "faith and morals." Faith means revealed truth. Morals theoretically means revealed moral principles, but it has long been understood as moral judgment in any area of human conduct; thus, the Roman Catholic Church not only prohibits contraception for its members, but by declaring it contrary to "the natural law" the church declares contraception to be universally wrong. Thus, morals includes the declaration and interpretation of the natural law. The limits of faith and morals have never been declared by the Roman Catholic Church, and one cannot take the exercise of the teaching authority as a reliable guide. The teaching authority condemned the heliocentric theory of Galileo as contrary to the Bible. The teaching authority has always understood that revealed truth involves other propositions that are not themselves revealed but that must be affirmed or denied, at least in the present context of knowledge, because of revealed doctrine.

Dogma is the name given to a proposition that is proclaimed with all possible solemnity either by the Roman pontiff or by an ecumenical council. A dogma is a revealed truth that the Roman Catholic Church solemnly declares to be true and to be revealed; it is most properly the object of faith.

Infallibility

The first Vatican Council declared that the pope, when he teaches solemnly and in the area of faith and morals as the supreme universal pastor, teaches infallibly with that infallibility that the church has. The infallibility of the church has never been defined, and its extent is understood by theologians in the sense of pontifical infallibility as limited to faith and morals. These terms are ambiguous, as noted above. Infallibility is actually hedged in with many reservations; nevertheless, pontifical documents often have an aggressive tone that may mislead the incautious reader. The real problem is how a teaching authority that can and does make errors in doctrinal teaching can be called infallible, even with numerous and serious reservations. In the early 1970s some Catholic theologians (*e.g.*, Hans Kiing) suggested that the church should be understood as indefectible (*i.e.*, not able to fail or he totally led astray) rather than infallible.

The proper response of the Roman Catholic to authoritative teaching that is "ordinary" and does not clearly deal with "faith or morals" is religious assent. This is extremely difficult to define; it admits dissent under poorly defined conditions. But the theory of religious assent does in fact permit the somewhat massive dissent from the authoritative teaching of Paul VI in 1968 against contraception. Religious assent is particularly relevant to the pontifical document called the encvclical, a type of document that first appeared in the 18th century and became the normal mode of pontifical communication in the 19th century. The encyclical letter is a channel of ordinary teaching, not solemn and definitive and somewhat provisional by definition. Religious assent may be withheld, in popular language, by anyone who in good conscience thinks he knows better. The traditional discipline has made Roman Catholics slow to say this; in modern times they say it more quickly. At the same time, the documents of the second Vatican Council indicate that the authoritative teaching body will be slower to assert itself in the future.

MAJOR DOGMAS AND DOCTRINES

The Roman Catholic Church in its formula of Baptism still asks the candidate to recite the Apostles' Creed as a sign that he believes what he must believe. The early Church Fathers made the creed the basis of the baptismal homilies given to catechumens, those preparing for the rite. The homilies, like modern Roman Catholic doctrine, went considerably beyond the bare articles of the creed.

Grace

Roman Catholic faith incorporates into its structure the books of the Old Testament. From these books it derives its belief in original sin, conceived as a hereditary and universal moral defect that makes man incapable of achieving his destiny and even of achieving basic human decency. The importance of this doctrine lies in its ex. planation of the human condition as caused by the failure of man and not by the failure of God (nor, in modem Roman Catholic theology, by diabolical influence). Man can be delivered from the human condition only by a saving act of God. This act is accomplished by God in the death and Resurrection of Jesus. In Jesus, God is revealed as the Father who sends the Son on his saving mission, and through the Son the Spirit comes to dwell in the redeemed. Thus, the Trinity of Persons is revealed, and the destiny of man is to share the divine life of the three Persons. The saving act of Jesus introduces into the world grace, a theological idea that has been much and hotly disputed. Grace signifies in Roman Catholic belief both the love of God and the effect produced in man by this love. The response of man to the presence of grace is the three theological virtues of faith, hope, and charity; these enable him to live the Christian life. Man is introduced to grace and initiated into the church by Baptism, which must be preceded by repentance and faith. The life of grace is sustained in the church by the sacraments.

The life of grace reaches its fulfillment in eschatology; in this area of belief about the end of the world and "the last things," there is some uncertainty in modern theology. Most theologians recognize the mythological character of most of the imagery of heaven, hell, and purgatory. The peculiarly Roman Catholic belief in purgatory was an effort to state that most men at death are neither good enough for heaven nor bad enough for hell. The theology of the last things is still unable to cope with the implications of this statement. Belief in a resurrection to eternal life has never been easy, and modern times have produced more difficulties than solutions. Christianity, in fact, shows oscillation between a transcendental direction and an immanent direction; in modern times the emphasis is on immanence—that is, on the meaning of religion in the world. The second Vatican Council reflected this in its statements on the "secular" and the response of the church to the secular.

This summary can state no more than the basic elements of the Christian fact. The complex Roman Catholic dogmatic structure has been mentioned several times, and probably no two statements of "major dogmas and doctrines" would be the same.

V. Worship

THE LITURGY

Cultic worship is so universal in religion that some historians of religion define religion as cult. Cultic worship is social, and this means more than a group worshipping the same deity in the same place at the same time. Cult is structured with a division of sacred personnel (priests) who lead and perform the cultic ceremonies for the people, who are in a more distant relation with the deity. The sacred personnel are designated by the choice and acceptance both of the deity and of the worshipping group. The words and actions of the cultic performance are divided into roles assigned to the leaders and to the worshippers. It is the tendency of cultic worship to replace spontaneity, which it once had, with set and even rigid forms of words and acts. These are preserved by tradition, and they generally have a sacredness that is based on the belief that the directions for cultic worship came ultimately from the deity.

Central act of liturgy

The eucharistic assembly or mass. Roman Catholic liturgy has its roots in Judaism and the New Testament. The central act of liturgy from earliest times was the eucharistic assembly, the commemorative celebration of the Last Supper of Jesus. This was set in a structure of liturgical prayer. The first six centuries of the Christian Church saw the development of a rich variety of liturgical systems, many of which have survived in the Oriental churches. In the West the Latin liturgy appeared fully developed in Rome in the 6th and 7th centuries. From the 8th century the Roman liturgy was adopted throughout western Europe. In this same period, however, liturgy developed in Frankish territories; and the Roman rite that emerged as dominant in the 10th century was a Roman–Frankish creation. The Roman rite was reformed by the Council of Trent by the removal of some corruptions and the imposition of uniformity; after Trent the Roman see was the supreme authority over liturgical practice in the entire Roman Catholic Church.

By the 11th century, Roman liturgy had acquired the classic form that it retained up to the second Vatican Council. The fullness of the liturgy could be witnessed only in some cathedrals, collegiate churches, and monastic churches. The full liturgy included the daily celebration of the solemn high mass and recitation of the divine office in choir. The solemn high mass was performed by at least three major officers (celebrant, deacon, and subdeacon), assisted by many acolytes and ministers. Except during the penitential seasons of Advent and Lent, the altar was decorated, and numerous candles (in the Middle Ages for light rather than ornamentation) and incense were employed. The singing and chanting were accompanied by the organ and in modern times even by orchestral music; Mozart once complained that the Archbishop of Salzburg compelled him to compose a mass without the resources of a full symphonic orchestra.

The divine office. The divine office was a legacy to the clergy from the monks. From the beginnings, monks assembled several times daily for prayer in common. This developed into set common prayer at stated times each day (Matins, midnight; Lauds, first daylight; Prime, sunrise; Terce, midmorning; Sext, noon; None, midafternoon; Vespers, sunset; Compline, before retiring). The divine office consisted basically of the chanting of the Psalms (in a weekly cycle), the recital of prayers, and the reading of the Scriptures (to which were later added selections from the writings of the Church Fathers, probably instead of a homily given by one of those present). Together with the mass, the office has been the only "official" prayer of the Roman Catholic Church; all other prayer forms are "private," even if several hundred people recite them together. For this reason, clerics in major orders for centuries since the Middle Ages have been obliged to recite the divine office, or "breviary," privately if they are not bound to attend the office in choir. It was long recognized that there is an inconsistency in the private silent reading of a prayer structure that is intended for choral chanting, and the second Vatican Council recommended a reform that was not yet implemented in the early 1970s. Many priests, however, had abandoned the breviary.

Official prayer of the church

The cycle and the language of the liturgy. The liturgy has long been arranged in an annual cycle that is a re-enactment of the saving events of the life, death, Resurrection, and glorification of Jesus Christ. Even many Catholics do not realize that the cycle has an eschatological outlook; the events are re-enacted as an assurance that the saving act will reach its eschatological fullness, and the liturgy is an expression and a support of the Christian hope. The cult of the saints is an intrusion into the liturgical cycle, and it has been much reduced in the contemporary liturgical reforms.

Latin did not become the language of the Roman rite until the 6th century; the language of imperial Rome was Greek. As a sacred language Latin really has no parallel. Jews have always made a genuine effort to learn some Hebrew, and other sacred languages are archaic forms of the vernacular; the English of the Authorized Version of the Bible became the language of prayer in many Protestant churches. The effect of Latin was to make the liturgy the preserve of the clergy, and the laity became purely passive. This was countered by the efforts to use sound and spectacle in the performance of the solemn liturgy. The Canon of the mass, the central eucharistic formula, for centuries was recited by the celebrant inaudibly; this was a kind of verbal "sanctuary" that the laity were not even supposed to hear. The abandonment of Latin as a result of the second Vatican Council excited deep antagonisms; one sees in the Latin liturgy an image, cherished by many, of the timeless and changeless Roman Catholic Church. Yet, the restoration of the vernacular should restore to the liturgy two functions that it had in the early centuries: to instruct converts and to confirm members in their faith.

THE SACRAMENTS

The sacraments in general. In Roman Catholic theology a sacrament is an outward sign, instituted by Jesus Christ, that is productive of inner grace. The number is seven (defined by the Council of Trent against the Reformers, who reduced the number). The number seven does not appear in Roman Catholic teaching before the 11th century, and it is an example of truth for which the Roman Catholic Church relies on its own tradition.

The sacrament as an encounter with mystery

The sacrament in modern theology is frequently described as an encounter with mystery, the mystery being the saving act of God in Christ. Theological studies have been directed to the exploration of the idea of sign and significance. The traditional Roman Catholic statement of the effectiveness of the sacraments (defined by the Council of Trent) is described by the untranslatable ex *opere operato,* which is best explained briefly by saying that the faith and virtue of the minister neither add to the sacrament by their presence nor detract from it by their absence. The minister is merely the agent of the church, and the effectiveness of the sacrament is based on the saving act of God in Christ, which is signified by the rite and applied to the recipient of the sacrament.

Protestant theologians formerly charged the Roman Catholic Church with a belief in magic; this controversial angle has generally been abandoned, but the theological explanation of the sign that effects by signifying is still difficult. Roman Catholic theologians remark that the mystery of God's saving act is not capable of complete rational explanation. There are analogies, however, in common experience, and there is no society that does not employ effective signs. These signs are not merely for display. The inauguration of the president of the United States makes the man president; the sign is effective because it signifies the reality of the election that this individual won. The sign of the coronation of a monarch is equally effective, but it is more difficult to define the reality signified. Such effective symbols are a part of human society.

The Roman Catholic Church adheres strictly to the external sign. Traditionally, the church attributes the institution of the sign to Jesus Christ (although this has been the subject of discussion among modern theologians), and this removes the right of anyone to tamper with it. The Roman Catholic Church believes that, if God gave a sign, the alteration of the sign so that the significance is lost might render the sign ineffective. Hence, the use of the proper material and the retention of the traditional formula are treated as sacred. The Roman Catholic Church maintains its own exclusive competence to supervise matter and form "in detail," a competence not precisely defined. Since Thomas Aquinas the material used is called matter, and the words are called form; the terms are borrowed from the Aristotelian theory of the constitution of matter. The material becomes sacred and salutary only by its conjunction with the proper words. The effect produced has been called for centuries grace, but it is difficult to assert a single effect and still explain why there are seven symbols.

Sacramentals

The term sacramental is used to designate verbal formulas (such as blessings) or objects (such as holy water or medals) to which a religious significance has been attached. These are symbols of personal prayer and dedication, and their effectiveness is measured by the particular dispositions of the person who uses them. Although superstition has arisen in connection with sacramentals, the Roman Catholic with elementary instruction knows the difference between these things and sacraments.

Baptism. Baptism is the sacrament of regeneration and initiation into the church. According to a theme of St. Paul, probably influenced by Jewish belief in the circumcision of adult proselytes, Baptism is death to a former life and the emergence of a new person, signified by the conferring of a new name; it is the total annulment of the sins of one's past and the emergence of a totally innocent person. One becomes a member of the church and is incorporated into the body of Christ, thus becoming empowered to lead the life of Christ. Nothing but pure natural water may be used, and Baptism must be conferred in the name of the Father, the Son, and the Holy Spirit. Baptism is normally conferred by a priest, but the Roman Catholic Church accepts the Baptism conferred by anyone having the use of reason "with the intention of doing what the church does." As the sacrament of rebirth it cannot be repeated. The Roman Catholic Church baptizes conditionally in case of doubt of the fact of Baptism or the use of the proper rite.

Points of controversy

Two points of controversy still exist in modern times. One is Baptism by pouring rather than immersion, although immersion was probably the biblical and early Christian rite. The change was almost certainly the result of the spread of Christianity into Europe north of the Alps and the occurrence of the baptismal feasts, Easter and Pentecost, often in early spring. The Roman Catholic Church simply asserts that the symbolism of the bath is preserved by a ritual infusion of water.

The second is the Baptism of infants. There is no certain evidence of this earlier than the 3rd century, and the ancient baptismal liturgies are all intended for adults. The liturgy and the instructions clearly understand the acceptance of Baptism as an independent adult decision; without this decision the sacrament cannot be received. The Roman Catholic Church accepts this principle by introducing adults (sponsors, godparents), who make the decision for the infant at the commission of the parents. In Roman law as in modem law, adults are empowered to make decisions for minors. It is expected that the child will accept the decision made for him and will thus supply the adult decision that was presumed.

Until the recent liturgical renewal, Baptism did not have the religious and ceremonial importance that it had in the early church; the ceremonies were intended to make the adult aware that he had made the most important decision of his life, and the whole church witnessed the ceremony, performed only twice a year on a group of catechumens. Doubtless the Baptism of infants contributed to this loss of ceremonialism and to a corresponding lower esteem of Baptism.

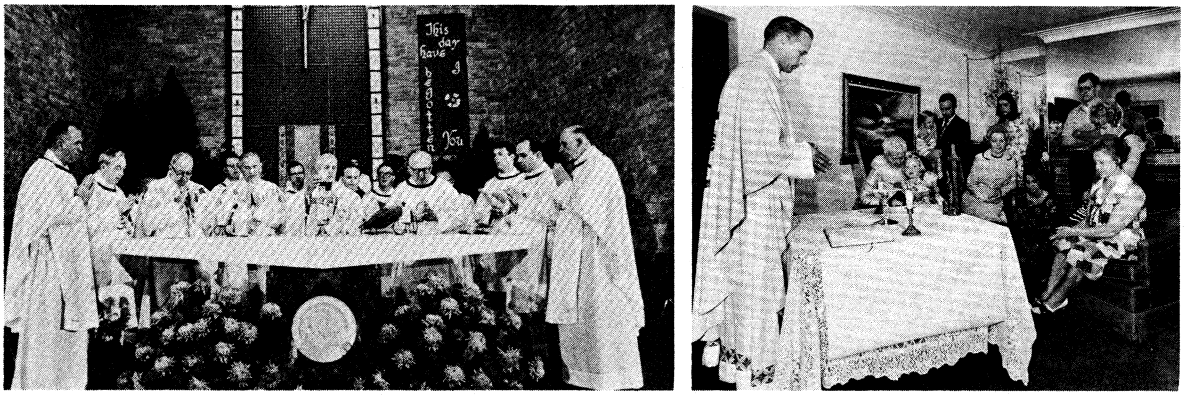
Confirmation. Confirmation since the 11th century has been conferred by the bishop through the anointing with oil and the imposition of hands; the words are a declaration that the Holy Spirit is conferred. This is an echo of the accounts in the Acts of the Apostles (chapters 8 and 19) in which a distinction is made between Baptism and the conferring of the Spirit. In Acts, however, the reception of the Spirit meant the reception and the manifestation of charismatic gifts (*e.g.*, prophecy, speaking with tongues, ecstasy); something else is now meant. Confirmation is normally conferred at or near the beginning of adolescence. The modern liturgical renewal has empowered pastors of parishes to confer confirmation.

Neglect of the theology of confirmation has left some ambiguities. The Oriental churches confer it on infants as a part of the initiation rites of Baptism. The postponement of confirmation has led many Roman Catholic theologians to interpret it as a rite of passage from childhood, like the Jewish Bar Mitzwa ceremony; such rites of passage are common in tribal cultures. Early Christian Baptism, however, was conferred on adults; the catechumenate was the period of "immaturity." It seems that there should be a return to the theology of the Spirit and a consideration of confirmation as the sacrament that empowers the Christian to take an active part in the church. The traditional Roman Catholic view of the laity as passive has contributed to the neglect of the theology of confirmation; it left no room for a charismatic laity.

The Eucharist. The Eucharist (the Lord's Supper, Holy Communion) is with Baptism one of the two sacraments most clearly found in the New Testament; most Christian churches have it in some form. The Roman Catholic Church distinguishes the Eucharist as sacrifice (mass) and sacrament (Communion).

The formula of institution of the Eucharist and the command to repeat it are found in the three Synoptic Gospels (Matthew, Mark, Luke) and in Paul. Originally, the Eucharist was a repetition of the common meal of the local group of disciples with the addition of the bread and the cup symbolizing the presence of Jesus. Even in the 2nd century the meal became vestigial and was finally abandoned. The Eucharist was originally celebrated every Sunday; by the 4th century it was celebrated daily. The eucharistic formula was set in a framework of biblical readings, psalms, hymns, and prayers that depended in form somewhat on the synagogue service. This remained one basis of the various liturgies that arose, including the Roman rite.

The sacrificial character of the Eucharist was determined by its relation to the death of Jesus. This is not seen as sacrificial everywhere in the New Testament, but the theme is so clearly elaborated in the Letter to the Hebrews that it is universally accepted as Christian belief. The Protestant churches denied the sacrificial character of the Eucharist and rejected the mass. Roman Catholic theology has never reached a universally accepted theory explaining the connection between the death of Jesus and the mass, but it has firmly insisted that the mass repeats the rite that Jesus told his disciples to repeat and that the rite is an effective symbolic commemoration of his death. The mass is the only act of worship that the Roman Catholic Church imposes upon its members. Historically,



Two innovations now in practice in the Roman Catholic mass. (Left)The concelebrated mass; (right)mass in a private home.
(Left) Religious News Service, (right) Algimantas Kezys, S.J.

the Roman Church has attached great importance to the mass, conceding almost anything to secure its celebration.

Roman Catholicism believes in the "Real Presence," and this has dominated Catholic–Protestant controversies about Holy Communion. Protestant belief can generally be called dynamic as contrasted with Catholic realism. The celebrated term transubstantiation is defined as the change of the substance of bread and wine into the substance of the body and blood of Jesus Christ. Protestants believe that Jesus is experienced as present. The Roman Catholic theory is difficult to explain in terms other than antiquated Aristotelian physics, and recent theories, not yet successful, have attempted to explore sacramental symbolism in other ways. The realism of belief in the presence is associated with the Roman Catholic practice of distributing only the bread to the laity, a serious modification in the sacramental sign. Not yet universally restored, Communion under both species has become much more common since the second Vatican Council.

Sacrament of unity

Neither in Roman Catholic nor in Protestant eucharistic practice does the sacrament retain much of the symbolism of Christian unity, which it clearly has in the New Testament. Originally, the symbolism was that of a community meal, an accepted social symbol of community throughout the whole of human culture. Roman Catholic efforts to restore this have included the use of the vernacular and the active participation of the laity. Furthermore, the ancient rite of concelebration—*i.e.*, several priests or bishops jointly celebrating a single eucharistic liturgy—was restored by the second Vatican Council as a means of symbolizing unity; and the practice of celebrating the Eucharist in an informal setting—*i.e.*, in private homes or classrooms—was instituted in some places as a way of drawing the laity more intimately into the rite. But a great obstacle to the symbolism of unity remains the liturgical isolation of the celebrant and the passive silence that suited the atmosphere of mystery and the presence of God.

Church law obliges the Roman Catholic to receive Communion once a year (during the Lent–Easter season). Practice of frequency has varied over the centuries; the present law reflects the infrequency that was common in the Middle Ages. The symbolism of the sacrament as nutrition becomes rather feeble with such infrequency; it was rationalized both by the theology of the power of the sacrament and by considerations of the general unworthiness of Christians to receive it.

Penance. The name of the fourth sacrament, penance, reflects the earliest discipline of the penitential rite. Those who sinned seriously were excluded from Communion until they showed repentance by undergoing a period of public penance that included such things as fasting, public humiliation, the wearing of sackcloth, and other austerities. At the end of the period they were publicly reconciled to the church. There were some sins, called capital (murder, adultery, apostasy), for which certain local churches at certain times did not perform the rite; this did not mean that God did not forgive but that good standing in the church was permanently lost. Elsewhere, it was believed that the rite of penance could be performed only once; relapsed sinners lost good standing permanently. Rigorist sects that denied the power to forgive certain sins were regarded as heretical. The penitential rite did not endure beyond the early Middle Ages, and there can be no doubt that it was too rigorous for most Christians. It may also be noticed that the penitential discipline did not reflect the forgiveness of Jesus in the Gospels with all fidelity.

Confessing of sins

It is impossible to assign an exact date for "auricular confession"—the confessing of faults by an individual penitent to a priest—but it must have arisen in the early Middle Ages with the disappearance of the penitential system. This is the penitential rite that has endured into modern times. It was rejected by most of the Reformers on the ground that God alone can forgive sins. The Roman Catholic Church claims that the absolution of the priest is an act of forgiveness; to receive it the penitent must confess all serious (mortal) sins and manifest genuine "contrition," sorrow for sins, and a reasonably firm purpose of amendment. No quality or quantity of sin is too great for sacramental absolution. Roman Catholic theologians have not arrived at an explanation of the process of absolution. They do not admit that absolution is merely a recognition by the priest of dispositions on the part of the penitent that merit forgiveness nor that it is merely a process whereby the penitent is reconciled with the church. There seems to be an unspoken belief that it is a rare person who is really sorry for his sins and that the sacrament is a manifestation of the graciousness of God to human weakness.

Indulgences, which caused such a stir at the beginning of the Reformation, are neither instant forgiveness to the unrepentant nor licenses of sin to the habitual sinner. They are declarations that the church accepts certain prayers and good works, listed in an official publication, as the equivalent of the rigorous penances of the ancient discipline.

The anointing of the sick. This sacrament was long known in English as "extreme unction," literally rendered from its Latin title, *unctio extrema.* This non-English designation concealed the meaning of the Latin, "last anointing." It is conferred by anointing the sense organs (eyes, ears, nostrils, lips, hands, and formerly the feet and the loins) with blessed oil and the pronunciation of a formula. It may be conferred only on those who are seriously ill; seriousness is measured by the danger of death, but a danger, however certain, from external causes (such as the execution of the death sentence) does not render one apt for the sacrament. It may be administered only once during the same illness; recovery renders one apt again. Its effects are described as strengthening both of soul and body; it is an ancient rite that continues Jesus' ministry of healing. The sacrament is directed against "the remains of sin," an ill-defined phrase; but it was long ago recognized that illness saps one's spiritual resources as well as one's physical strength, and one is not able to meet the crisis of mortal danger with all of one's powers. In popular belief anointing is most valuable as a complement to confession or, in case of unconsciousness, as a substitute for it.

The anointing is not the sacrament of the dying; it is the sacrament of the sick. The New Testament passage (James 5:14–15) to which the Roman Catholic Church appeals for this rite does not envisage one who is beyond recovery. Postponement until the patient is critically ill in modern medical terms means that the sacrament is often administered to an unconscious or heavily sedated patient. Under such circumstances, the rite can no longer be effective as a sacrament of the sick, and to the uninformed a magical rite of forgiveness is suggested.

The claim of jurisdiction over marriages

Marriage. The inclusion of marriage among the sacraments gives the Roman Catholic Church jurisdiction over an institution that is of concern to the state and to other persons and groups within society. The Roman Church claims complete jurisdiction over the marriages of its members, even though it is unable to urge this jurisdiction in modern secular states. The sacrament in Roman Catholic teaching is administered by the spouses through the exchange of consent; the priest, whose presence is required, ıs an authorized official witness; in addition, the church requires two other witnesses. Marriage is safeguarded by a number of impediments that render the marriage null and void whether they are known or not, and the freedom of the spouses must be assured. This means that the Roman Catholic Church demands an unusually rigorous examination before the marriage, and this in turn means that it is practically impossible to marry on impulse in the Catholic Church. All of this is for the purpose of assuring that the marriage so contracted will not be declared null in the future because of some defect.

The rigid Roman Catholic rejection of divorce has been a major point of hostility in the modern world. Absolute indissolubility is declared only of the marriage of two baptized persons (Protestants as well as Catholics). The same indissolubility is not declared of marriages of the unbaptized, but the Roman Church recognizes no religious or civil authority except itself that is empowered to dissolve such marriages; this claim is extremely limited and is not used unless a Roman Catholic is involved. Because of its rigorous conditions for contracting marriage, the Roman Catholic Church finds grounds for nullity that do not exist in civil law, and it is willing to make a more searching examination. Declarations of nullity, however, should not be confused with divorce nor be thought a substitute for divorce. Some Roman theologians have suggested that Roman Catholic rigour is based on a misunderstanding of the Gospel texts that reject divorce; but a position maintained for centuries is not easily modified.

The onerous conditions that Roman Catholicism formerly imposed upon non-Catholic partners in mixed marriages have been notably relaxed since the second Vatican Council, particularly as regards written promises that the children would receive religious education in the Roman Catholic faith. The former coldness of the Roman Church toward such marriages is also relaxed; they may be celebrated in church during the mass, and a Protestant minister or a Jewish rabbi may share the witness function with the priest.

Power to administer the sacraments

Holy orders. This sacrament confers upon candidates the power over the sacred, which means the power to administer the sacraments. There are four minor orders (porter, reader, exorcist, acolyte) and four major orders (subdeacon, deacon, priest, bishop), but theologians regard only the last three as sacramental orders. In spite of this hierarchy of orders, the Roman Catholic Church maintains that holy orders is only one sacrament. The minor orders are anachronisms. They represent church services that are still rendered (except for exorcism, the casting out of evil spirits) but not by ordained persons. Ordination is conferred only by the bishop; the rite includes the imposition of hands, anointing, and the delivery of the symbols of the order. The power of the sacred peculiar to the bishop is shown only in the sacraments of confirmation and orders. Ordination can neither be repeated nor annulled. Priests who are suspended from priestly powers or laicized (permanently authorized to live as a layman) retain their sacred power but are forbidden to exercise it except in emergency. The priest is always ordained to a "title," meaning that he is accepted in some ecclesiastical jurisdiction.

Theological developments following the second Vatican Council concerned the ordination of women, against which no solid theological objection has been shown; the restoration of the permanent diaconate (with the powers to baptize, preach, and administer the Eucharist), to which both married and single men are admitted; and the idea of ordination for a fixed period of service. Except for the diaconate, these are radical suggestions in Roman Catholicism.

PARA-LITURGICAL DEVOTIONS

In the Roman Catholic Church, liturgy in the proper sense is the liturgy of the mass, the divine office, and the sacraments. The Latin language, the clerical character of the liturgy, and the search for novelty for hundreds of years have combined to produce forms of worship that are para-liturgical—by which is meant that they lie outside the liturgy and in some cases in opposition to it. These acts are also known as devotions or devotional practices, by which is meant that they are accepted voluntarily and not from obligation.

Benediction and "exposition"

Eucharistic devotions. A number of eucharistic devotional practices arose in the Middle Ages, when Catholics rarely received the Eucharist more than once a year. These were cultic forms that were directed to the Real Presence of Jesus in the Eucharist rather than to sacrifice and Communion. Such were Benediction of the Blessed Sacrament and "exposition." Benediction was a blessing conferred by a priest holding a consecrated Host in a vessel of display called the monstrance; the priest's hands were covered to signify that it was the blessing of Jesus and not his own. This blessing was accompanied by hymns and the use of the organ and incense. Exposition was the public and solemn display of the eucharistic bread, again with the accompaniment of hymns, the organ, incense, and processions. The reservation of the Eucharist in churches was a way in which Catholics could address themselves in personal prayer to Jesus really present. These have often functioned as substitutes for mass and Communion, and since the modern renewal of liturgy they occur much less frequently.

Cult of the saints. Other devotions revolve about the cult of the saints, a practice repudiated by the Reformers as a denial of the total mediation of Christ. This objection oversimplified Catholic practice, but the devotions did sometimes approach superstition. Catholic theologians distinguish (by Greek technical terms) the worship paid to God (latria, "adoration") from the veneration addressed to Mary (hyperdulia, "super-service") and the saints (dulia, "service"). Protestants do not disagree with the principle of admitting the saints as examples of genuine Christianity, but they reject the intercession of the saints as utterly superfluous and ineffective. The Roman Catholic understanding of the intercession of the saints is an extension of the belief in the communion of saints. Although such veneration does tend to multiply mediators, it has often fostered a simple and not unpleasing familiarity with the world of the supernatural. The excesses of the cult of Mary have stirred up controversy, and the tendency to superstition and the deification of Mary have sometimes been painfully present. Mary represents the feminine principle in Roman Catholicism; often in other religions this principle has been personified as a goddess. Mary is given the feminine traits of sympathy and tenderness that are not improper to the deity but are somewhat improper to the father figure and the king figure. The multitude of apparitions of Mary (*e.g.*, at Lourdes, France and Fatima, Portugal) come from the need of a local and national symbol of presence, which enables the Roman Catholics of a nation or region to identify with Mary. Because Mary as a historical person is almost totally unknown, Catholics have been able to find in her all the traits of the ideal person that they needed to find.

Roman Catholicism has always insisted on its right to official supervision of devotional cults, and only approved forms of devotions may be used in the churches or under clerical auspices. Approval does not imply the historical reality of the vision or apparition involved; no Roman Catholic is obliged to believe that Mary appeared to anyone at Lourdes or Fatima, that the rosary (prayer beads) was delivered by a private revelation, or that Jesus manifested himself as the Sacred Heart. Nor is any Catholic obliged to practice any of these devotions. Generally, they serve the purpose of emphasizing some element of Christian faith that is obscured in the preaching and the liturgy at a particular time and place. Devotion to the Sacred Heart, for example, turned the attention of Catholics to the humanity of Jesus and to Christian love in the somewhat arid spirituality of the 17th and 18th centuries. It may be urged that more authentic biblical proclamation would have brought out these things; Roman Catholicism has often manifested itself through devotions when authentic biblical preaching was not available. In approving devotions the Roman Catholic church simply declares that they are not in conflict with Roman Catholic faith and morals. It does not deny that they may be entirely products of the imagination.

Direct experience of God

Mysticism. The search for God through mysticism has never been received cordially by the official Roman Catholic Church. In general terms, the mystical experience can he described as a direct experience of the reality of the divine. A sufficient number of mystics have been proved fraudulent to justify caution but not to justify a blanket antecedent disapproval. Every saint who has been recognized as a mystic had some trouble with church authority. Indeed, one may see in the mystical experience of God something that the official church can neither furnish nor control. In addition, mystics have often had a prophetic character that expressed itself in criticism of abuses in the official church. Whatever the explanation, mystical phenomena have become extremely rare in the modern Roman Catholic Church.

VI. Practices

MISSIONS, EDUCATION, AND ELEEMOSYNARY ACTIVITIES

Missions. From its beginnings Christianity alone among the great religions has regarded itself as a true world religion that appeals to all men without distinction of race, nation, or culture. Roman Catholicism believes that it has preserved this missionary thrust more faithfully than any of the non-Roman churches. From the 4th to the 10th century the Roman Church devoted itself to the evangelization of the barbarians. The barbarians wished to become "Roman," and they accepted the church as a component of Roman civilization. The spread of Islām was met with crusades and not with missionaries, and the Roman Catholic Church has never mounted more than a feeble missionary effort toward Muslims. Thus, the missionary movement languished from the 10th to the 16th century; but the ages of the expansion of Europe, in which the Catholic countries were the early leaders, spread Roman Catholicism to the Americas, Asia, Oceania, and Africa.

Centralized control by the Roman see

This missionary effort differed from both the New Testament missions and the missions to the European barbarians in its very close, centralized control by the Roman see. Missionary churches have begun to achieve that independence proper to the diocesan structure only in the 20th century. It has been difficult for the Roman Catholic missions to divorce themselves from colonialism, and many missionaries did not want the divorce. Again until recent times most of the clergy and all the hierarchy in mission countries were European or American, as were the heads of educational and benevolent operations. Even the peoples of the mission countries, including their clergy and religious personnel, generally wished to give their church a European identity rather than an Asian or African identity. The Roman see, which had suppressed efforts to admit Chinese rites in the 18th century, was unsympathetic to what appeared to be "non-Roman" practices. The second Vatican Council officially ended the colonial phase of missions; in practice, however, the end will take longer. Where possible—meaning where the personnel are available—the operation of the mission churches has been given to native hierarchy and clergy.

Education. Between the barbarian invasions and the Protestant Reformation, education in Europe, except for the Arabic and Jewish centres of learning, was conducted by Roman Catholicism. Learning during the early Middle Ages was preserved by the monasteries; and, although the monks did little more than copy the manuscripts of Greek and Latin pagan writers and the Church Fathers, they educated the few people who had any learning. The foundation of the European universities after 1200 was also the work of Roman Catholicism; these institutions were stimulated by the learning of Arab scholars, through whom Europeans learned the philosophy of Aristotle and produced the learning of Scholastic philosophy and theology. The cultivation of literature and the arts in the 15th century flourished under the patronage of the papacy and Catholic princes and prelates.

The birth of modern science was coincidental with the Reformation and the age of the expansion of Europe. The Roman Catholic response to the new science, accompanied by new philosophical systems, was hostile; and the world of European learning after 1600 was dissociated from the Roman Catholic Church, which patronized only defensive learning. At the same time, movements of general education among the poor began in this period under Roman Catholic auspices. The invention of printing had diffused education far beyond earlier possibilities, and the churches were all interested in reaching the minds of the young. This interest was matched after the French Revolution by the modern states, which in the 19th century moved toward the exclusion of church influence from education. But the Roman Catholic Church, through its religious communities, was a pioneer in the elementary education of the children of the poor.

Critical problems for Catholic schools

In the 20th century the Roman Catholic educational endeavour in many European and American countries, particularly in the United States, had become a vast enterprise. In the second half of the 20th century, mounting costs and diminished religious personnel created critical problems for Catholic schools, and even their survival was at stake in many regions. The problems were not lightened by the realization that Roman Catholic education, even where it was strongest, reached only a minority of Catholic students; and the Roman Church had to face its established reputation as an adversary of the intellectual freedom that the modern academic world cherishes.

Eleemosynary activities. Institutional benevolence to the poor, the sick, orphans, widows, and other helpless has been characteristic of the Christian Church from its beginning. It involved organized assistance, supported by the contributions of the entire community and rendered by dedicated persons. The church in this way fulfilled the duty of "the seven corporal works of mercy" mentioned in the Gospel According to Matthew (chapter 25) and carried on the healing mission of Jesus. Protestant churches continued the works of institutional benevolence after their separation from the Roman Church, and the history of Christian benevolence is a noble portion of church history. Institutional assistance to the helpless is a legacy from the church to modern governments.

This work, which would seem to be above criticism, was beset with troubles in the latter half of the 20th century. Costs for these works, like the costs for education, soared beyond the possibilities of individual contributions. The assumption of benevolence as a government responsibility both rendered the necessity of church works doubtful and narrowed the base of contributions. Church organizations as they existed were not well equipped to deal either with modern urban poverty or with the problem of international poverty.

SOCIOPOLITICAL VIEWS AND PRACTICES

Church and state relations. The most important modification in Roman Catholic theory and practice of church–state relations was the declaration of the second

Recognition of the pluralistic state

Vatican Council in which the Roman Catholic Church recognized the modern, secular, pluralistic nation as a valid political society. Union of church and state had dominated the history of the Roman Catholic Church since the era of Constantine, and all pontifical declarations of the 19th century rejected separation of church and state as pernicious teaching. This position was steadfastly maintained in spite of the fact that the union of church and state had been accepted by the Protestant countries of Europe; it reflects a long history of domination of the church by the state and of the church's involvement in political power struggles. The second Vatican Council declared that the Roman Catholic Church is not a political agent and will not ask political support for ecclesiastical ends. A significant change in the Roman attitude towards the state is the council's express declaration of freedom of religion. (For further treatment see CHURCH AND STATE.)

Economic views **and** practice. In the centuries when the Roman Catholic Church was Christendom, there was a place for every member that corresponded to his place in the social structure. In modern times the hierarchy became identified with the landed aristocracy; this dangerous identification led the Revolutionaries of 18th-century France to attempt to destroy the Roman Catholic Church with other components of the old order. The Roman Catholic Church entered the 19th century with a firm official bias against revolutionary movements, and the brief liberalism of Pius IX was ended with his experiences in the Italian revolution of 1848. The Roman Catholic Church was inflexibly opposed to all forms of Socialism, and its opposition to Marxist Communism was implacable. Thus, the Roman Catholic hierarchy was identified with the new capitalist classes of the industrial society. In many European countries this meant that it lost membership among the working classes. Leo XIII in *Rerum Novarum* (1891; "Of New Things") was the first pope to speak against the abuses of capitalism. Social teaching was further elaborated by Pius XI in *Quadragesimo Anno* (1931; "In the 40th year"), John XXIII in *Mater et Magistra* (1961; "Mother and Teacher") and *Pacem* in *Terris* (1963; "Peace on Earth"), and Paul VI in *Populorum Progressio* (1967; "The Progress of Peoples"). Catholic opposition to Socialism has gradually been diminished, although Catholic teaching tends in the direction of the diffusion of capital and not in its nationalization. In some features, however, such as the recommendation of labour unions, Catholic teaching has reached points that Pius IX would have regarded as Socialism.

In its own practices the Roman Catholic Church has accepted the ownership of property and of productive investments. It has admitted no responsibility to the laity for its funds, which are managed by the hierarchy; hence, the wealth of the Catholic Church has long been a mystery, often attractive to greedy anticlerical governments. Their raids as well as some public disclosures indicate that popular belief exaggerates the wealth of the Catholic Church. Following the second Vatican Council, there was a strong movement in Catholicism for public financial reports.

Stability of the family

The family. Roman Catholic teaching on the family is conservative and attributes to the family a social and moral centrality that many people think it no longer has. The teaching has grown up around a number of factors, not all of which come from the New Testament. The medieval family (whether nobles or commoners) is seen in the guiding principle of Roman Catholic teaching, the stability of the family. The stability of the family forbids divorce. It is preserved by a strong authority structure in which the father is the head; this reflects not only the Old Testament but Roman law. The family is child-centred; traditional Catholic teaching makes the primary end of marriage the procreation and rearing of children. Only recently have Catholic theologians begun to speak of mutual love as an end "equally primary." In earlier cultures, marriages were of concern to the two families involved; and, if they agreed, the social purpose of the marriage was as well fulfilled without love as with it. Rigid monogamy was not unrelated to the common and widely tolerated practice of adultery, which the Roman Catholic Church regarded as more tolerable than divorce. The stability of the family also presupposed a certain tribal view of the family; one's chief security was not in the law and the courts but in one's kinsmen. Such a view of stability is not well adapted to the mobility of the modern family. Nor is it well adapted to the independence possible to the person who wishes it badly enough. In the early 1970s the Roman Catholic Church was faced with the problem of preserving for its members the unquestioned values of mutual love and responsibility without imposing an antiquated authoritarian structure. But the major problem was certainly the practice of birth control. The moral arguments for the Catholic position had suffered general erosion, and many Catholics regarded the declaration of Paul VI in 1968, reiterating the traditional prohibition of birth control, as a simple exercise of authority.

VII. Roman Catholicism following the second Vatican Council

The Roman Catholic Church has been experiencing a renewal that reached its official peak in the second Vatican Council. Renewal has brought benefits, but it has also brought internal disturbances greater than any the church has known since the Protestant Reformation. There has been a clear polarization between liberal and conservative wings of the type that tends to leave no room for moderates. Such disunity is a real threat of schism, but there have been no group departures except in a few instances. The number of individual departures has been large enough to cause concern, but their number is unknown; discontented Catholics in modern times leave quietly.

The Roman Catholic Church has officially abandoned its "one true church" position. It has entered ecumenical conversations with the Protestant churches that could lead to Christian union; the Catholic Church has expressed a readiness to make doctrinal and disciplinary concessions, but how far these may go is not yet clear. The church has even made gestures of friendliness to Islām and Jewry and does not speak of the great Oriental religions as simple paganism. The openness of the Catholic Church toward social movements has been mentioned; this has taken a surprising form in some unexpected places such as Spain and Latin America. The edge of Catholic opposition to Marxism has been taken off, and the Roman see has been engaged in unobtrusive diplomatic conversations with some Communist governments.

Problems within the church

Problems are, however, more in evidence than progress. The long, latent conflict between hierarchy and lower clergy has become open. Priests are resistant to the traditional total obedience in style of life and ministry. This conflict has come to a focus in clerical celibacy; without sure statistics it is a reasonable assumption that at least half of the Catholic clergy wish celibacy to become an option. The discontent with life and ministry has led to a large number of losses in the priesthood and in religious communities, some of which face the possibility of extinction. Much of this discontent revolves around ministry as much as around life style; many religious workers feel that the conventional ministries are not reaching enough people and are not touching their most urgent needs. The desire to work "in the world," while hardly alien to the New Testament ministry, is not adaptable to traditional clerical and religious rules. What might appear to be a minor point in some places has become major; priests and religious (women religious in particular, who have had more of a problem) no longer wish to wear the identifying garb; they believe that it immediately places an obstacle to personal relations. Actually, there is a widespread but not explicit, perhaps not even recognized, rejection of the traditional use of authority and obedience in Roman Catholic clergy and religious communities.

Roman Catholic liturgy has been profoundly changed. The results have not been altogether satisfactory, and some observers say that the effects of the new liturgy cannot be assessed until a new generation has grown up that knows no other liturgy. On this point minor local schisms have occurred, led by reactionary Catholics. Others find the new liturgy stodgy; but the degree to which liturgy ought to be exciting has never been known.

The place of the laity, like the place of the clergy, in church decisions remains uncertain. Bishops, clergy, and laity generally are timid in undertaking a modification in church government for which nothing in their previous church experience has prepared them. They seem to hesitate to employ their experience in government and business, where shared responsibility is the rule rather than the exception. Many Roman Catholics find it difficult to examine the role of their hierarchical officers without also questioning their credibility. Yet the direction of the movements where the problems lie is toward greater responsibility of each member of the Catholic Church—hierarchy, clergy, and laity, each in its own way.

BIBLIOGRAPHY. The best and most recent reference works are the *New Catholic Encyclopedia,* 15 vol. (1967), which treats all phases of Roman Catholicism and topics related to the study of Roman Catholicism; and *Sacramenturn Mundi: An Encyclopedia of Theology,* ed. by KARL RAHNER *et al.* (1968), which deals with Catholic doctrine and theological thought. An excellent brief compendium of doctrine is *De Nieuwe katechismus* (1966; Eng. trans. by KEVIN SMYTH, *A New Catechism: Catholic Faith for Adults,* 1967), the celebrated and controversial Dutch catechism. Roman Catholic theology of the church is discussed by HANS KUNG in *Die Kirche* (1967; Eng. trans., *The Church,* 1967). The contemporary Roman Catholic Church is surveyed by JOHN L. MCKENZIE in *The Roman Catholic Church* (1969).

(J.L.McK.)

Roman Catholicism, History of

Historians generally agree that, in the first half of the 11th century, a new epoch began in the church of the West. Though no modem scholar would hold that Western Christendom suddenly became Roman Catholic or that the Roman Church preached a new creed at that time, it was certainly then that the papacy reorganized itself and proclaimed a doctrine of authority—traditional indeed in its elements but new in its logical and comprehensive dynamism—and, by asserting rights half-forgotten or ill-defined, gathered under its direct and complete control the framework of the Western Church. The main characteristic of that church, from the 11th to the 20th century, is that of a body of Christians claiming to be the only church body with a right to the title of the authentic, apostolic church of Christ and basing this claim and this unity on a recognition of the bishop of Rome, or pope, as the successor of St. Peter and head of the church, with jurisdiction over all its members. Though epochs, eras, and periods are artificial divisions of the continuous stream of history and each so-called revolution is foreshadowed in its causes and imperfect in its effects, some moments of change nevertheless are decisive, and in the 11th century the structure of the institutional church, the ideas of its leaders, and the monuments and records of their thoughts and actions all have a character markedly different from what had gone before. The present article is concerned with the history of this church, which in fact, between 1000 and 1517, contained within its obedience all the professed Christians of Europe, west and north of the domain of the Orthodox Church in Russia, the Balkans, and Greece, save for a few numerically minute groups of Christian "heretics," and (for a time) Bohemia in the 15th century.

This article is divided into the following sections:

I. The Latin Church of the West (1000–1517)

THE CHARACTER OF LATIN CHRISTIANITY IN THE 11TH CENTURY

The church and the social order. At the beginning of the 11th century, Latin Christendom was a rather poorly organized body. Though a large part of Italy was within the papal sphere of control, the degraded papacy—after over a century of occupancy by several weak or immoral popes—remained inactive, and its jurisdiction was overlaid, partly by powerful lay rulers at home and partly by the kings or emperors of Germany, who controlled many ecclesiastical activities within their own territories. France was fragmented into many feudal domains, but this allowed the ecclesiastical hierarchy there a certain independence and cohesion, while the growth of the French reform-oriented monastery at Cluny laid the country open to the message of reform when it finally came. In England there was a unique intermingling of ecclesiastical and royal administration that, in fact, left the church entirely free. On the fringes of Christendom—Scandinavia, Scotland, Ireland, and northern Spain—there was little hierarchical development.

The condition of the papacy

Although the papacy had virtually no practical significance for the Christian population and its priests, Rome was everywhere recognized by the bishops and other leaders as the ultimate source of doctrine and discipline; her permission was traditionally recognized as necessary for the creation or transference of a see, and archbishops sought from Rome the pallium (a mantle that symbolized their office and jurisdiction). Moreover, popes, even of the most decadent period, possessed the archives of the Roman see and issued on request grants of immunity and confirmation of property. The city of Rome also remained a goal of pilgrimage for all the western European peoples. In short, papal jurisdiction, though largely dormant, was not challenged save when, in German lands, the king or emperor claimed or tacitly exercised superior powers. The liturgy, administration of the sacraments, and general canonical (church law) principles were basically identical everywhere, though modified in places by local peculiarities or customary law or (as with clerical celibacy) inveterate contrary practice.

In the year 1000 the church in continental Europe—northern Italy, the German lands, France, and northern Spain—lay under a regime of feudal organization and lay ownership. This last had gradually replaced the ecclesiastical ownership of the late Roman Empire. Whereas Roman law recognized a church as a legal *persona*, the customary law of the invaders (such as the Germanic tribes) assumed that all upon the land belonged to its lord. Churches were treated as private real property, and the priest, appointed by the lord, received a part only of the revenues. Simultaneously, the feudal system, now accepted almost everywhere save in Germany, regarded bishoprics and abbeys as fiefs (feudal estates) or benefices (ecclesiastical estates), bestowed by the ruler in return for loyal service or—as in part of Germany—the private possession of lay proprietors. At the highest level of all, the papacy, which for a century had been the pawn of Roman political factions, was regarded by the newly founded German Empire (the Holy Roman Empire, founded by Charlemagne in 800) as an endowment over which it had rights.

The Ottonian Empire. German kingship had entered upon a new epoch in the 10th century, and, under Otto I, the Great, the bishops and greater abbots were drawn into royal service and enriched with estates and counties, for which they did feudal homage. Otto conquered northern Italy and extracted from the pope an imperial coronation (962). Both he and his grandson Otto III regarded the papal territory as part of their realm; they appointed and removed popes and presided at synods.

Otto III, an enlightened ruler, appointed as pope his old tutor, Gerbert of Aurillac—who took the name Sylvester

11—whose brief reign (999–1003) was a shaft of light between two periods in which Roman factions dominated the papacy.

Otto's successors continued to keep control over both the church and the degraded papacy. Emperor Henry III was a capable reformer who appointed energetic bishops in his dominions, but a new age began when the Emperor appointed his relative—the non-Italian Bruno, known as the reforming bishop of Toul—as Pope Leo IX (1002–54) in 1049.

Forces of reform

Meanwhile, the forces of reform had been building up in Italy, beginning with the monks from Cluny and Greeks from southern Italy and developing into the austere reform orders at Camaldoli and Vallombrosa (Italy). The reform found active expression among a group of monks who were related by their zeal to purify the church, including the Italian hermit–monk and cardinal, St. Peter Damian; the Lotharingian cardinal, Humbert of Moyenmoutier; and Hildebrand, archdeacon of the Roman Church and later Pope Gregory VII. This great reform movement began as a moral campaign; the two enemies were nicolaism (marriage and incontinence of clergy) and simony (the purchase of ecclesiastical office). The remedies proposed were a monastic discipline and piety and the code of the ancient canon law revived; the instruments were to be a reformed and powerful papacy and a hierarchy freely elected.

Popular Christianity. By 1050 the population of Europe west of the German marches and the hinterland of Scandinavia was Christian, though pockets of paganism and many heathen practices remained. Of this Christendom, the greater part had been divided into bishops' dioceses and individual parishes, though in the northern and western regions the proliferation of small private churches had not yet been wholly absorbed, and the existence of proprietary and exempt enclaves continued to the Reformation and beyond. The priest, in rural districts usually a villein of the lord (subject to the lord but not to others), cultivated his acres of glebe (revenue lands of the parish church), celebrated mass on Sundays and feasts, recited some of the hours (liturgical or devotional services for use at certain hours of the day, according to the monastic daily schedule), and saw that his flock was baptized, anointed, and buried. Lay people normally received communion four times a year—Christmas, Easter, Pentecost, and Assumption (August 15). Auricular (privately heard) confession was widespread but not universal.

Education in 1000 was at a very low ebb outside the monasteries. Cathedral schools were few, and rural priests who could read Latin easily were rare. Almost all literary work came from the monasteries and, in Celtic lands (mainly Ireland), from the half-monastic Culdees (religious recluses). The larger monasteries, such as Cluny or St. Gall (Switzerland), were towns in miniature with social services; they were also the only reservoirs of learning and artistic skill. On the land, pious practices and beliefs often merged into superstition or "white" magic; and marriage customs, together with the complicated degrees of prohibited relationships, provided endless problems in an epoch when the presence of a priest was not necessary for a valid union. In an age of protective lordship, heavenly patrons were highly valued, and the body or relics of a reputed saint made him the *persona,* a quasi-living protective presence, of a church or abbey. This aspect of belief explains the popularity of pilgrimages to shrines such as that of the apostles at Rome, St. James of Compostela (Spain), the Magi at Cologne (Germany), and countless others. Monastic piety, in addition to the liturgy, was expressed in "little offices" (liturgical or devotional services) of the Blessed Virgin, of the cross, of all saints, and of the dead; and the primary reason for a monastery's existence was intercessory prayer—hence the numerous monastic foundations by royal and noble families.

THE INVESTITURE CONTROVERSY (1049–1122)

The first reformers: Leo IX and Nicholas II. Leo IX was the first pope to impress his authority upon the church in general; he achieved this policy by a tactic of lengthy tours beyond the Alps, punctuated by synods, in which decrees both dogmatic and disciplinary were passed. He also began the practice of appointing non-Romans to curial (papal administrative) posts and sending legates (papal representatives) to carry out his decrees. A man of great energy and spiritual purpose, he must nevertheless bear the responsibility for a disastrous war that ended in capitulation to the Normans and for choosing the rigid and violent Humbert for the mission to Constantinople in 1054, the year from which the Schism between the churches of the East and West is dated. In the confused years that followed, the papal election decree of Nicholas II in 1059 stands out: it gave the right and duty of papal election to the cardinals, tacitly eliminating the king of Germany. The same pope shortly afterward renewed earlier decrees on simony and clerical celibacy but avoided the issue of pope and empire.

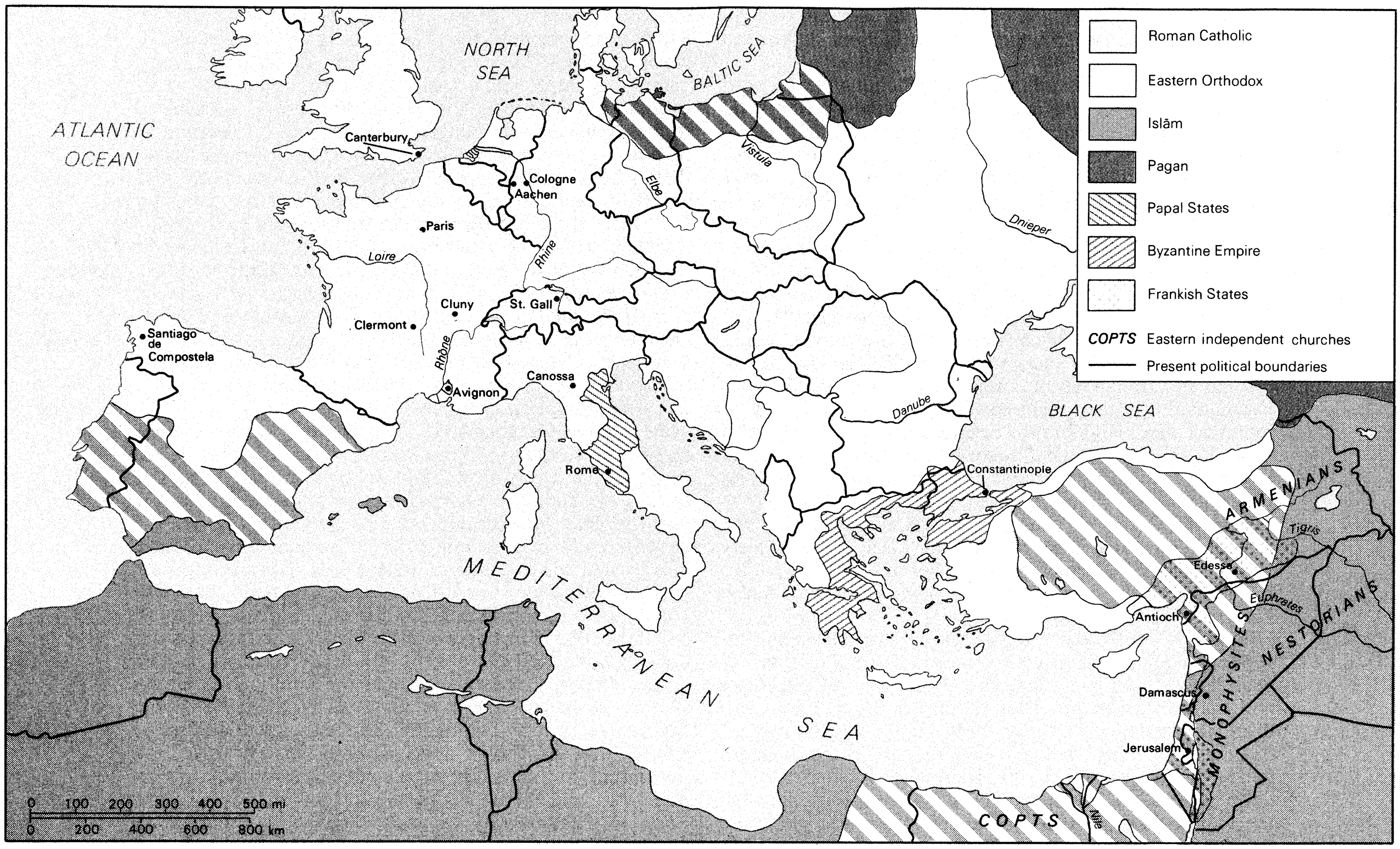
The Schism between the Eastern and Western churches

The reign of Gregory VII. Hildebrand, who succeeded in 1073 as Gregory VII, proved to be one of the greatest of his line and had more influence than any other person of his time upon the external fabric of the church. In his long struggle with the German king, Henry IV, he suspended and excommunicated his opponent, pardoned him as penitent at Canossa, Italy (1077), excommunicated him again (and was himself twice deposed), and was finally driven from Rome by Henry to die in exile at Salerno (1085). In opposition to Henry's claim to be the divinely appointed vice regent of Christ over the activities of the church, Gregory presented the unlimited commission of Christ to Peter over all souls (Matt. 16:18–19). Beneath these lofty claims lay the resistance of the ruler, who did not want to be deprived of his ancestral right of appointing to office his most influential subjects (who often also held the richest fiefs), and the insistence of the Pope on the authority of ancient canon law and papal decrees. If the King's claims were inconsistent with the current conception of a free church, the Pope's claim and actions were without precedent within the memory or records of his age.

Gregory was defeated by force of arms, but his principles had-been established: Even more directly influential was Gregory's centralization of the church. Plenipotentiary legates (representatives with full power to negotiate); the immediate control of diocesan bishops, canonical elections, and Roman and local synods; the publication of canonical collections and polemical manifestos—all combined to fashion a web in which every thread led to Rome. The scattered priests and the distant bishops were gradually becoming a class, the clergy, distinct from others and with a law and a loyalty of their own. Gregory died a lonely exile, yet his principles of reform had found reception all over Europe, and the new generation of bishops was Gregorian in sympathy and obedient in practice to papal commands, in a way unknown to their predecessors.

The Investiture Conflict (1085–1122). The efforts of the reformers to make the church independent of lay control inevitably centred upon the appointment of bishops by the ruler of the country or region. In ancient canon law, election of bishops had been by clergy and people; entrance upon office followed lawful consecration. Feudalism and royal claims had transformed election into royal appointment, and admission to office was by means of the bestowal, or investiture, by the lord of ring and staff (symbols of the episcopal office), preceded by an act of homage. This savoured of simony, both because a layman bestowed a spiritual benefice and because money was often offered or demanded. The conservatives appealed to immemorial practice, accepted and even enjoined by the papacy.

Gregory VII, though asserting the principle of freedom, was in fact tolerant of royal appointments free from simony. Pope Urban II (reigned 1088–99) was equally ambivalent, though in other ways he was a reformer. Pope Paschal I1 (reigned 1099–1118) at once condemned lay investiture, thus precipitating the crisis in England between Anselm, archbishop of Canterbury, and King Henry I. This and a similar crisis in France were settled



The extent of Christianity during the period of the Crusades.
From F.W. Putzger, *Historischer Weltatlas*

by a compromise. Election (by the cathedral chapter) was to be free; investiture was waived, but homage before the bestowal of the fief was allowed. Meanwhile Paschal, at issue with the German king Henry V, who was demanding imperial coronation, suddenly offered to renounce all church property held by the king if investiture were also abandoned. Henry accepted, but the bishops refused the terms; thereupon the Emperor seized the Pope who, under duress, allowed investiture. By this time, however, a large majority of the bishops were Gregorians, and the Pope was persuaded to retract. Eleven years later, Pope Gelasius II provisionally accepted the Concordat of Worms (1122). According to this agreement, free election by ecclesiastics was to be followed by investiture (without staff and ring) and homage to the king.

The Concordat of Worms

This ended a strife of 50 years, in which pamphleteers on both sides had revived every kind of claim to supremacy and God-given authority. Nominally a compromise, the concordat was in effect a victory for the monarch, for he could in fact usually control the election. Nevertheless, the war of ideologies had exposed the weakness of the emperor who, in the medieval context, had in the last resort to admit the spiritual authority of the pope, and the struggle left intact the claim of the church to moderate the whole of society.

The Crusades. The authority of the papacy and the relative decline of the empire also became clear in the unforeseen emergence of the Eastern Crusades as a major preoccupation of Europe. The idea of a holy war, blessed by the pope, had encouraged volunteers in the reconquest of Spain and had been exploited by William the Conqueror in his invasion of England (1066). Moreover, the papacy had been stirred more than once by the disasters that befell Eastern Christians, such as their defeats by the Seljuq Turks at Manzikert (1071) and Antioch (1085) in Asia Minor. The Byzantine emperor Alexius I had appealed for help to Pope Urban II, and this may have been the decisive motive, though the advantages of diverting the Normans of Sicily and other turbulent warriors away from Europe to wage a sacred war were obvious. Urban's celebrated call to the Crusade at Clermont (France) in 1095 was unpredictably effective, placing the Pope at the head of a large army of volunteers. The capture of Jerusalem (1099) and the erection of a Latin kingdom in Palestine were balanced by disasters and quarrels, but the papacy had gained greatly in prestige. Though Germany as a whole remained aloof, a pope had for the first time stood out as the leader of European endeavour. The Crusades, with their combination of idealism, ambition, heroism, cruelty, and folly are essentially medieval and, as such, are outside modern man's experience, but they were part of the religious background for two centuries and were to add greatly to the anxieties, both spiritual and financial, of the papacy.

THE 12TH CENTURY

The Proto-Renaissance. The 12th century, or, more correctly, the century 1050–1150, has been called the first Renaissance. A more accurate title would be the adolescence of Europe, in which higher education, the techniques of thought and speech, and a fresh attack upon the old problems of philosophy and theology appeared for the first time in postclassical Europe. All these activities were exercised by clerics and controlled by churchmen. The focus was the cathedral school, and the new agency was the semiprofessional, unattached teacher, such as the French philosopher–theologians Berengarius, Roscelin, and Abelard, though monks still had a share, such as Lanfranc, Anselm of Canterbury, and Hugh and Richard of the Monastery of St. Victor, Paris.

Significance of churchmen in the first Renaissance

Philosophy was revived with logic and dialectic, which were applied to doctrines of the faith, either as formal exercises, Augustinian speculation, or critical reformulation. From 1100 onward theology, in the modem sense of the word (first used by Abelard), emerged. The teaching of Scripture and of the early Church Fathers on the various doctrines were consolidated and organized in works called Sentences. Then masters gave judgments and opinions, and the first handbook of theology was composed by Abelard. Finally, Peter Lombard (bishop *c.* 1159) published his Books of Sentences, which summarized the Christian faith, using the Sic et *non* (Yes and No) dialectic popularized by Abelard and the canon lawyers, and himself pronounced on vexed questions. His classic manual may be said, in modern terms, to have created the syllabus of theological study for the age that followed. Together with the expansion of logic—brought about by the arrival (through Muslim sources) of what was called the new logic of Aristotle—and the emergence of the university, the Sentences ended the era of literary, humanistic, and monastic culture and opened that of the formal, impersonal, Scholastic age.

Proliferation of monasteries

Reformed monasticism. The most distinctive feature of the century 1050–1150, according to some scholars, was the appearance and diffusion of reformed monasticism. Beginning with a few relatively small quasi-hermit orders in Italy, such as the Camaldolese and the Vallombrosans, it spread to France with the extreme eremitical Grandmontines (founded in 1077) and the eremitical Carthusians (founded in 1084), and became as wide as Christendom with the multiplication of the daughter monasteries of Citeaux (founded in 1098). The keynote of the Cistercians (based at Citeaux) was exact observance of the Rule of St. Benedict, with emphasis on simplicity, poverty, and manual work. The addition of lay brothers tapped a large reservoir in an age of economic and demographic expansion, and the organization of the order—with annual visitations and a general chapter—ensured good discipline and enabled the order to accommodate itself to the strain of a vast family of houses, scattered throughout the Latin Church. The success of Citeaux owed much to the genius of St. Bernard, abbot of Clairvaux from 1115 to 1153, who was for 30 years the untitled religious leader of Europe. Owing to his influence, other new orders, such as the Premonstratensians, the English Gilbertines, and the military Knights Templars, accepted or imitated Cistercian practices. All these and others had a popularity that in any other age would have seemed miraculous, since they practiced austerity. By the end of the 12th century the saturation point for monasticism had been reached all over Europe, save in a few peripheral regions, and the golden age of monasticism had passed.

The papacy. Popes, often elected as elderly men, normally had a brief span of power, and the 12th century saw few of individual significance. In its early decades the papacy was neo-Gregorian in character; that is, papal policy continued that of Gregory VII, but with an emphasis on the political and legal side. There was a great increase in bulls (papal documents) confirming the possessions and privileges of churches and abbeys and a gradual increase of appeals to Rome that led to the use of local judges who were delegated as fact finders and who presided over courts of first instance (first trials). Legates were sent out to visit and to hold synods, and every effort was made to ensure free elections of bishops, while the papal control of regional churches increased. At the same time the pressure against a married clergy continued with some success and the private ownership of churches by laymen decreased, through gifts to bishops and monasteries and through the foundation of religious communities in churches with large incomes.

In the second half of the century the spread of canon law increased rapidly, despite resistance from the newly organized royal administrations. There were few distinguished popes between Paschal II and Innocent III (reigned 1198–1216). Three, however, stand out: Eugenius III (reigned 1145–53), a disciple of St. Bernard, the holiest of the group; Adrian IV (reigned 1154–59), the Englishman, whose promising rule was cut short; and Alexander III (reigned 1159–81), whose pontificate was the longest and most influential. The first of the former university canonist popes, Alexander III settled many disputed points of discipline and sacramental practice. A succession of short-lived pontiffs followed him, who were unable to cope with the difficulties of their time.

The age of faith

Popular Christianity. The 12th century, perhaps more than any other, was an age of faith in the sense that all men, good or bad, pious or worldly, were fundamentally believers, and religious causes and interests (crusades, monastic foundations, building churches, and assisting education and charities) made up much of the life of the literate and administrative classes. Lay religion was, as never before or since, permeated with monastic ideals. Prodigious numbers of the populace became monks, knights (members of military-religious orders), labourers (lay brothers), and lay people who followed monastic rules, and the favourite lay devotions were short versions of monastic offices. Almost every church—whether cathedral, monastic, parochial, or private—was built or rebuilt between 1050 and 1200. Almost all baronial families founded a monastery, and townspeople not only paid for their cathedrals but often supplied materials and labour.

Heresy. Heresy on a large scale was unknown in the West before the middle of the 12th century. The early dissenters were often radical reformers such as the Italian canon, Arnold of Brescia (d. 1155), an outspoken critic of clerical wealth and corruption. Then there appeared in north ltaly and southern France the sect, Eastern and Manichaean in origin, later known as the Cathari (the "pure," from the ascetic lives of their leaders). This sect had an organization and liturgical life that imitated Christianity but that overtly denied many key doctrines, such as the incarnation of Christ, and was dualistic in that it regarded matter and the human body as evil and the spirit as good. Its emphasis on poverty and its genuine solidarity of mutual assistance appealed to many by contrast with the luxury and wealth of the Catholic hierarchy. A little later, another type of dissent appeared with the Waldenses (founded by a French reformer named Peter Valdes) of the Rhône Valley and Piedmont. These groups, basically and professedly orthodox, together with the reform-minded Humiliati of Lombardy (Italy), practiced poverty, scripture reading, and preaching. The Cathari were proscribed as heretics by the papacy, attacked by a crusade and later by the Inquisition, and gradually disappeared. The Humiliati remained orthodox as a quasi-religious order. The Waldenses, largely through mismanagement by the bishops, drifted away from the church and remained throughout the Middle Ages and beyond as a non-Catholic body. These heretical movements, together with numerous legal disputes between monks and bishops, and bishops and metropolitans (ecclesiastical provincial leaders), imparted a sense of decline and peril to the last decades of the 12th century, which were notably barren of saints and great men. The church was too rich and too set in its hierarchical ways to meet the demands of larger populations and economic stresses, especially in urban conditions. Reformers demanded a spirit of poverty and a fresh wind of spirituality.

THE 13TH CENTURY (1198–1303)

The significance of Innocent III

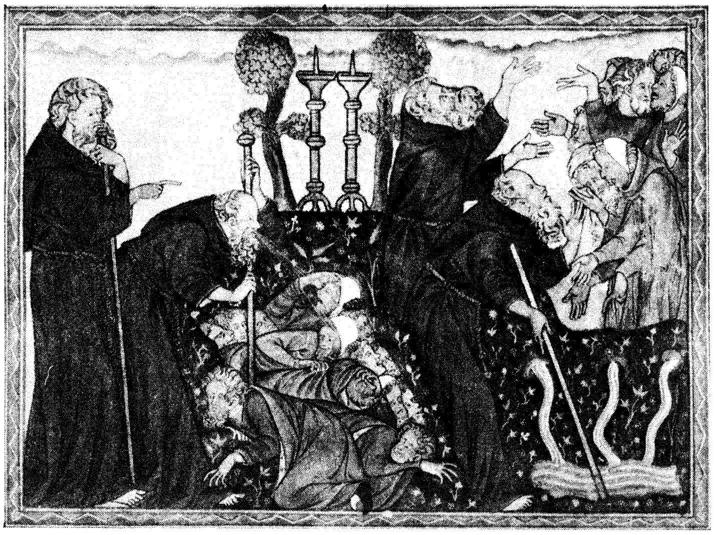
The papacy. Elected at the age of 37 and youngest of the cardinals, Innocent III (reigned 1198–1216) was one of the greatest of the popes, capable of realizing and satisfying the needs of his age. His program was the reform of Christendom, lay and clerical, and he attempted to influence, in Christian terms, its political life by papal action, direct and indirect. A legislator and reformer rather than a theologian and evangelist, lacking the charisma of sanctity yet a leader of genius and integrity, Innocent III came nearer than any medieval pope to being the acknowledged arbiter of Europe in all matters affecting men as Christians. He made mistakes: he condoned too readily the rape of Constantinople by the Crusaders (1204); he turned too lightly to force after failing to convert the Cathari; he misunderstood English affairs. But on the credit side must stand the great Lateran Council of 1215; the acceptance of the mendicant (begging) religious orders of St. Francis and St. Dominic; and a long series of wise judgments and decretals. His pontificate ended and began an epoch; it was in many ways a summit between them.

There followed a series of great administrators and lawyers: Gregory IX (reigned 1227–41), Innocent IV (reigned 1243–54), and Alexander IV (reigned 1254–

61), who practiced as popes what they had learned or taught in the schools. They exalted the powers of the papacy and curia in a centralized church and used spiritual sanctions to enforce conformity and later to obtain from their unwilling subjects the funds necessary for a government that was not only a great administration but also the political and diplomatic centre of Europe. By the midcentury, many were beginning to feel and to say that the papacy, which under Innocent III had guided the church wisely, was now, under Alexander IV, beginning to fleece rather than feed the sheep. The canon lawyers continued to exalt the powers of the pope as universal ordinary (jurisdictional leader), with a real if ill-defined right of control in political life. Boniface VIII (reigned 1294–1303), personally ambitious and intemperate, pressed his extreme claims against the first "secular" monarch, Philip IV (1268–1314) of France, and was overcome by force.

The mendicant orders: the friars. Below the level of the papacy, a spiritual revival took place. The pontificate of Innocent 111 saw the appearance of a totally new form of religious life, that of the penniless or mendicant friar. Francis of Assisi (1181/82–1226), a personality of magnetic originality who believed that he was called by Christ to preach poverty, bad no thought of founding an order; but his message and his genius exactly suited his age, and the vast concourse of his followers gradually changed from a homeless, penniless band of preachers and missionaries in Italy into an international body, governed by a single general and devoted to the service of the papacy. Dominic of Spain (c. 1170–1221,) on the other hand, with a vocation to preach doctrine to heretics and with followers keeping a canonical rule, changed his existing institute into one of friars. Gradually, through mutual influence and pressure of circumstances, the two groups became similar: international, articulated groups of men, bound to an order but not to a community. They took the customary monastic vows of poverty, chastity, and obedience but dropped the vow of *stobilitas,* stability, in favour of mobility, and were governed by elected superiors under a supreme chapter and general. Unpredictably, first the Dominicans and then the Franciscans entered and soon dominated the theological schools of Paris and Oxford. Two similar bodies joined them, the Carmelites and Austin Friars, and for almost a century the friars were the theologians, the preachers, and the confessors of the Christian people.

Published on Bodleian Lbrary colour filmstrip 164A



Mendicant friars preaching the gospel; manuscript illumination by an unknown artist, 13th century. In the Bodleian Library, Oxford (MS. Douce 180).

The golden age of Scholasticism. The 13th century was an age of fresh endeavour and splendid maturity in the realms of thought, theology, and art. Philosophy, hitherto almost exclusively devoted to logic and dialectic, had stagnated in the later 12th century. It was revived by the gradual arrival from Spain and Sicily of translations of the whole corpus of Aristotle's writings, often accompanied by Arab and Jewish commentaries and treatises. Aristotle, especially in his *Metaphysics* and Ethics, opened the whole field of philosophy to the schools. After a short period of hesitation they were used by theologians, at first eclectically, and then systematically. The great German philosopher and theologian Albert of Cologne (known as Albertus Magnus) and his more famous pupil Thomas Aquinas rethought the system of Aristotle in Christian idiom, pouring into it a fair dose of Neoplatonism from St. Augustine. Aquinas, in some 25 years of work, set theology firmly on a philosophical foundation. The Italian theologian Bonaventure (1217–74), in an even shorter career, renewed the traditional approach of Augustine and the Victorine monks, regarding theology as the guide of the soul to the vision of God. At the same time, masters in the arts school of Paris used Aristotelian thought to present a naturalistic system that clashed with orthodox teaching. The condemnations that ensued in 1272 and 1277, coinciding with the deaths of Bonaventure and Aquinas (1274), included some Thomist theses. This apparent victory of conservatism ended the long era in which Greek thought was regarded as right reason and foreshadowed the age of individual systems and the divorce of philosophy from theology.

Revival of philosophy

Ecclesiastical life. The coming of the friars and the legislation of the fourth Lateran Council in Rome (1215) —including requirements of annual confession and communion and a reduction in number of the impediments to marriage—saved for the church the lower classes of the population and silenced many of the critics of the establishment. Well-trained mentally and socially and extremely mobile, the friars were able to reach and hold sections that the static monks and clergy had failed to move. The 13th century in Europe as a whole was a time of pastoral endeavour in which bishops and university-trained clergy perfected the diocesan and parish organization and reformed many abuses. It was an age of active and spiritual bishops, many of them masters in theology and themselves friars. There also were controversies. The early friars served and were welcomed by the bishops and parish clergy, but clashes soon occurred; the papacy gave the friars exemptions and privileges so wide that the basic rights of the secular clergy were threatened. An academic war of pamphlets led to an attack on the vocation and work of the friars. A compromise was finally arranged by Boniface VIII (*c.* 1235–1303) that was just and workable, and under a revised form it lasted for two centuries. The bishop could refuse friars entry into his diocese, but once they had been admitted the friars were free from his control.

Troubles of the church. The last quarter of the 13th century was a time of growing bitterness and harshness. The golden age of Scholastic theology had come to an abrupt end. The troubles of the Franciscans—divided into those who stood for the absolute poverty prescribed by the rule and testament of Francis (the Spirituals) and those who accepted papal relaxation and exemptions, (the Conventuals)—were a running sore for 60 years, vexing the papacy and infecting the whole church. The Inquisition (the ecclesiastical tribunal instituted in 1229 to deal with heretics) and the papal court incurred odium for their inhumane and inequitable treatment of those suspected of heresy. Beneath the beauty of the Italian poet Dante's great poem, *The Divine Comedy* (begun in 1300), one can see a whole world of hatred and cruelty in the city-states of Italy. The papacy—held after about 1260 by a succession of popes of mediocre talent—was surrounded by a small group of feuding cardinals. The debacle of Boniface VIII at Anagni (Italy) in 1303, when he was made a prisoner by his enemies, was only the climax of extravagant claims—such as the bull Unam *Sanctam* (1302), which asserted papal supremacy—arousing equally violent resistance.

Persecution of the Jews

Another instance of hardening sentiment is seen in the treatment of the Jews. Between 800 and 1200 the Jews had multiplied in Lombardy, Provence, and the towns of the river valleys of the Rhône, the Rhine, and the Dan-

~ b eThey entered England only after the Norman Conquest (1066.) Everywhere they became wealthy and were used by governments that were in need of funds. Apart from heretics such as the Cathari, they were the only "foreign body" in Western Christendom, and as such served as an irritant to the ignorant and brutal. There were shocking massacres of Jews when the crusades were preached, especially in the Rhineland, and after various instances of panic on the part of Christians, Jews were accused of sacrilege and child murder. These, however, were all mob movements, resisted by kings and bishops. Later, the Jews suffered from suspicions that were aroused by the Cathari. The fourth Lateran Council gave the Jews a distinguishing badge and forbade their employment by governments. This established once and for all the ghetto system in large towns but did not at first impair Jewish prosperity. Later on, the growing class of Christian merchants became jealous and hostile, and in 1290 and 1306 the Jews were expelled from England and France. This swelled their numbers in Germany, thenceforward called "the classic land of Jewish martyrdom." Groups remained in Italy, and the Roman colony was never disturbed. In Spain, toleration gave way to widespread persecution and conversion under duress that left a heritage of sorrow for the future.

THE 14TH CENTURY (1303–78)

The intellectual enterprise. In the early decades of the 14th century, the Aristotelian abstraction of the essence from the thing perceived, which had been the philosophical method of the previous century, was challenged and the emphasis shifted to the individual, either as a "thisness" known by the mind (as taught by Duns Scotus) or as an object of intuition (as taught by William of Ockham). William of Ockham regarded all general terms, such as man or animal, as mere names (thus the term Nominalism) given by man's mind to his experience. He thus eliminated all natural theology—which claimed that man could know God through reason—and left no bridge between experience and divine revelation. The existence of God, he claimed, could not be proved. Causality and value were words only; the good was simply what God commanded. Ockham cut the Aristotelian link between philosophical statement and scientific observation and eliminated speculative theology, but his emphasis on the individual "thing" led to a gradual interest in scientific method and in nature.

Removal of the papacy from Rome

The Avignon papacy (1309–1377). After the death of Boniface VIII (1303) the papacy, through a series of accidental circumstances, settled at Avignon, a papal enclave in French territory, where it remained for 75 years. This, the Babylonian Captivity (named for the 70 years of the Jewish exile in Babylon in the 6th century BC), was bitterly denounced at the time and, until recent times, historians have concurred in this judgment. Viewed in the perspective of history, however, the popes of Avignon are now generally seen as personally devout and even reformist, with a sense of universal responsibility. They created a government machine of great complexity and efficiency, comprising a judicial system, chancery, and financial system that raised papal income to a higher level than ever before, and the monastic reforming decrees of Benedict XII (reigned 1334–42) endured until the Council of Trent (1545–63). Nevertheless, it was considered disgraceful that the bishop of Rome should live permanently in French territory, in a wealthy papal city surrounded by a small body of rich cardinals jealously divided along national lines; the Avignon residence of the popes was held largely responsible for the disasters that followed.

Attacks on the papacy. The disputes of the Franciscans, crystallizing finally upon the teaching of the Spiritual Franciscans that their absolute poverty was that of Christ, were harshly settled (1322) by the irascible octogenarian John XXII (reigned 1316–34), but a group of them, led by Michael of Cesena, general of the order, and William of Ockham, became bitter and formidable critics of the papacy. With them for a time was the Italian political philosopher, Marsilius of Padua, a Paris master who, in his *Defensor pacis* (1324), outlined a secular state in which the church was a government department, the papacy and episcopate human institutions, and the spiritual sanctions of religion relegated to a position of honourable nonentity. Between them, Ockham and Marsilius used almost all the arguments that have ever been devised against the papacy. Condemned more than once, Marsilius had little immediate effect or influence, but during the Great Schism of the papacy (1378–1417) and later, in the 16th century, he and Ockham had their turn.

Advocacy of a secular state

The apptoaching storm. With the papacy "in captivity" and Nominalism capturing the universities, Europe and the church entered upon an epoch of disasters, of which the Hundred Years' War between England and France (began 1337) and the Black Death (1348–49) were the most clearly seen by contemporaries. When the papal states in Italy had been recovered, Urban V (reigned 1362–70) returned to Rome, but only for a short time (1367–70). Gregory XI (reigned 1370–78) made the final move in 1377, but then an even greater disaster occurred immediately: the papal schism, in which there were usually two, but sometimes three, popes.

Christian life. For all this, Christian life in the first half of the 14th century changed little. Many of the largest parish churches of Europe date from this time, as do many popular devotions, prayers, hymns, and carols; also, many hospitals and almshouses were founded. Though the relations between the friars and the secular clergy had been canonically settled, friction still continued. The friars came under wider criticism for worldliness and immorality, but they still remained popular. Though heresy and antisacerdotal (anticlerical) sentiment became almost endemic in the cities of Belgium and the Netherlands, the 14th century was a period that produced some of the greatest mystical writers of the church's history: Johann Tauler and Jan van Ruysbroeck in the north, Catherine of Siena in Italy, the author of *The Cloud of Unknowing,* and Walter Hilton in England.

The spread of the faith. The missionary enterprise during the period 1000–1350 involved three principal fields of work: Spain, central Europe, and Asia. In Spain the absorption of the Mozarabic Church (the Arabic term for Spanish Christians under Moorish rule) and the reestablishment of Catholic practices was accomplished by Spaniards who followed the crusade ideal and by volunteers, partly monastic, from beyond the Pyrenees. In central Europe, Pope Sylvester II (reigned 999–1003) had founded the ecclesiastical hierarchies of Hungary and Poland. The space between these countries and Germany was gradually conquered and Christianized by neighbouring bishops and German missionaries. The Baltic lands were won by a mixture of preaching and the swords of the Teutonic Knights (a military monastic order) between 1100 and 1400. Purer in motive and magnificent in design were the efforts of the Franciscans and Dominicans in the Near and Far East. Both orders preached to the Muslims, and early in the 13th century the Franciscans were in Georgia (now in the U.S.S.R.) and Persia and the Dominicans in Syria. In midcentury the Franciscans penetrated Mongolia and established a church in China with an archbishop and ten suffragan bishops, and under John XXII there was a hierarchy in Persia. All this might well have endured, had not the last of the great invasions (1383), under the Turkic conqueror Timur, or Tamerlane, broken all links between Europe and the East. (M.D.K.)

THE LATE 14TH TO EARLY 16TH CENTURIES (1378–1517)

The Great Schism and conciliarism. The desire of Christendom for the return of the pope from Avignon to Rome was at last fulfilled by Urban VI (reigned 1378–89), unfortunately an extreme believer in the principle of papal power. Because the Romans had exerted pressure on the conclave in which Urban was elected, and because the French cardinals opposed both the return to Rome and the extremely impetuous reform demands of the new Pope, the election was declared invalid. Within the same year Clement VII (reigned 1378–94), a relative of the French king, was elevated as antipope; he returned

The origin of the Western Schism

to Avignon and there formed a second Curia. The Western Schism had begun.

The German kings had long since made Christendom familiar with the ominous figure of an antipope; Ockham had, indeed, put forward the thesis that there could certainly be several popes simultaneously—this would, however, have signified the destruction of the unity of the church. But never before had the split been so deep as now, for each pope had a successor upon his death, and Christendom divided itself into two allegiances of approximately equal size, one Roman, one Avignonese. The adherence of the Western countries to the one or the other was decided for the most part on purely political grounds —for example, England belonged to the Roman allegiance, France and Scotland to the Avignonese. The consequence was that antipathy to the papacy grew, and ecclesiastical punishments against sovereigns lost all power, since the two popes excommunicated and interdicted each other. The whole church was thus under excommunication and appeared to be disintegrating. A merely human institution would fail in trying to revive itself by its own force. The church (interpreted as a divine–human institution), however, was able to do so. It regained its unity—for unity is the very life of the church—although it was a lengthy and tiresome process. In the many writings that ensued (particularly the ones produced at the University of Paris), those that dealt with the possibility of "root and branch" reform—an idea familiar from the Defensor pacis—were more and more emphasized, especially those that held that a general council, as the supreme authority of the church, would be the means of re-establishing unity. After mutual stubbornness had brought to naught all negotiations between the Popes, the two parties of cardinals finally united, after 31 years of schism, in calling a general council at Pisa (1409). Here both of the reigning popes, Gregory XII in Rome, Benedict XIII in Avignon, were deposed and Alexander V (reigned 1409–10) was elected. This was an unheard of action, *i.e.*, for the council to sit in judgment on the popes.

But the problem was too deeply rooted; neither would yield. Instead of two popes, there were now three (Alexander V resided at Bologna). Alexander was followed by a very unworthy successor (John XXIII, who reigned 1410–15 and is not considered a legitimate pope), and the demand for a new council became stronger. The German king proved once again to be the supreme protector of the church. King Sigismund (1368–1437) was able to compel the summoning of a general council at a German city, Constance. When John XXIII tried, by shameful flight, to break up the council, the King kept it together.

The Council of Constance

Conciliarism. The Council of Constance (1414–18) was one of the truly great church assemblies; in it the Christian West once again presented itself as a single whole. The problem of unity was solved: two popes, John XXIII and Benedict XIII, were deposed; the third, Gregory XII, withdrew. Martin V (reigned 1417–31), a member of the Colonna family of Rome, was newly elected and generally acknowledged. (Ironically, after his election, he worked, first and foremost, against the Constance theory of the superior authority of the council over that of the pope.) But the second task of this council, that of reform, remained unfulfilled. And for precisely this reason the oppositional movements in the church, which were calling for reform, remained alive; at the same time, the conflict with the Bohemian reformer Jan Hus (c. 1370–1415) and his followers finally broke out openly. Constance was thus both the symbol of unity and the beginning of the great convulsions that grew from the culpable failure to satisfy the church's demands for reform. At Constance it was decided that a general council should assemble every ten years. These councils, however, which were indeed summoned (at Pavia in 1423 and Basel in 1431), remained unfruitful because they were burdened with national controversies and the attempts of various democratic elements to limit the papacy. In the structure of the church as it was instituted, there must be room for primacy (papal authority) and for collegiality (collective authority). In the antagonism between these two poles of the church, the council at Basel actually became schismatic; instead of thoroughgoing reforms, the actions of that council culminated in the formation of another—the last—antipapacy (Felix V, reigned 1439–49).

Eugenius IV (reigned 1431–47), with whom the Renaissance became established in Rome, tried to achieve unity with the Greek Church. In order to secure the help of the West against the encroaching Turks, the Greek Emperor and the Greek Patriarch at Constantinople signed an agreement, but Eastern churchmen would have none of it; after the fall of Constantinople (1453) to the Turks, it became only a scrap of paper. The winners in the schism of the Western Church were the sovereigns of Europe; the influence upon the churches of their countries that they acquired in this period paved the way for the politically-based religious decisions made in the Reformation. The great loser was church reform.

Heretical movements. Because a genuine religious reform within the church, bearing the stamp of sacrifice and deepened faith, thus failed to come about, there developed a new kind of reform that became a most severe attack upon the church—the Reformation.

Reform movements led by Wycliffe

In England. Prologues to the Reformation were the movements headed by John Wycliffe (*c.* 1330–84) in England and Jan Hus in Bohemia. Besides the biblical basis of their theology, the power of these movements was characterized by a national coloration, which was particularly prominent in that of Hus. The resistance that England had shown to Rome ever since Innocent III's interference in English affairs (*e.g.*, his prohibition of the Magna Carta of 1215) grew with the increasing national consciousness and was nourished during the Hundred Years' War by mistrust of the French pope resident in Avignon. In 1366 the English Parliament refused the pope feudal tribute. The preacher and professor John Wycliffe provided a theological basis for this anticlerical and anti-Roman resistance. He taught that the church had no right to worldly power and wealth but was subordinate to the state. In an extreme emphasis on the Augustinian doctrine of predestination (*i.e.*, that men are foreordained to salvation or damnation,) Wycliffe finally opposed monasticism, indulgences, and sacraments. There was, he claimed, no need of a professional priesthood; the papacy was unnecessary, indeed, it stemmed from the antichrist. Wycliffe found adherents in all strata of the population. There was a Roman condemnation and later—after the uprising of the Lollards (poor farmers and the unemployed of the towns)—an English condemnation as well, but he himself was not molested. In 1417, however, when the connection between Wycliffe's doctrine and that of Hus had been recognized, his bones were dug up and burned.

In Bohemia. John Hus (c. 1370–1415), a professor of the University of Prague, protested in 1403 against the condemnation of Wycliffe by the German majority at the university; he was the Czech leader in the university struggles between Bohemians and Germans, which temporarily ended in 1409 with the exodus of the Germans from Prague. As a relentless preacher of reform (against the wealth and unspiritual conduct of the mostly German prelates), Hus drew upon himself the wrath of his archbishop, who brought his case before the Pisan pope Alexander V. Against Alexander's order for him to recant, Hus appealed to the new pope, John XXIII; the excommunication that the latter pronounced against him Hus answered by passionate sermons against papal politics and simony. The German king Sigismund (1368–1437), who wanted to be king of Bohemia as well, attempted to suppress the growing unrest; executions were already taking place. At the king's summons Hus came to Constance to "bear witness for Christ and his law." After three days of public hearings he was condemned as a heretic and, in violation of the royal safe-conduct given him, was burned at the stake.

Assessment of the Council of Constance

The proceedings of the council could be defended from a formal juristic standpoint, but the execution of Hus has been regarded as a frightful, unchristian, and historically dangerous act. Hus adopted much from Wycliffe, but, contrary to Wycliffe, he retained the Catholic concept of the sacraments, which indeed acquired substantial importance in his movement. The fate of their leader embittered the Bohemians; a whole people then rose against the rest of Western Christendom, whose sense of unity, already seriously weakened by the Great Schism, was now shaken again most severely. (Other forerunners, some of them from much earlier times were: Joachim of Fiore and his followers; some mystics; and, in an ambivalent way, nominalism.)

Problems and accomplishments of the medieval church. Alongside the far-reaching phenomena of decav, there were-in the 14th and 15th centuries a large number of significant beginnings of deeper religious life within the church. In almost all the monastic orders, reform efforts were seen: the Franciscan Observants, the observers of the old Rule; the alliance of Benedictine monasteries with reform congregations in Germany; the reform-minded Cardinal Ximénez in Spain. Important preachers of repentance also came to the fore, and the evangelistic movement took hold in France and Italy. A fruitful intensification of evangelical, biblical, and sacramental devotion found expression in many new foundations, such as that of the congregation of the Brethren (later also Sisters) of the Common Life founded by the Dutch preacher Gerhard Groote. Independent of clerical leadership, a new piety developed, called the *devotio rnoderna* (modern devotion), which was nourished by a personal imitation of Christ and which anchored church life in the continuity of faith and the sacraments. From this movement came Thomas à Kempis (*c.* 1380-1471) who wrote the *Imitatio Christi* ("Imitation of Christ"), the last universally accepted book of Christendom before it was split. This book of biblical and eucharistic prayer was, within a few years after its appearance, translated into many European languages; next to the Bible, it was the most widely read book in European literature. Strongly influenced by the *devotio rnoderna* were such diverse men, important for modern times, as the philosopher and cardinal Nicholas of Cusa, the Nominalist philosopher Gabriel Biel, Pope Adrian VI (reigned 1522–23), the Humanist Erasmus, and the reformer Martin Luther. In addition to the monastic reform congregations and the *devotio rnoderna,* there arose numerous brotherhoods devoted to charitable works and the religious education of the people. In view of these facts the total picture was not entirely negative, for the crucified Lord remained the centre of faith. At the same time, however, the interrelation of spiritual and material aspects had very unfortunate consequences.

The coming of a new spirit in Western Christendom

A great longing for ecclesiastical and worldly renewal existed in a large part of the population of the West. This led, initially in Italy, to retrospection toward the past. The ground appeared ready for a new culture. Since, however, disintegration had made such headway in the church, the new spirit did not fulfill itself primarily within the church but instead went more and more against it. The new elements, together with the medieval heritage, were undergoing a reordering that was at once decay and reformation; from it modern times were born. The most important transition of society, learning, and the church from the Middle Ages into modern times heralded itself intellectually (though not ecclesiastically or politically) in what became known as Humanism, the Renaissance attitude emphasizing man and his affairs.

II. Roman Catholicism in Europe in modern times

CATHOLIC REFORMATION AND COUNTER–REFORMATION

The condition of religious life on the eve of the Reformation. *Papal power and lack of spiritual care.* The great schismatic-papacy movement within the late medieval church, which found its most significant expression in the idea of conciliarism, had collapsed. The influence of the sovereigns upon the churches of their respective countries had grown extraordinarily, but the papacy had also increased in power. To be sure, its gains in power were primarily political, economic, and cultural, and not religious. The church's surrender to the culture of the Renaissance deepened still further the gulf between the religious view of Peter's office and its actualization. The Curia's excessive claim to power called forth opposition of the canonists in various countries and of their sovereigns. They demanded the limitation of the papal *plenitudo potestatis* ("fullness of power"); this division between national concerns and papal concerns likewise weakened the church.

The overall result of late medieval developments was, thus, the obscuring of the concept upon which the papacy had risen to prominence and influence. The concept of an institution founded by Jesus Christ—unique, unassailable in religious terms, its catholicity firmly based upon the Bible and standing above all one-sidedness—had been dangerously weakened. This widespread theological confusion was one of the factors that made the Reformation possible. Fateful also was the financial conduct of the Curia, which exploited Christendom, and the privileges of nobility attached to the lucrative ecclesiastical hierarchical positions, which were seldom regarded as spiritual offices for the cure of souls, but for the most part were used as sources of income for a life of pleasure.

Since a genuine development of spirituality was lacking, the growing masses in the cities found not even an approximately adequate spiritual care—in spite of enormously increasing numbers in religious professions (up to 10 percent of the city populations consisted of parish clergy, monks, and nuns); in the countryside, the economic misery of the parish priests and priests consecrated only for reading the mass made any constructive work impossible. In spite of some important exceptions, the lower clergy in general were without a "calling," without formal knowledge or dignity, despised by the people, and ridiculed by the Humanists. The abuses outweighed the still-existing islands of deeper piety. The proliferation of religious activities, including those of the pious foundations, brotherhoods, and the like, reflected an excessive preoccupation with the idea of gaining merit by good works and an unhealthy increase in the number of ecclesiastical pardons granted.

Deterioration of religious life

Inner. decay. Hand in hand with the multiplication of religious activities went an inner decay; the dogmatic doctrines of baptism, the mass, the church, and salvation became shallow. The catchword associated with piety was "being saved," in an outward sense, rather than primarily that of being justified through the crucified Jesus in repentance. Nevertheless, the doctrinal teaching of the redeeming life of Christ spoke to the people from innumerable pictorial representations and was thus a reality to them; the sublime liturgy of the Holy Mass continued to bear witness to the people the doctrine that all things happen "through Christ our Lord." Thus, in spite of all the disintegration and confusion, the negative side of medieval church life was not the whole picture, and the church's practices should not be characterized simply as the authentic product of the church's doctrines, as many of the propagators of the Reformation so declared.

A striking proof of the dislocation of the religious life of this period is supplied by the increased prevalence of heretical and sectarianizing tendencies (as in the teachings of Wycliffe and Hus and also in penitential movements) that drew their religious strength from the Bible. The New Testament praises poverty (see the Sermon on the Mount, Matt. 5–7; Parable of the Rich Man, Matt. 19:23–24), but the ecclesiastical hierarchy had devoted itself to riches. The disinherited confronted the hierarchy with the word poverty; there developed a Christian socialism (peasant uprisings in southern Germany and by the Lollards in England). While this movement took hold of the lower classes and brought them into confrontation with the hierarchy, the Renaissance and Humanism led the educated classes—who until then had been exclusively in the service of the church—to new secularized (but by no means exclusively non-Christian) conceptions of the world and human culture.

In the pious—but not entirely well-balanced—Dominican monk Girolamo Savonarola (burned at the stake in 1498, after having tried to reform the Italian city of Florence through his moral preaching and the passing of strict laws for conduct), the conflict of all these forces was expressed as if concentrated into one symbol. Savonarola was a striking embodiment of the rousing but, in its effect, vague repentance preaching of the time and also a revelation of its sharp contradictions. His personality and his fall from power and public support penetratingly illuminate the confusion of the period, particularly within the church, which had elected the notorious Alexander VI (Rodrigo Borgia) in 1492.

The Catholic Reformation. Doctrinal issues. Since the Reformation, according to many of the Protestant Reformers, was concerned with doctrinal issues primarily and with moral and practical issues secondarily, an understanding of the Reformation should necessarily begin with doctrinal matters. Martin Luther (c. 1483–1546), the founder of the Protestant Reformation, understood correctly that the gospel hinged upon but few major points, indeed, upon only a single one: the justifying faith that is the gift of God. His doctrinal approach could thus be strikingly simple. Together with the uncompromisingness and the impressive oratorical power with which he set forth his views, this simplicity made his preaching an uncommon attraction; it seemed to produce effective phrases of its own accord: "the Word," the "pure doctrine," the "freedom of the Christian man," "the gospel" (as opposed to the Law).

The Catholic "and" and the Protestant "alone"

Luther seized upon a central point of the Christian message—that of trusting faith in God the Father through the crucified Christ—and pronounced it to be the whole. Accordingly, the relationship of Reformation to Catholic doctrine may be correctly expressed in the statement: the Catholic "and," which illustrates its inclusive quality, stands opposed to the Protestant "alone," which is concerned with the centrality of a particular doctrine. (The Protestant Reformation battle cries were "Scriptures Alone," "Faith Alone," and "Grace Alone.") But this Catholic formulation is more "evangelical" than has long been thought by some church historians, theologians, and sectarian polemicists. The exclusive formulations of Protestantism place especially strong emphasis on what is the central point, yet, correctly understood, this central point is also Catholic. The Catholic "and" does not denote an addition but rather a dynamic-functional relationship and development. Scripture and tradition do not stand opposed as things foreign to each other; on the contrary, tradition in the church is viewed as the whole living heritage, transmitted primarily through inspired scripture. The functional meaning of the Catholic "and" leads to a concept of the church that affirms the sacramental priesthood (the office of those who by the power of their ordination as priests can officiate at the mass and perform other sacramental acts), the teaching office of the church, and the general priesthood of all believers, who have access to God through prayers and other such acts, thus pointing to the final decisive difference between Reformation-oriented and Catholic. It is a different concept of the church that is involved. Catholic doctrine tries to realize fully the biblical command: to be "hearer" of the word preached by the Apostles and their successors the bishops; *i.e.*, to take into account all the texts and books of the Bible according to the teaching of tradition and magisterium. The undermining or rejection of the magisterium in favour of "Scriptures alone" brought along increasing insecurity and separations within Protestantism. Contrary to Luther's intention—*i.e.*, the restoration of the pure gospel in one church—he became in a certain sense the father of modern liberalism.

Beginnings of Catholic reform. The history of the Reformation can only have a depressing effect upon Roman Catholics. One Catholic defeat, ecclesiastical and political, followed another, and Catholic failures were to blame for many of them. Catholic strength appeared to be exhausted. But this was only apparently so. For there was also—besides the Protestant Reformation—a Catholic Reformation of the 16th century that grew out of independent roots within the church and was a positive Catholic achievement, not merely a reaction to the Protestant assault. Important beginnings of the Catholic Reformation dated, in fact, from before the Reformation; others occurred later, independently of it (particularly in Spain, as in the work of lgnatius of Loyola, the founder of the Jesuits, and the mystical writer Teresa of Avila). The results, too, were not merely negatively defensive, but were in large part extensions based on Catholicism's own central foundations; *e.g.*, a new birth of Catholic piety.

At the same time, of course, the impact of the Reformation itself aroused many Catholic forces and induced and accelerated Catholic reform. Destruction threatened the church as it had been viewed for centuries, but the Catholic will to live responded. This reciprocal effect may be seen most clearly in the Council of Trent (1545–63). This will to live expressed itself in holiness, which is the very root of the entire Catholic reform of the 16th century, as at other turning points in church history. The church found the strength to renew itself. From the beginning the reformers, to a great extent, did not criticize others but themselves. They did not begin to alter the institution of the church, but rather their representatives.

The role of the papacy in the Protestant Reformation

Reform of the priesthood—which had become secularized—became the motto of the Catholic renewal. In spite of all the reservations and detailed criticisms that can be made, the work of reform was unexpectedly successful, although certainly far from uniform in the different countries. South of the Alps, secularization had gone much further than in Germany, yet it turned out that in that northern country, where the ecclesiastical organization—centred in a foreign city, Rome—was much more severely shattered, the Catholic religious renewal advanced at a significantly slower pace. It was quite otherwise in the south—in spite of the fact that, into the middle of the 16th century, the papacy displayed alarmingly little understanding of the dangers with which the church was threatened by the Reformation. By its altogether politically motivated actions, the papacy repeatedly compelled the Holy Roman emperor to put aside the struggle against the Protestants, so that the papacy inadvertantly became the very saviour of the Protestant Reformation. With the exception of Pope Adrian VI (reigned 1522–23), the papacy had no part in the early stages of the Catholic reform.

Forces of religious *renewal.* Not until the devastation of Renaissance Rome during the sack of Rome in 1527, by the unpaid German and Spanish mercenaries of the emperor Charles V (1500–58), was the negative prerequisite created for the papacy's collaboration with reform-minded groups. Just as in the Middle Ages, the forces of religious renewal did not emanate directly from the papacy and even less so from the bishops. They rather came from the strength of the faith of the religious community or of certain leading groups. Again, as in the Middle Ages, the forces of reform first reached their full strength through union with the papacy.

For a historical understanding of the Catholic Reformation it is important to note how the Catholic renewal was able to succeed even against very strong obstacles within the church itself and with numerous setbacks. The centre of resistance was, above all, the families and international politics of the popes, as well as curial resistance to the council for which reformers were calling. The reforms expected from such a council were extraordinarily feared by the curial officeholders. The effect of this state of mind was that, in Rome, the danger of incipient Protestant reforms to the Catholic Church was at first simply not seen. Thus, resistance was offered to the reforming zeal of Pope Adrian VI, and, as late as 1534, a man like Paul III (reigned 1534–49), entirely a product of the acutely secularized Renaissance, was elected pope. Even Paul IV (reigned 1555–59), zealous for reform, lapsed into a disgraceful nepotism. The same impression of ineradicable corruption is given by the style of living of the clergy who were members of the high nobility in France, Germany, and Poland right up to the end of the 16th century. With such a worldly, immoral attitude, such members of the hierarchy acted as if the Reformation had never occurred, as if the church were not fighting for its life.

Intrachurch reform. Hence, the intrachurch reconstruction of the 16th and 17th centuries was a complex process. The intrachurch reform was, in the first instance, the fulfillment of the manifold reform stirrings of the late Middle Ages. The basically nonclerical groups of the devotio *moderna*, which practiced a lay piety that was humanistic in form, were the forerunners of those religious associations from whose spirit new churchliness and Catholic piety arose in Italy. An important role in the reform was played by the Oratory of Divine Love, an ordinary religious order founded by laymen to aid in the moral and spiritual improvement of its members. The number of priests it admitted was limited. As early as 1497, a brotherhood founded at Genoa bore this name, which was adopted also by similar brotherhoods in other Italian cities. Supporters of the Catholic reform in Italy, such as Gaetano da Thiene (1480–1547) and Gian Pietro Carafa, later Paul IV (1476–1559), emerged from this oratory. The pious practices it prescribed for its members (*e.g.*, attendance at mass, communion, fasts, prayer, care of the sick) were the model for many associations subsequently founded. Similar reform groups were formed in Venice, led by Paolo (formerly called Tommaso) Giustiniani (1476–1528), Vincenzo Quirini (1479–1514), and Gasparo Contarini (1483–1542); the latter was a liaison man for the Italian reform forces. The specific goal of the Catholic reconstruction was the restoration of that which was most lacking in the church—the cure of souls. This was also the goal of the Theatine Order, founded by Gaetano da Thiene and Gian Pietro Carafa. Among the many abuses in the church, the two most ruinous had been clearly recognized as interconnected: the irreligious clergy and excessive wealth. Hence this new order (the Theatines) was expected to be, in complete poverty, a living example of irreproachable churchly spirituality, and an example that would encourage reform of the clergy from the ground up.

The Theatine Order

All these foundations were not bellicose in outlook; they wished to render positive services to church renewal. In many cases their connection with Humanism was determined already by the identity of their leading members. The piety that these groups practiced was not infrequently misunderstood and deprecated as protestantizing. Indeed, an often too-amorphous "evangelism" made itself felt—but in spite of some crises (such as the conversion to Protestantism of the Franciscan Bernardino Ochino), loyalty to the church remained unbioken. The positive role that Humanism played through its revivification of the Holy Scriptures was very evident. The spirit of this reform bore fruit in the regulations established at Trent regarding the education and life of the clergy.

Long before, Adrian VI had wanted to introduce reform. The last non-Italian Pope, he combined German intensity and Spanish churchliness. He was a great but tragic figure. He had the courage to publicly acknowledge (at the Diet of Niirnberg of 1523) the responsibility of the clergy—and in particular of the Roman Curia—for the religious-ecclesiastic revolt; but the time—or more correctly, the church—was not yet ready for his reforms. He was given no appropriate response, and his early death was a setback for reform. His successor, Clement VII (reigned 1523–34) was an entirely political figure, who vehemently rejected the council that persons all over Europe were demanding.

The Council *of* Trent. Only under Paul III (reigned 1534–49) did a slow change begin. Personally still living in the spirit of the Renaissance, yet thinking and acting with the political outlook of a modern monarch, he encouraged the newly formed orders (Theatines, Capuchins, Ursulines, and later the Jesuits). He revitalized the Sacred College of Cardinals by appointing the most important religious figures of the time (Gasparo Contarini, Giovanni Morone, Reginald Pole, and others) and finally—after much dangerous delay—convened the Council of Trent.

The council came too late for the re-establishment of church unity, but it undoubtedly gave Catholics new strength and reinforced the will to reform. To be sure, it achieved a transformation in Catholic self-confidence but only with difficulty and over a long period of time. The 'council continued through three periods of sessions: 1545–47/48, 1551–52, 1562–63. The core of its membership--which particularly in the first period was numerically very small—always consisted of Spaniards (Diego Lainez, the successor of Ignatius as general of the Jesuits, Domingo de Soto, Alfonso Salmerón, and Melchor Cano); the theological achievements of the general of the Augustinians, Seripando (1492–1563), were probably of the greatest value. Voting rights were no longer extended to the theologians, as at the Councils of Constance and Basel, but were held only by prelates having their own jurisdictions. The influence of political considerations was evident in the transfer of the council from German-controlled Trent to the Italian city of Bologna (1547), precisely at the moment when the Emperor, triumphant in Germany, had secured the participation of the Protestants in the council. With this transfer the Emperor's attempt to mediate a reconciliation of the Protestants with the church decisively failed, since the Protestants refused to go to a city that was controlled by the papacy. At the second session, again in Trent, representatives of the German Protestants appeared, but without success. This time the German princes conspiring against Charles V compelled the breaking off of the consultations. The third period of sittings brought the council to its conclusion and a profusion of reform decrees,

From the beginning, questions of reform and of faith were treated concurrently. Of especial importance was the decree of justification, in which this question, which was at the centre of Reformation doctrine, was finally clarified in Roman Catholic theology. The council declared that, in accord with scriptural doctrine, God and his mercy alone are decisive, but God works mysteriously to enable man to accept faith so that nothing in justification is other than God's work, and the very merit is a gift of God. The questions of reform were dealt with primarily in the third period of the council; a general cleanup of abuses was at least in principle agreed upon. Regulations regarding multiple benefices (having more than one revenue-producing ecclesiastical office), commerce in benefices, and, above all, the cure of souls, were reformed; through the issuance of the decree on the establishment of seminaries, the creation of a new clergy was begun.

Significance of the Council of Trent

The execution of the Trent decrees—that is, the actual achievement of the Catholic reform—required decades and, in some countries, centuries. The deepest meaning of the Council of Trent lay in that it contributed decisively to the clarification of the Catholic concept of the church; it was a victorious conclusion—in favour of the papacy—of the conflict that had begun in the 13th century.

The Counter-Reformation. Unfortunately, the expression Counter-Reformation is often used inexactly. Here it will be used literally to mean the totality of Catholic efforts directed against the Protestant movements. They were, of course, connected with the concept of power that since the Middle Ages had been central to the thought of the leading elements of the church and particularly of the popes. But these efforts can by no means be described as simply the expression of egotistic and nonreligious tendencies. The totality of the Catholic action directed against the Protestant Reformation was governed not only by the instinct of self-preservation but also by the church's comprehensive missionary mandate.

Intellectual achievements. Accordingly, it brought to light an amazing religious strength among Catholics. In the defense against the Protestant theological attack, there came forward in the first instance scholars who remained loyal to the Catholic Church. In first place among these was the highly learned Johann Eck (1486–1543), who was the first to recognize in Luther's theses the sign of a revolutionary attack. Others prominent in the first years after Luther's public fame were: the polemicizing chaplain of Duke George of Saxony, Hieronymus Emser; the satirical preacher and folk writer Thomas Murner; the Franciscan Kaspar Schatzgeyer; and the schoolmaster Johannes Cochlaeus. In Italy there was Cardinal Cajetan (who had been involved in a discussion with Luther in Augsburg in 1518) and the previously mentioned Contarini and Seripando, who were notable for their deep appreciation of the problems that the Reformation leaders had newly raised on the basis of the Bible. In England there was Bishop John Fisher of Rochester (1469–1535), and, initially, King Henry VIII (1491–1547).

Many noteworthy theological works, springing from devoted loyalty toward the church, could be cited; but none even approached the power of creative genius of Luther or Calvin. There was too little positive contribution issuing from the very centre of Catholic life and instead too much polemic. A large part of the struggle against Protestantism took place in the political sector. The outstanding figure, Luther's only opponent of equal genius, was Charles V, in whom deep Christian faith was combined with the concept of a Western Empire. In taking the side of the Catholic faith, Charles realistically took setbacks and detours in his stride, including the insufficient support accorded him by the popes of his time. When, as a result of the Peace of Augsburg of 1555, the continued existence of the Lutheran confession—and therewith the Schism—was recognized in imperial law, Charles abdicated. His son Philip II (reigned 1556–98) successfully fought for Catholicism in Spain (using the notorious Inquisition). In the German Empire Charles's brother Ferdinand I made concessions to Evangelical (Protestant) princes on the "heathen" principle ***cujus regio ejus religio*** (*i.e.*, the religion of the prince is the religion of the land). At this, the Counter-Reformation definitely began to take hold. Because its rulers remained Catholic, Bavaria remained with the Catholic Church. Their intervention also kept the Rhine area as ecclesiastical territories for Catholicism. Later the Austrian lands were recatholicized.

Importance of the Jesuits

In alliance with the princes, the Jesuit Order in particular became religiously effective. Founded in 1534 by Ignatius of Loyola (1491–1556), this order united many inspired and inspiring pioneers of the Catholic renewal on the basis of stern discipline of the will and unconditional self-sacrifice. The Jesuits worked as pioneer detachments in contested centres; through their schools and colleges, through preaching and the cure of souls, they ecclesiastically and religiously consolidated the ground won.

In England and France. In England there was only a short period during which the Counter-Reformation and Catholic Reform could operate: in the reign of Henry VIII's Catholic daughter Mary, from 1553 to 1558. She at once sought, with the help of her cousin, the influential Cardinal Pole, to restore and build up what remained of Catholicism. Here, too, from February 1555, violent methods were adopted, which claimed 273 victims. A symbol of the tragic character of Mary's effort is the double appellation with which she is recorded in history; one side called her Mary the Catholic, the other, Bloody Mary.

After her early death, her sister and successor, Elizabeth I (reigned 1558–1603), slowly and cautiously but effectively again eliminated all Catholic activity; only as underground groups and in emigration did Catholic remnants persist. A papal bull of anathema against Elizabeth in 1570, which seemed to sanction even the murder of the Queen, contributed greatly to making the lot of the Catholics in England yet worse. Furthermore, after the execution of their pretender to the throne, Mary Stuart, in 1587, Catholics no longer had political representation.

In France the question of creed submerged in that of politics. The royal house at first took severe measures against Calvinism, but at the same time Henry II (reigned 1547–59)—out of political considerations—supported the German Protestant princes in their struggle against the Catholic emperor. The Pope was threatened by Gallicanist (French antipapal) manoeuvres, with a national (French) council, and even with schism. In the cruel Huguenot Wars (French religious wars of the 16th century), the question of creed eventually became unimportant.

Papal activities of the Counter-Reformation. The collaboration of the popes with reform elements in the Counter-Reformation began with the establishment of the Roman Inquisition, instituted in 1542 to destroy heresy. Its creator was Cardinal Carafa, who as Paul IV (reigned 1555–59) put it to work, and most frightfully. He was a genuine reformer who eliminated many abuses but finally, through his cruel excess of zeal, did more harm than good. In contrast to him, the personal focus of Pius IV (reigned 1559–65) was worldly. By wise management, however, he saw the Council of Trent to its conclusion. Under him there also began, properly speaking, the ecclesiastical Counter-Reformation, in that he installed important men in the right places; *e.g.*, the Jesuit Peter Canisius in Germany and Cardinal Hosius in Poland. Even his practice of nepotism worked out well for the church. His nephew, whom he made a cardinal at the age of 14, was Charles Borromeo (1538–84), a saint who was to be of immense importance for the spiritual renewal of the church. Above all, to Charles Borromeo must go the credit for the election of Pius V (reigned 1566–72), the first pope of modern times to have been considered saintly. With him occurred the definitive reversal of the papal policies that had originated in the Middle Ages and had been secularized in the Renaissance, that is, there was a turning away from political goals. Pius V enforced the Tridentine decrees (those of Trent) in Rome and concerned himself with ending the serious abuses in the Curia. His reign was the first high point of Catholic reform and the Counter-Reformation.

Missions in Asia and Latin America

Pius V's successors, Gregory XIII (reigned 1572–85) and Sixtus V (reigned 1585–90), were no saints, but they continued his work. In Rome, seminaries were founded for the training of the European clergy (Germanicum, English College, Collegium Romanum); these were all headed by Jesuits. The extra-European missions (India, Latin America, China), again led principally by Jesuits, achieved extraordinary results at first, as in the work of Francis Xavier. But now there arose the problem still familiar to modem missions—the propagation of Christianity, which had Western characteristics, among peoples with entirely different traditions. In Asia, an attempt was made to solve the problem by extensive adaptation to existing customs. This magnificent experiment, entirely in the tradition of the Christianizing work of the early church, was later forbidden by Rome; thereafter the decline of Asiatic Christianity was assured. In Latin America, where Christianization was coupled with brutal suppression of the native population, including forced baptism, the bishops (such as Bartolomé de Las Casas) were the only defenders of the Indians, and Christianity was only very slowly able to take root. Through the founding of the Congregation for the Propagation of the Faith (Congregatio de Propaganda Fide; 1622), the papacy sought to unify and organize its missionary work.

Europe, however, remained the centre of the church's activity. The church had reconsolidated itself, and a whole band of saints (including Philip Neri, Francis de Sales, Frances de Chantal, Vincent de Paul, and many others) had revived its strength. The popes of the first half of the 17th century were mostly of lesser importance, and a certain reduction in the energy of the Curia was evident. An ill omen for the future relationship of the church to the fast-developing modern world was the rejection of the new natural science by theology in the trial of the astronomer Galileo (1633).

The Thirty Years' War

European religious and political antagonisms. In Bohemia, the heartland of Europe, the difficulties caused by the religious schism, which had increasingly turned into political antagonisms, led to a partly religious, partly dynastic conflict involving the destruction of Protestant churches in Catholic territories, the deposition in 1619 of the Catholic Habsburg ruler, Ferdinand II, and the calling of Frederick V, the "Winter King," to the throne of Bohemia. In the course of the following 30 years, nearly all of Europe was drawn into a terrible war, in which all fronts finally became entangled. Only England remained uninvolved, because of its own difficulties. In the struggle against the German Emperor—who claimed to be fighting for the cause of the church and who proposed an Edict of Restitution in 1629 to compel the return of the stolen church properties—the lead was eventually taken by France, whose all-powerful prime minister was Cardinal de Richelieu. Acting out of purely political motives and unhindered by creed, Richelieu personified the force that was to be largely decisive in the development of the church during the next century and a half, the age that would be characterized by the state church. This period, known as the Baroque era, produced, in the Catholic countries, splendid achievements in religious education, art, intellectual and literary production, and spirituality. Yet it must be acknowledged that Catholic thinkers and scholars lagged behind the forefront of European leadership, and that even many theological disciplines stagnated.

The Thirty Years' War in Germany, precipitated by religious reasons, was at last ended by the Peace of Münster and Osnaburg (1648)—known as the Peace of Westphalia—which finally acknowledged the equal rights of Catholics, Lutherans, and Calvinists in the Holy Roman Empire. The automatic assumption, however, was still that there was a universal validity of the Christian confession. There was no question of general tolerance in the modern sense of the 20th century. (J.L.)

ROMAN CATHOLICISM IN THE 17TH AND 18TH CENTURIES (1648–1789)

Emphasis on "Roman" Catholicism

Changing conditions. The signing of the peace in 1648 may have meant that the era of the Reformation had ended, but, for those who remained loyal to the see of Rome it meant that what had been thought of as a temporary flurry would now be a permanent condition—in a word, that henceforth Catholicism would have to be seen as "Roman" Catholicism. The church still claimed to be the only true church of Jesus Christ on earth, but, in the affairs of men and of nations, it had to live with the fact of its being one church among several. The Roman Catholic Church was also obliged to deal with the nations and national states of the modern era one by one. To understand the history of modern Roman Catholicism, therefore, it is necessary to identify trends that went beyond geographical boundaries and to consider particular states or regions—such as France, Germany, the New World, or the mission field--only as illustrations of tendencies that permeated the entire life of the church. Most of the development of Roman Catholicism since 1648 makes sense only in the light of this changed situation.

The papacy of the 17th and 18th centuries manifested the results of the change. On June 6, 1622, Gregory XV (1621–23) created the Congregation for the Propagation of the Faith (Congregatio de Propaganda Fide, hence "propaganda"). Its responsibility was, and still is, the organization and direction of the missions of the church to the non-Christian world, as well as the administration of the affairs of the church in areas that do not have an ordinary ecclesiastical government (for example, the United States as late as 1908). It has therefore played an important role in the efforts to restore Roman Catholicism in Protestant and, to some degree, in Eastern Orthodox territories.

The incumbent of the papal throne at the time of the Peace of Westphalia, Innocent X (1644–55), formally protested the provisions of the Peace of Westphalia as a violation of church law, but without avail. Under several of the popes of the 17th century, notably Alexander VII (1655–67), the artistic embellishment of the city of Rome, begun in the Renaissance but interrupted by the Reformation, was resumed and advanced. Under Alexander VII, Queen Christina of Sweden was converted from Lutheranism to Roman Catholicism. The internal reform of the life and law of the church was a continuing issue. Clement XI (1700–21), who is perhaps best remembered for his role in the Jansenist controversy (see below *Jansenism*), also engaged in conflicts with the rulers of Prussia, the Holy Roman Empire, Savoy, and Spain. That the process of electing Benedict XIV (1740–58) began February 17, 1740, and took six months, with (according to contemporary accounts) 255 votes before the deadlock was broken, was symptomatic both of the party spirit within the church and of its precarious relation to the cultural and political forces around it.

Developments in France. ***The Gallican problem.*** In many ways, however, it was this relation to political powers that determined the course of church history more than did the leadership of the popes. Not only the shrinking authority of the church as a consequence of the Reformation but also the expanding ambition of the state as a consequence of the growth of nationalism put ecclesiastical and secular government on a collision course throughout Europe. France, "the first daughter of the church," was the national state whose development during the 17th and 18th centuries most strikingly dramatized the collision, so much so that Gallicanism, as the nationalistic ecclesiastical movement was called in France, is still the label put on the efforts of any national church to achieve autonomy.

Nationalistic ecclesiastical movements

Usually, the autonomy from Rome implied subjection to the French crown, particularly during the reign of Louis XIV, who sought to extend still further the so-called prerogatives of France when Rome resisted. A conclave of bishops and deputies met on March 19, 1682, in Paris and adopted the Four Gallican Articles, which had been drafted by Jacques-Bénigne Bossuet, a French bishop and historian. These asserted that: (1) In temporal matters rulers are independent of the authority of the church. (2) In spiritual matters the authority of the pope is subject to the authority of a general council, as had been declared at the Council of Constance. (3) The historic rights and usages of the French church cannot be countermanded even by Rome. (4) In matters of faith the judgment of the pope is not irreformable but must be ratified by a general council. The next move was up to the papacy: Innocent XI and Alexander VIII rejected Louis' candidates for bishoprics in France, and only in 1693, when Innocent XII was pope, was this all but schismatic conflict resolved. Though Gallicanism was in part an expression of the distinctive traditions of French Catholicism and in part a result of the personal power of Louis XIV, the Sun King, it was, perhaps even more fundamentally, a systematic statement of the inevitable opposition between the papacy and a series of rulers from Henry VIII (1491–1547) of England to Joseph II (1741–90) of Austria, who, though remaining basically Catholic in their piety and belief, wanted no papal interference in their royal business but insisted on the right of royal interference in the business of the church.

Jansenism. The church in France was the scene of controversies other than these administrative and political ones. In 1640 there was published, posthumously, a book by a Dutch theologian Cornelius Jansen, entitled ***Augustinus,*** a defense of the theology of Augustine against the dominant theological trends of the time within Roman Catholicism. Its special target was the teachings and practices associated with the Jesuits. Jansen and his followers claimed that the doctrine of grace defined by the theologians of the Counter-Reformation in their opposition to Luther and Calvin had erred, in their view, in the other direction; *i.e.*, emphasizing human responsibility at the expense of the divine initiative and thus relapsing into the Pelagian heresy, against which Augustine had fought in the early 5th century. Over against this emphasis, Jansenism asserted the Augustinian doctrine of original sin, including the teaching that man cannot keep the commandments of God without a special gift of grace and that the converting grace of God is irresistible. Consistent with this anthropology was the rigoristic view on moral issues taken by Jansenism in its condemnation of the tendency, which it claimed to discern in Jesuit ethics, to find loopholes for evading the uncompromising demands of the divine law. When it was espoused in the ***Lettres Provinciales*** ("Provincial Letters") of Blaise Pascal, a French philosopher, this campaign against Jesuit theology became a cause célèbre. The papacy struck out against Jansenism in 1653, when Innocent X issued his bull, *Cum Occasione* ("With Occasion"), and again in 1713, when Clement XI promulgated his constitution, *Unigenitus* ("Only-Begotten").

Theologically, Jansenism represented the lingering conviction, even of those who refused to follow the Reformers, that the official teaching of the Roman Catholic Church was Augustinian in form but not in content; morally, it bespoke the ineluctable suspicion of many devout Roman Catholics that the serious call of the Gospel to a devout and holy life was being compromised in the moral theology and penitential practice of the church. Though Jansenism was thus condemned, it did not remain without effect, and, in the 19th and 20th centuries, it contributed to an evangelical reawakening, not only in France but throughout the church.

Quietism. Quietism, another movement within French Roman Catholicism, was far less strident in its polemics and far less ostentatious in its erudition but no less threatening in its ecclesiastical and theological implications. Quietism was, in many ways, yet another form of the Augustinian opposition to any recrudescence of the "Pelagian" idea that man's religious activity can make God propitious to him. In Quietism this belief was associated with the development of a technique of prayer in which passive contemplation became the highest form of religious activity. Christian mysticism had always combined, in an uneasy alliance, the techniques of an aggressive prayer that stormed the gates of heaven and a resigned receptivity that awaited the way and will of God, whatever it might be. In the theology of François de Fénelon, a French archbishop and mystical writer, Quietism was combined with a scrupulous orthodoxy of doctrine to articulate the distinction between authentic Catholic mysticism and false spiritualism. Nevertheless, as scholars of medieval mystical movements have suggested, Quietism showed the great gulf fixed between the Roman Catholicism that came out of the Counter-Reformation and the spirituality of the preceding centuries, both Greek and Latin. A devotion such as that of St. Gregory of Nyssa and Evagrius of Pontus, Greek theologians of the 4th century, was completely ruled out by the legalistic theology that condemned Quietism.

Adaptation to non-Western religious traditions

Controversies involving the Jesuits. ***The Chinese rites controversy.*** An analogous judgment would have to be voiced concerning the Chinese rites controversy, centring on Matteo Ricci, an Italian Jesuit missionary in China. Decades of scholarly research into Buddhist and Confucian thought had prepared Ricci for a campaign that sought to attach the Roman Catholic understanding of the Christian faith to the deepest spiritual apprehensions of the Chinese religious tradition; the veneration of Confucius, the great Chinese religious and philosophical leader of the 6th century BC, and the religious honours paid to ancestors were to be seen not as elements of paganism to be rejected out of hand, nor yet as pagan anticipations of Christianity, but as rituals of Chinese society that could be adapted to Christian purposes. Ricci's apostolic labours won him many converts in China, but they also won the suspicion of many in the West that the distinctiveness of Christianity was being compromised in syncretistic fashion. The suspicion did not assert itself officially until long after Ricci's death; but, when it did, the outcome was a condemnation of the Chinese rites by Pope Clement XI in 1704 and again in 1715 and by Pope Benedict XIV in 1742. Ancestor worship and Confucian devotion were said to be an inseparable element of traditional Chinese religion and hence incompatible with Christian worship and doctrine. Here again, the embattled situation of the Roman Catholic Church in the 17th and 18th centuries helps to account for an action that seems, in historical perspective, to have been excessively defensive and rigoristic.

Suppression of the Jesuits. Among the repercussions of the controversy over Chinese rites was an intensification of the resentment directed against the Society of Jesus, to which some of the other movements mentioned above also contributed. The widespread support enjoyed by Jansenism was due in part to its attack on the moral theology associated with the Jesuits. Pascal's ***Lettres Provinciales,*** although placed on the Index in 1657, voiced an opposition to Jesuit thought and practice that continued to be read throughout the century that followed. The political role played by members of the Society most probably evoked the campaign to suppress it. The Portuguese crown expelled the Jesuits in 1759, France made them illegal in 1764, and in 1767 Spain and the Kingdom of the Two Sicilies also took repressive action against them. But the opponents of the Society achieved their greatest success when they took their case to Rome. Pope Clement XIII is said to have replied that the Jesuits "should be as they are or not be at all" and refused to act against them. But his successor, Clement XIV (1769–74), whose election was urged by the anti-Jesuit forces, finally did take action. On July 21, 1773, he issued a brief, ***Dominus*** ac ***Redemptor*** ("Lord and Redeemer"), suppressing the Society for the good of the church. Frederick II of Prussia and Empress Catherine II of Russia—one of them Protestant and the other Eastern Orthodox—were the only monarchs who refused to promulgate the order to suppress the Jesuits when it was issued. In these lands and in others, the Society of Jesus maintained a shadow existence until, on August 7, 1814, Pope Pius VII restored it to full legal validity. Meanwhile, however, the suppression of the Jesuits had done serious damage to the missions and the educational program of the church, and this at a time when both enterprises were under great pressure.

Religious life. Yet, it would be a mistake to allow the narrative of these controversies to monopolize one's attention. Less dramatic but no less important was the continuing life of the Roman Catholic Church during these centuries as "mother and teacher." Bossuet was not only the formulator of Gallican ideology; he was also one of the finest preachers of Christian history, addressing king and commoner alike and asserting the will of God with eloquence, if sometimes with undue precision. Together with Jean Mabillon, a Benedictine monk and scholar, Bossuet helped to lay the foundations of modern Roman Catholic historiography. During the 18th century their work was continued and expanded, especially by Mabillon's confreres, the Maurists, a Benedictine group that edited the works of the Greek and Latin fathers. Both Jansenism and Quietism must be seen not only as parties in a controversy but also as symptoms of religious vitality. Engaging as they did considerable segments of the Roman Catholic laity, they expressed "the practice of the presence of God" with a new vigour. The Roman Catholic Church of this period exercised a profound influence on culture and the arts. Indeed, the spirit of Baroque is inseparable from the Counter-Reformation, as is visible, for example, in the Gesù Church in Rome and in the sculpture and architecture of Gian Lorenzo Bernini. Among the literary figures of the time, Pascal and Cervantes are notable examples of Roman Catholic thought and piety expressed through writing. The most fateful of the church's conflicts with modern culture in this period took place in the natural sciences. The condemnation of Galileo in 1616 and again in 1633 as "vehemently suspected of heresy" was more important symbolically than intrinsically, as a sign of the alienation between science and theology. From this period came the establishment or further development of several major religious orders, including the Daughters of Charity, founded by Vincent de Paul in 1633, and the Trappists, who take their name from the Cistercian abbey of La Trappe, which in 1664 was transformed into a community of the Strict Observance.

Scholarly and literary activities

ROMAN CATHOLICISM IN THE REVOLUTIONARY ERA

The church in France (1789–1801). The period of the Reformation and the Counter-Reformation was a time of convulsion for the Roman Catholic Church, but the era of revolution that followed it was, if anything, even more traumatic. This was partly because, despite the polemical rancour of Reformation theology, both sides in the controversies of the 16th and 17th centuries still shared much of the Catholic tradition. Politically, too, the assumption on all sides was that rulers, even when they opposed one another or the church, stood in the Catholic tradition. In the 18th century, however, there arose a political system and a philosophical outlook that no longer took Christianity for granted, that in fact explicitly opposed it, compelling the church to redefine its position more radically than it had done since the conversion of the Roman emperor Constantine in the 4th century.

What made the relation of the Roman Catholic Church to the ***ancien régime,*** the political and social system before the French Revolution in 1789, so problematical at the time of the revolution was a subtle but fundamental difference. Although the rhetoric of the revolution spoke as though the church and the old order had been one, no one could study the history of the church under (or over against) Louis XIV and accept so simplistic an interpre-

Underlying causes of anti-ecclesiastical sentiments

tation. Conflict there had been, bitter and uncompromising conflict—and yet conflict within the context of given presuppositions. It is significant, for example, that the French aristocracy, soon to become the hated object of revolutionary zeal, constituted the source of almost all the bishops of the church in the *ancien régime.* This also meant that positions of authority in the church were largely foreclosed to the lower clergy because of their class. The theological and ecclesiastical parties identified with opposition to Rome were frequently those that drew the support of the laity; Jansenism, for example, was identified as the position of the lay lawyers who spoke for the French courts of justice over against the hierarchy. In spite of the hostility between church and state, therefore, the old regime appeared to its critics to be a monolith. Thus, when the French philosopher Voltaire said, "Écrasez l'infâme" ("Crush the infamous one"), he may have meant superstition, ignorance, and tyranny, but what it added up to concretely in the minds of the revolutionaries was the supposed alliance of the monarchy with the Roman Catholic Church. This identification was only confirmed when the defenders of the established order, both lay and clerical, spoke out against the threat of revolution with a greater awareness of its dangers than of its justification.

Complicating the predicament of the church in the old regime was the corrosive influence of the Enlightenment on the religious beliefs of much of the lay intelligentsia. Enlightenment rationalism took hold among many defenders of the political status quo, as well as among clerical scholars, helping to produce the beginnings of critical biblical scholarship and of religious toleration. It would be an oversimplification, therefore, to put the Enlightenment unequivocally on the side of the critics and revolutionaries. Perhaps no one embodied the spirit of the Enlightenment more completely than Frederick II the Great of Prussia. But the confidence in reason and the hostility to "superstition" cultivated by the Enlightenment inevitably clashed with the Christian reliance on historical revelation and with the belief in supernatural grace as communicated by the sacraments. The political and social prerogatives of the church were threatened by the Enlightenment, especially when it was allied with the expanding claims of an autocratic "enlightened despotism." The brotherhood taught by such groups as the Freemasons, members of secret fraternal societies, and the Illuminati, a rationalistic secret society, provided a rival to the Catholic sense of community as found in the church. In the *Magic Flute,* the Austrian composer Wolfgang Amadeus Mozart (who wrote his *Requiem Mass* in the same year) celebrated the Masonic alternative to the mass of the church. When the church condemned Galileo, it confirmed the suspicion of many that traditional theology was inseparably wedded to an obsolete world view and that science had to be freed from its tyranny if it was to progress.

Although leaders of the state were often more hospitable to the ideas of the Enlightenment than were leaders of the church, the latter proved to have been more accurate in their estimate of the revolutionary implications of these ideas. The "heavenly city of the 18th century philosophers" may originally have been intended as a substitute for the City of God, but it also provided much of the ideological rationale for the attack upon the *ancien régime.* In the familiar epigram of the Swiss writer, Jacques Mallet du Pan, after the French Revolution, "philosophy may boast her reign over the country she has devastated." The action of the French Revolution against the church took many forms, but the most significant was the "Civil Constitution of the Clergy" of 1790. In it, a Gallicanism originally enunciated in the name of the absolute French monarchy attempted to subject the church to the National Assembly. The entire church in France was reorganized, with the authority of the pope restricted to doctrinal matters. Later in that year, a constitutional oath was required of all the French clergy, most of whom refused. Pope Pius VI (1775–99) denounced the "Civil Constitution" in 1791, and Catholic France was divided between the adherents of the papal system and the proponents of the new order. The closing decade of the 18th century was dominated by this conflict, and no resolution was provided by either church or state. The ultimate humiliation of the church came when Pius VI was driven out of Rome by the French armies in 1798 and in the following year was taken captive by them and dragged back to France, where he died. Not since the Great Schism and the Babylonian Captivity had the prestige of the papacy sunk so low.

The Concordat of 1801

Napoleon I—exportation of the Revolution. As it was obvious that the French Revolution itself had to be carried to some more permanent settlement, so it was recognized on all sides that a more stable arrangement of church–state relations was essential. This was achieved by Napoleon Bonaparte in a concordat concluded with Pope Pius VII on July 15/16, 1801. It recognized that Roman Catholicism was the faith of most Frenchmen and granted it freedom of worship. All incumbents of bishoprics were to resign and were to be replaced by bishops whom Napoleon, as first consul, would nominate. The properties of the church that had been secularized during the Revolution were to remain so, but the clergy were to be provided with proper support by the government. Many historians maintain that the Concordat of 1801 has been as decisive for modern church history as the conversion of Constantine had been for ancient church history. As Constantine first recognized and then established Christianity in the Roman Empire, so a series of concordats and other less formal agreements have created the modus vivendi between the church and modern secular culture. What this meant for the papacy was the realization that most of the temporal holdings of the church in Europe would have to be surrendered. The eventual outcome of this realization was to be the creation of Vatican City as a distinct political entity, but only after a long conflict over the States of the Church during the unification of Italy in 1869–70. First, however, came the period after the fall of Napoleon, when those who had emerged victorious at the Battle of Waterloo (1815) attempted to restore the previous condition. The Society of Jesus was revived in 1814, and the Congress of Vienna in 1814–15 helped to establish a basis for the recovery of the church during the 19th century. Temporary though these supposed settlements were, they made it clear to those living in the following period that the church would continue to be a force to be reckoned with in the affairs of Europe and America.

THE 19TH CENTURY

Much of the history of Roman Catholicism in the 19th century is identified with the pontificates of two men: Pius IX, who was pope for a third of a century (1846–78), and Leo XIII, who was pope for a quarter of a century (1878–1903).

Promulgation of the dogma of papal infallibility

The reign of Pius IX (1846–78). Few popes of modern times have presided over so momentous a series of decisions and actions as Pius IX. During his reign the development of the modern papacy reached a kind of climax with the promulgation of the dogma of papal infallibility. It had long been taught that the church, as "the pillar and bulwark of the truth," could not fall away from the truth of divine revelation and therefore was "indefectible" or even "infallible." Inerrancy had likewise been claimed for the Bible by both Roman Catholic and Protestant theologians. As the visible head of that church and as the authorized custodian of the Bible, the pope had also been thought to possess a special gift of the Holy Spirit, enabling him to speak definitively on faith and morals. But this gift had not itself been identified in a definitive way. The outward conflicts of the church with modern thought and the inner development of its theology converged in the doctrinal constitution *Pastor Aeternus* ("Eternal Shepherd"), promulgated by the first Vatican Council on July 18, 1870. It asserted that "the Roman Pontiff, when he speaks *ex cathedra,* that is, when in discharge of the office of pastor and teacher of all Christians, by virtue of his supreme apostolic authority he defines a doctrine regarding faith or morals to be held by the universal Church, by the divine assistance promised to him in blessed Peter, is possessed of that infallibility with which the divine Redeemer willed that his Church should be endowed." The decree was, of course, retroactive, even though there were historical incidents that appeared to contradict the retroactivity, such as the condemnation of Pope Honorius I by the third Council of Constantinople in 680, which were cited by opponents of the decree. This opposition was, however, ineffective, and the dogma of infallibility became the public doctrine of the church. Those who continued to disagree withdrew to form the Old Catholic Church, which was centred in The Netherlands, Germany, and Switzerland.

Even before the promulgation of this dogma, Pope Pius had exercised the authority that it conferred on him. In 1854, acting on his own prerogative and without any council, he defined as official teaching the doctrine of the immaculate conception of the Virgin Mary, "that the Most Blessed Virgin Mary, at the first instant of her conception, was preserved immaculate from all stain of original sin, by the singular grace and privilege of the Omnipotent God, in virtue of the merits of Jesus Christ." This put the church unequivocally on one side of a debate over the doctrine of Mary that had been going on since the Middle Ages. Ten years later, Pius issued a document that was in some ways even more controversial, the ***Syllabus of Errors*** (December 8, 1864). In it he condemned various "errors" characteristic of modem times, including pantheism, Socialism, civil marriage, secular education, and religious indifferentism. By thus appearing to put the church on the side of reaction against the forces of liberalism, science, democracy, and tolerance, the ***Syllabus*** seemed to be part of the retreat of Roman Catholicism from the modem world. At the same time, it did seek to clarify the identity of Roman Catholic teaching at a time when it was being threatened on all sides.

This combination of reactions to modern thought and society came to a head in the conflict over "Americanism," which was condemned by Leo XIII in 1899, and even more vigorously in the ***Kulturkampf*** (*i.e.*, struggle in Germany with Catholicism). Prince Otto von Bismarck, both because he was a Prussian and because he was a Protestant, resisted the basic trend of the developments just traced. In the Roman Catholic parties of the centre in the German states, he saw an obstacle to the form of German reunion to which he was dedicated, viz., a predominantly Protestant Germany without Roman Catholic Austria. The ***Syllabus of Errors*** and the dogma of infallibility represented the hostility of Roman Catholicism to the very sort of state he was trying to establish. Much of the theological opposition to papal infallibility came from German thinkers, notably Ignaz von Döllinger, to whose defense Bismarck sprang. The conflict between church and state came in several principal areas. The ***Kulturkarnpf*** began with the exclusion of the Roman Catholic Bureau from the Ministry of Culture and Cultus in the Prussian state. Bismarck asserted the authority of the state over all education in Prussia and had the Society of Jesus expelled. Then, in direct defiance of the ***Syllabus of Errors,*** he required civil marriage of all, regardless of whether or not they had also exchanged their vows before a clergyman. Laws were passed compelling candidates for the Roman Catholic priesthood to attend a German university for at least three years. Bismarck summarized his defiance of the Pope in an allusion to the conflict between Pope Gregory VII and Emperor Henry IV in the 11th century: "We are not going to Canossa!" When Pius IX died in 1878, the conflict was still unresolved.

Rise of the Modernist movement

The reign of Leo XIII (1878–1903). Although Leo XIII was no less conservative in his theological inclinations than his predecessor, his positive appreciation of the church's opportunities in modem society gave his pontificate a significantly different cast from that of Pius. On issues of church doctrine and discipline, his administration was a strict one. It was during his reign that the Modernist movement, which advocated the use of biblical and historical criticism and freedom of conscience, arose within Roman Catholicism; and, although the formal condemnation of its tendencies did not come until 1907, four years after his death, he had made his opposition to this trend clear by the establishment of the Pontifical Biblical Commission as a monitor over the work of scriptural scholars. The positive side of his theology came to voice in the encyclical ***Aeterni Patris*** ("Eternal Father") of August 4, 1879, which, more than any other single document, provided a charter for the revival of Thomism (the medieval theological system based on the thought of Thomas Aquinas) as the official philosophical and theological system of the Roman Catholic Church. It was to be normative not only in the training of priests at the seminaries of the church but also in the education of the laity at universities. To this end Leo also sponsored the launching of a definitive critical edition of the works of Thomas Aquinas. In 1895 Pope Leo appointed a commission to decide the long-mooted question whether, despite the separation from Rome in the 16th century, the priestly ordination of the Anglican communion was valid, as, for instance, that of the separated Eastern churches was; in 1896 he issued ***Apostolicae Curae*** ("Apostolic Concerns"), which denied the validity of Anglican orders and was a setback for ecumenical hopes on both sides.

The "pope of peace"

Nevertheless, Leo XIII is best remembered for his social and political thought, which earned him the sobriquet the "pope of peace." He managed to mollify the church's position toward the policies of Bismarck, and the Chancellor in turn moved toward a compromise. Diplomatic relations between Germany and the Vatican were restored in 1882, and gradually the restrictive laws were lifted. But the greatest achievements of Leo's work in the relation between the church and modern culture were his social and political encyclicals. Without repudiating the theological presuppositions of the ***Syllabus of Errors,*** these encyclicals articulated a positive social philosophy, not merely a defensive one. In ***Libertas*** ("Liberty"), an encyclical issued on June 20, 1888, he sought to affirm what was good about political liberalism, democracy, and freedom of conscience. Above all, the encyclical ***Rerum*** *Novarum* ("Of New Things") of 1891 put the church on the side of the modern struggle for social justice. Though rejecting the program of 19th-century Socialism, the Pope was also severe in his condemnation of an exploitative laissez-faire capitalism and in his insistence upon the duty of the state to strive for the welfare of all its citizens. The social thought of Leo XIII helped to stimulate concrete social action among Roman Catholics in various lands, such as the Christian Social Movement. When he died, soon after the close of the 19th century, the church seemed in many ways to be entering a new era of respect and influence, but the turmoil of war, depression, and revolution in the 20th century intervened.

THE 20TH CENTURY

Two historical forces, one external and the other internal, have dominated the development of Roman Catholicism during the 20th century: the world wars of 1914–18 and 1939–45, with the accompanying upheavals of politics, economics, and society; and the second Vatican Council of 1962–65, with upheavals no less momentous in the life and teaching of the church.

The period of the world wars. Pope Pius X (1903–14) symbolized the transition from the 19th century to World War I. In his encyclical, ***Pascendi Dominici Gregis*** ("Feeding the Lord's Flock"), of September 8, 1907, he formally condemned Modernism as "the résumé of all the heresies," and, in 1910, he prescribed that clergy and seminary professors take an oath abjuring Modernism and affirming the correctness of the church's teachings about revelation, authority, and faith. He sponsored the revision and clarification of the code of canon law. More perhaps than any of his immediate predecessors or successors, Pius X gave attention to the reform of the church's liturgy, especially to the Gregorian chant and advocated early and frequent reception of communion. Yet hanging like a cloud over his pontificate was the growing threat of the world war, which neither diplomacy nor piety was able to forestall. The last major document issued by Pius X was a lament over the outbreak of war, dated August 2, 1914; less than three weeks later he was dead.

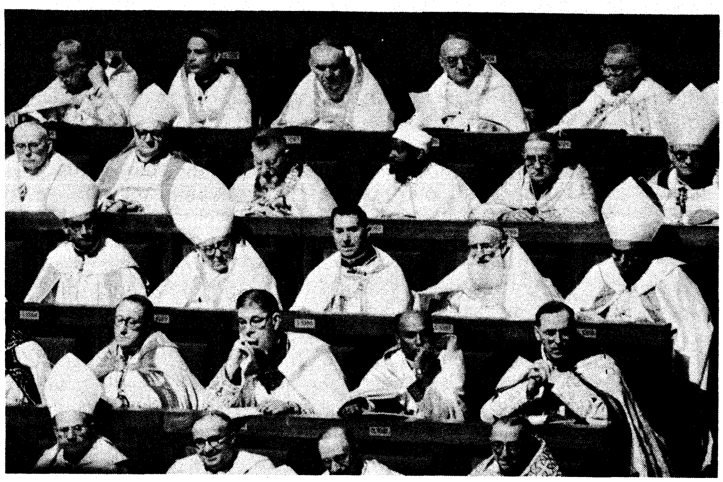
World War I, often called the real end of the 19th century, was also a major turning point in modern Roman Catholic history. Ever since ancient times, the church had been accustomed to order its relations to human society by negotiations with kings and emperors, preferably members of its own fellowship. The war and the revolutions attending it meant the end of the Hohenzollern (Germany), Habsburg (Austria-Hungary), and Romanov (Russia) dynasties, obliging the church to come to terms with the new realities of democratic, Communist, and Fascist regimes.

The Lateran Treaty

Of special significance was a series of pacts with the Fascist Italy of Benito Mussolini. In 1929 the church and the Italian government signed the Lateran Treaty, which finally regularized relations between them and gave Vatican City independent status. In 1933 the church went on to conclude a concordat with Nazi Germany, hoping to protect its own interests and those of minorities; but this hope proved to be ill founded, and the church's relation with Hitler and his regime deteriorated. Although Pius XI (1922–39) and Pius XII (1939–58) both spoke out several times against the excesses of the regime, they did little to restrain it. The papacy spoke out much more often, for example, during the Spanish Civil War (1936–39), against the dangers of Communism, the eventual dominance of which over Poland, Hungary, and other strongly Roman Catholic lands was a major setback to the church of the 20th century. As a diplomat and former papal secretary of state, Pope Pius XII was obliged, under the pressures of World War II, to clarify and redefine the church's teachings on war and peace, as well as to work out a strategy of survival. In 1950 he became the first pope since the first Vatican Council to exercise the right of defining doctrine, proclaiming the bodily assumption of the Virgin Mary to be a dogma binding on all members of the church. Earlier in that same year, in the encyclical ***Humani Generis*** ("Of the Human Race"), he had given a reproof to various theological trends that appeared to be reviving the ideas and methods of Modernism.

Vatican II. From these two papal promulgations of 1950 many observers were ready to conclude that, in the second half of the 20th century, Roman Catholicism would assume an essentially defensive posture in relation to the modern world. Those who had come to that conclusion were compelled to revise it by the pontificate of John XXIII (1958–63) and by the second Vatican Council (1962–65). During his brief reign, Pope John issued several important encyclicals. Of special interest was ***Mater et Magistra*** ("Mother and Teacher"), published in 1961, which explicitly attached itself to the ***Rerum Novarum*** of Leo XIII in calling for justice and the common good as the norms of social conduct. Two years later, in ***Pacem in Terris*** ("Peace on Earth"), the Pope addressed himself not only to members of the church but to "all men of good will." In this encyclical he formulated, more completely than any previous pope had done, a social philosophy for peace among men and between nations. This spirit of reform and concern came to expression in the council, which Pope John convoked but which he did not live to see to its conclusion. The council brought about drastic changes in the life and worship of the church, encouraging the use of the vernacular in the liturgy and greater lay participation everywhere. Perhaps even more historic were its actions toward those outside the borders of the Roman Catholic Church. To Eastern Orthodox and Protestant Christians, it extended the hand of fraternal understanding instead of denouncing them as heretics. To the Jewish community, it addressed words of reconciliation and regret for the anti-Semitism of the Christian past. To the world religions, it spoke of the church's admiration for the spiritual values that had been preserved in those traditions that did not know the name of Christ. And to all men, believers and unbelievers, it expressed its respect for the integrity and freedom of man and its repudiation of coercion as a means for bringing men to faith. In its importance for the development of the church, the second Vatican Council will probably rank with the councils of Nicaea (325), Chalcedon (451), and Trent (1545–63). **(J.J.Pe.)**

Influence of Vatican II



Delegates to the second Vatican Council, 1962.
NC Photos—KNA

III. Roman Catholicism outside Europe in modem times

THE NEW WORLD: THE SPANISH AND PORTUGUESE EMPIRES

Colonial period. The Western Hemisphere was discovered by Europeans immediately before the Protestant Reformation began in Europe. The fact of that discovery at that moment in history and the original development of the New World by Roman Catholic empires (*e.g.*, Spain) is of major significance in the religious history of the hemisphere. The only part of it that was to be non-Catholic in its general cultural outlook was the area of those colonies that was to become the United States. Spain and Portugal were in their prime as sea powers in the late 15th and early 16th centuries, and they were most responsible for exploring, colonizing, and establishing the Christian faith in the southern two-thirds of the American half of the world.

Relations between missionaries, colonists, traders, and indigenous peoples

The chief institutions for Catholicizing were the Franciscans, Dominicans, Augustinians, Jesuits, and other religious orders. Well-trained and self-sacrificing representatives of the orders were able to go wherever Spanish and Portuguese ships went. Sometimes they could be accused of serving as religious supporters of anything the Crown desired. Because the missionaries were in quest of souls, however, even though the imperial powers often wanted merely to exploit the bodies and material resources of New World natives, there were also clashes between Catholic churchmen and colonizers or traders.

The establishment of Catholicism in Central and South America, then, did not always mean that there was a moderation of the empires' exploitations of the natives, who came to be called Indians. There were, however, some missionary efforts and successes among these natives. At times Catholicism was able to temper the inhumanity of the conquerors. Best known among the humane spokesmen for Indians was the Dominican Bartolomé de Las Casas (1474–1566), "the Apostle of the Indies."

In the course of the 16th through the 19th centuries, European colonists and immigrants from nations other than Spain and Portugal came to Latin America. Even when these movements were made up of Protestant minorities or when they included Protestant missionaries, however, they did little to disrupt the generally or nominally Catholic cultures.

Modern secular forces also served to jostle the Catholic settlements. The case of Mexico is illustrative; time and again its ruling powers have proscribed Catholic education and embodied anticlerical interests. Still, the Mexican people remained largely Catholic and blended some of their inherited native religious values and practices with distinctively Catholic forms. In Mexico as elsewhere, then, it is difficult to know just who is Catholic and who is non-Catholic. Estimates of the number of nominal Catholics in Latin America run as high as 200,000,000.

After independence. The inevitable reaction by Catholic and non-Catholic alike arose against the colonial powers. This took the form of movements of independence, anticlerical revolts that were directed against European powers. Some institutions, particularly those devoted to education, were opposed to the practices of Catholicism. Because so many of the clergy came from Europe, anti-European sentiment assured that the American fields were not attractive, and chronic clerical shortages prevailed. As was the case in Europe, the various revolutions were often concurrent with or encouraging to the various versions of Enlightenment thought, and this meant that they were expectably uncongenial to the truth claims of Christianity.

Effects of anti-clericalism and revolutionary Catholicism

By the middle of the 20th century, wherever Latin American Catholicism remained strong, it still tended to identify itself with increasingly reactionary national regimes and was then dismissed by much of the rest of the world as appearing to be uncongenial to the legitimate aspirations of majorities. Because of the cosmopolitan influences of the second Vatican Council (1962–65), however, the self-generated renewal of the church, and the presence of a new socially responsible leadership, there appeared during the 1960s a more radical Catholicism. Dom Helder Câmara of Recife, Brazil, exemplified the impulse toward drastic social reform. Camillo Torres, killed in the role of a Colombian guerilla, typified the association of a Catholic minority with violent revolutionary programs.

Spanish and French missions in North America. Though at the time of its settlement the United States under British and continental Protestant influences became a largely Protestant outpost, Spanish Catholics did establish missions in Florida and elsewhere. Franciscans began work in California in 1514 and in New Mexico in 1581; this work reached its greatest success when the Spanish missionary Fra Junipero Serra founded stations all along the California coast after 1769. Similarly, to the north, French explorers, traders, and conquerors settled much of eastern Canada and brought with them a Catholic Church that has remained dominant there until the present. French missionaries also penetrated the Great Lakes region and the Mississippi Valley, but their efforts left few traces when the North American interior came to be settled by English-speaking people late in the 18th century.

ROMAN CATHOLICISM IN THE UNITED STATES AND CANADA

United States. As far as the 13 colonies of the emerging United States were concerned, only Maryland, which had been settled in 1634 and established in 1649, included an appreciable number of Catholics before American indeyendence. Catholics were often unwelcome in and even excluded from many colonies, where Congregational or Episcopal churches were supported by law. According to some estimates, there were at most 25,000 Catholics in a colonial population of almost 4,500,000 at the time of independence after 1776.

Relationship of Catholics to Protestants

From the first, however, Catholic leadership enjoyed its place in the free society of the new United States. Bishop John Carroll, a representative of a notable colonial Catholic family, pioneered in exploring positive relations between Catholic religionists and their fellow citizens. Beginning in the 1830s and 1840s, the assurances of religious freedom were added attractions for millions of Catholic immigrants who had to make their way to the United States for economic reasons. Coming as most of them did from Ireland or the European continent to a nation of largely British and almost exclusively Protestant provenance, they awakened suspicion and hostility and were met by what has since been called a nativist Protestant Crusade.

Catholicism endured, however, and built impressive institutions including parochial schools. These elementary and secondary schools were formed late in the 19th century because Catholic leaders feared Protestant influences in the public schools. Through these Catholic agencies, Catholic leaders were able to help their people sustain religious loyalties to Rome and civil loyalties to America. The church was plagued by several issues: "trusteeism," a debate over lay versus clerical control of ecclesiastical institutions; "Americanism," the charge that American Catholics were innovating in doctrine and practice; immigration; and the rescue of souls. The Church prospered through all these adversities.

After World War I, anti-Catholicism declined. By 1960 a Roman Catholic, John F. Kennedy, had become president —an office previously thought to be out of range for Catholics. Tensions over church–state issues remained, but these were minimized or at least they grew more confused because neither Catholics nor their old opponents continued to present a united front. The ecumenical age also brought about better relations between the various faiths.

Canada. Farther north, in Canada, England came to dominance in 1713, but the Quebec Act of 1774 guaranteed Catholic rights. The period of new nationalisms after World War II found French Catholics in Quebec nervous about the assimilation and even possible disappearance of their culture. They took steps to assure the perpetuation of the faith, language, and outlook of the French-speaking Catholic millions in an otherwise largely Protestant nation. Some militant movements even asked for separation and the formation of a new nation in Quebec.

ROMAN CATHOLICISM IN AFRICA AND ASIA

Though Catholicism had shaped Latin American and eastern Canadian culture and though it came to be at home in the United States, it also found itself to be a worldwide presence for the first time in the 19th century. This expansion was the result both of Western nations' imperial intrusions into Africa and Asia and of the rebirth of a missionary spirit in Christendom.

Catholicism as a worldwide presence

Some of the expansive efforts may have been built upon the traces of 16th-century missionary activities, such as those of St. Francis Xavier, a Jesuit missionary to Asia; usually, however, they had to develop on the basis of original methods and in new territories.

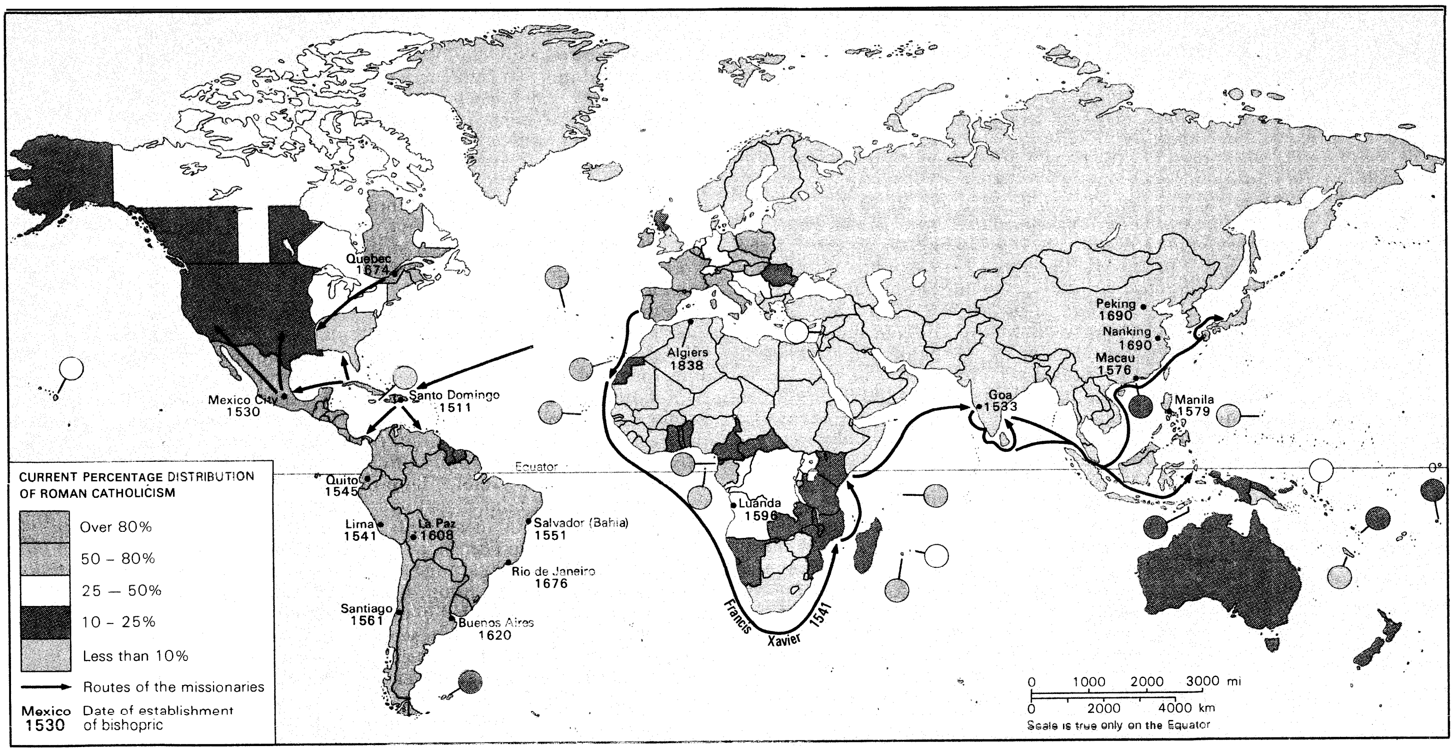
Early missions. ***Africa.*** In Africa almost nothing remained of the strong early Christian presence in the north. Through the centuries North Africa had become largely Muslim. The Muslim presence there offered more resistance than did native African religionists who lived on in the rest of the continent. Christians were not welcomed and were often persecuted. Even in partly Christian Abyssinia (Ethiopia), where the Coptic Church was prominent, Catholics were largely excluded except between 1702 and 1839. An archbishopric was established in Algiers, and in 1868 Archbishop Charles Lavigerie founded the White Fathers who were energetic but largely unsuccessful missionaries from that base.

West Africa presented obvious and persistent problems for all Christians, because it was from there that European nations had carried on most of the slave trade. Portuguese colonialists did help the Catholic Church establish itself in parts of West Africa, but progress was slow. Catholicism fared better in East Africa, particularly in Madagascar and around Lake Victoria. Uganda, Kenya, and Tanganyika (now Tanzania), for example, have thriving churches. The record was less triumphant farther south, in no small measure because of Dutch and British Protestant power. Yet there, as elsewhere, independent missionary societies worked desyite considerable hardshiy.

Asia. In Asia, Catholicism was able to profit from Portuguese and Spanish adventures from the 16th century on. In that part of the world, however, different styles of clashes occurred. Asians had not had contact, as Muslims had, with biblical views of history and destiny. Buddhists, Taoists, followers of Confucianism, and Hindus were devoted to world views uncongenial to Western attitudes toward God, time, and history. In the encounter, Catholicism was itself torn over debates concerning the permissible degrees of accommodation to Eastern ways and views of life, rituals, and terms.

Problems of accommodation

In India there were traces of missionary extensions from premodern centuries (*e.g.*, the Malabar Syrian Christians), and Catholicism here and there found new bases. But the suppression of the Jesuits in 1773 for reasons of European politics removed the most assertive group from



Routes of missionaries, dates of establishment of dioceses, and current distribution of Roman Catholicism,

the scene at the most inopportune moment. Catholics flourished under persecution in Indochina, in what is now called Vietnam. The major drama occurred, however, in China and Japan, which were opened to Westerners after centuries of relative isolation. In the 19th century, Catholic institutions became familiar sights on the Chinese landscape. Churches, hospitals, and schools were established. The Boxer Rebellion in 1900 symbolized the growing resistance of the Chinese to Western presences in their country.

In Japan little was left of the 16th-century missions except for an isolated sect of Catholics on an island near Nagasaki. In both China and Japan only a small percentage of the people ever became Catholic. The triumph of Communists in China in 1949 brought the end of Catholic missionary activity and proscriptions against native Catholic practices. Postwar Japan saw Catholicism engulfed by resurgent religions and a new secular spirit.

The Roman Catholic Church in Asia and Africa in the 20th century. *The development of indigenous clergy and native institutions.* Foreseeing some of the 20th-century difficulties, thoughtful Catholics began during the 19th century to argue that Western religions were not able to be appropriated directly and may not long be permitted in many places. Therefore, they began to advocate the development of indigenous clergy. The resultant native institutions often blended some elements of local cultures, but seldom were fusions of distinctive elements of Asian or African religion with Christian doctrine consciously permitted.

Conflicts and relations with national governments. If the recent centuries represented much promise for Catholicism's self-definition as a universal church, they also meant setbacks. Christians of the West had often exploited the developing nations, looted their resources, enslaved or demeaned their populations, and extirpated their religions and cultures. As colonial yokes were thrown off, new nations in quest of their own identities encouraged the renewal of the non-Christian religions that had long been part of their cultures. The Western Catholic could serve as a bogey. Overt anti-Christianity of most Marxist or Communist parties in these countries meant a rolling back of Catholicism.

The new world consciousness of Roman Catholicism. The second Vatican Council also saw the definition of more positive views of non-Catholic high religions, a fact that served somewhat to diminish the impulse to convert the whole world to explicit faith in Christ and obedience to Rome. Catholicism engaged in internal reforms that suggested a new responsiveness to revolutionary social situations. At least minor new local adaptations in Asian and African churches were permitted, and Western imperial pride was specifically condemned by modern popes. Although the dominating and conversionist impulses do not seem to have wholly died, nevertheless, in the majority of the world's nations Catholics have shown themselves more ready than ever before to be brothers to adherents of other religions and to have a new regard for secular human development. (M.E.M.)

Effectsof the second Vatican Council

BIBLIOGRAPHY

The Latin Church in the West (1000–1517): The only large-scale work covering the whole period (except for the century 1274–1378) is AUGUSTIN FLICHE and VICTOR MARTIN (eds.), *Histoire de l'Église depuis les origines jusqu'à nos jours*, vol. 8–10, 12–14 (1946–64). An earlier classic is ALBERT HAUCK, *Kirchengeschichte Deutschlands* (1887–1920), which covers most of continental Europe. Neither of these has been translated into English. There are useful chapters in *The Cambridge Medieval History*, vol. 5–7 (1929–32). HORACE K. MANN, *The Lives of the Popes in the Early Middle Ages*, 18 vol. (1902–32), is narrow in scope. The only history of medium size is DAVID KNOWLES, *The Christian Centuries*, vol. 2, *The Middle Ages* (1969), with bibliography. For a perceptive introduction, see RICHARD W. SOUTHERN, *The Making of the Middle Ages* (1953). GEOFFREY BARRACLOUGH, *Mediaeval Germany*, 911–1250, 2 vol. (1961), is the only adequate work in English. See also R.W. and A.J. CARLYLE, *A History of Mediaeval Political Theory in the West*, 3rd ed., vol. 3–5 (1938–50). ETIENNE GILSON, *History of Christian Philosophy in the Middle Ages* (1955), is a masterly summary with full bibliography to date; DAVID KNOWLES, *The Monastic Order in England*, 2nd ed. (1963), also covers part of Europe. Other recommended studies include: HENRY C. LEA, *The Inquisition of the Middle Ages*, a partial edition with introduction by WALTER ULLMANN (1963); GUILLAUME MOLLAT, *Les Papes d'Avignon, 1305–1378*, 9th ed. (1949; Eng. trans., *The Popes at Avignon*, 1305–1378, 1963); WALTER ULLMANN, *The Growth of Papal Government in the Middle Ages*, 3rd ed. (1970), indispensable, but in parts controversial; and SCHAFER WILLIAMS (ed.), *The Greporian Epoch* (1964), a useful collection of reprinted studies by early and recent authorities.

The late Middle Ages: WILLY ANDREAS, *Deutschland vor*

der Reformation, 6th ed. (1959), a presentation of history and culture at the end of the Middle Ages; GEORGE CLARK (ed.), *The Oxford History of England:* vol. 5, MAY MCKISACK, *The Fourteetzth Century,* 1307–1399 (1959); and vol. 6, E.F. JACOB, *The Fifteenth Century, 1399–1485* (1961); FLICHE-MARTIN, *Histoire de l'Église:* vol. 15, ROGER AUBENAS and ROBERT RICARD, *L'Église et la Renaissance, 1449–1517* (1951); vol. 16, E. DE MOREAU, PIERRE JOURDA, and PIERRE JANELLE, *La Crise religieuse du XVI^e siècle* (1950); and vol. 17, L. CRISTIANI, *L'Église à l'époque du concile de Trente* (1948); GEORGES DE LAGARDE, *La Naissartce de l'esprit laïque, au déclin du Moyen Age,* 3rd ed., vol. 1–5 (1956–63), a study of the lay movement in the Middle Ages; JOSEPH LORTZ, *Geschichte der Kirche in ideengeschichtlicher Betrachtung,* 22nd–23rd ed., 2 vol. (1965; Eng. trans. of 5th–6th ed., *History of the Church,* 1939), the history of the church from the history of ideas point of view; and "Zur Problematik der kirchlichen Missstande im Spat-Mittelalter," in *Trierer Theologische Zeitschrift* (1949), a presentation of the situation of the church in the late Middle Ages; HEIKO AUGUSTINUS OBERMAN, *The Harvest of Medieval Theology: Gabriel Biel and Late Medieval Nominalism,* rev. ed. (1967), a study of the theology of the late Middle Ages in its entirety, with special emphasis on Nominalism; WILLIAM A. PANTIN, *The English Church iiz the Fourteenth Century* (1955); REGNERUS RICHARDUS POST, *The Modern Devotion: Confrontation with Reformatioiz and Humanism* (1968), a history of the Brothers of the Common Life and their confrontation with Humanism and the Reformation; BRIAN TIERNEY, *Foundations of the Conciliar Theory: The Contributions of the Medieval Canonists from Gratian to the Great Schism* (1955), on the conciliar theories.

Roman Catholicism in Europe irt modern times: CONRAD BERGENDOFF, *The Church of the Lutheran Reformation* (1967), a study of Reformation history from its beginnings to the 20th century; INTERNATIONAL COMMITTEE OF HISTORICAL SCIENCES, *Bibliographie de la Réforme,* 1450–1648, vol. 1–7 (1958–70), a reference work for the investigation of the history of the Reformation; *The New Cambridge Modern History:* vol. 1, G.R. POTTER (ed.), *The Renaissance, 1493–*1520 (1957); and vol. 2, GEOFFREY R. ELTON (ed.), *The Reformation,* 1520–1559 (1958), a reference work; ARTHUR G. DICKENS, *Reformation and Society in Sixteenth-Century Europe* (1966), an account of the sociological relationships in the 16th century; GEOFFREY R ELTON, *Refortnation Europe, 1517–1559* (1963); HAROLD JOHN GRIMM, *The Reformation Era,* 1500–1650 (1964; with rev. bibliography, 1965), a study of the Reformation and the Counter-Reformation; HUBERT JEDIN (ed.), *Handbuch der Kirclzengeschichte:* vol. 3, pt. 1, *Die mittelalterliche Kirche* (1966); vol. 3, pt. 2, *Voin kirchlichen Hochmittelalter bis zum Vorabend der Reforznation* (1968); and vol. 4, *Reformation, katholische Reform und Gegenreforrnation* (1967); PHILIP HUGHES, *The Reformation in England,* 3 vol. (1950–54); JOSEPH LORTZ, *Wie Kam es zur Reformation?,* 3rd ed. (1955; Eng. trans., *How the Reformation Came,* 1964), on the causes of the Reformation in England; *Die Reformation als religiöses Anliegen heute* (1948; Eng. trans., *The Reformation:* A *Problem for Today,* 1964), thesis for ecumenical discussions; and *Die Refortization in Deutschland,* 5th ed., 2 vol. (1965; Eng. trans., *The Reformation in Germany,* 2 vol., 1968), standard work on the history of the Reformation; JAROSLAV PELIKAN, *Obedient Rebels: Catholic Substance and Protestant Principle in Luther's Reformation* (1964), an investigation of Luther's thought; SIR MAURICE POWICKE, *The Reformation in England* (1958); GOLO MANN and AUGUST NITSCHKE (eds.), *Propylaen Weltgeschichte,* vol. 7, *Von der Reformation zur Revolution* (1964); KARL SCHOTTENLOHER, *Bibliographie zur deutscken Geschichte im Zeitalter der Glaubensspaltung,* 1571–1585, 2nd ed., 7 vol. (1956–58, 1966), a reference work; HERBERT MAYNARD SMITH, *Henry VIII and tlze Reformation* (1962); GEORGES HUNTSTON WILLIAMS, *The Radical Reformation* (1962), a synoptic presentation of efforts for reform on the part of the "left wing" from the beginning until the end of the 16th century; JOHN D. MACKIE, *The Earlier Tudors, 1485–1558* (1952), textbook for the study of the history of the English Reformation; ERNST WALTER ZEEDEN, *Die Entstehung der Konfessionen* (1965), on the problems of the formation of the confessions of faith at the time of the development of different denominations; and *Das Zeitalter der Gegenreformation* (1967) covers the battle for the reorganization of the Roman Church.

Roman Catholicism outside Europe in modern times: The standard work on church history is KARL BIHLMEYER and HERMANN TUCHLE, *Kirchengeschichte,* 17th ed., 3 vol. (1951–62; Eng. trans., *Church History,* 3 vol., 1958–66). The documents of Roman Catholicism are conveniently assembled in HEINRICH J. DENZINGER, *Enchiridion symbolorum,* 33rd ed. (1965). ROLAND H. BAINTON, *The Horizort History of Christianity* (1964), is a well-written, beautifully illustrated, comprehensive introduction to Western Christianity through the centuries and includes references to modern worldwide Catholicism. Much more extensive and valuable, especially because of the excellent bibliographical clues it provides, is KENNETH SCOTT LATOURETTE, *Christianity in a Revolutiortary Age: A History of Christianity in the Nineteenth and Twentieth Centuries,* 5 vol. (1958–62). Volumes 1, 3, and 5 concentrate on Roman Catholic themes. The author was an advocate of the worldwide missionary activities of the church, and a bias in favour of Christian expansion colours his work. AUGUST FRANZEN and JOHN P. DOLAN, *Kleine Kirchengeschichte,* 2nd ed. (1968; Eng. trans., *A History of the Church,* 1969), is a convenient brief introduction that includes some modern materials. E.E.Y. HALES, *The Catholic Chrrrch in the Modern World* (1958), concentrates on the European and American settings. A number of works dealing with Catholicism outside that orbit fall into the category of histories of colonialism or missions. First to be noted are two works by STEPHEN C. NEILL: *Colonialism and Christian Missions* (1966) and *A History of Christian Missions* (1964). Though Neill is non-Catholic, he provides brief and generally fair comments on Catholic ventures. A much more conservative Protestant bias is measurable in the standard work by ROBERT HALL GLOVER, *The Progress of World-Wide Missions,* rev. ed. (1960). ROBERT L. DELAVIGNETTE, *Christianisme et colonialisme* (1960; Eng. trans., *Christianity and Colonialism,* 1964), is written by a Catholic and concentrates on Catholic experience, but in too brief a scope. Two works that make aspects of the American Catholic experience readily available to readers are JOHN TRACY ELLIS, *American Catholicism,* 2nd ed. rev. (1969); and the somewhat less adequate THEODORE MAYNARD, *The Story of American Catholicism,* 2 vol. (1960).

(J.L./M.E.M./M.D.K./J.J.Pe.)

Romance (Literature)

The subject of this article is medieval romance as it was developed in western Europe from the 12th century onward. It contains a brief history of romance as a literary form, of its extension and influence throughout western Europe up to and after the Renaissance, and of its re-emergence in the 18th century.

The Old French word *romanz* originally meant "the speech of the people," or "the vulgar tongue," in contrast with the written form of literary Latin. Its meaning then shifted from the language in which the work was written to the work itself. Thus, an adaptation of Geoffrey of Monmouth's *Historia regum Britanniae* (*c.* 1137), made by Wace of Jersey in 1155, was known as *Li Romanz de Brut,* while an anonymous adaptation (of slightly later date) of Virgil's *Aeneid* was known as Li *Romanz d'Enéas;* it is difficult to tell whether in such cases *li romanz* still meant "the French version" or had already come to mean "the story." It soon specialized in the latter sense, however, and was applied to narrative compositions similar in character to those imitated from Latin sources but totally different in origin; and, as the nature of these compositions changed, the word itself acquired an increasingly wide spectrum of meanings. In modern French a *roman* is just a novel, whatever its content and structure; while in modern English the word "romance" (derived from Old French *romanz)* can mean either a medieval narrative composition or a love affair, or, again, a story about a love affair, generally one of a rather idyllic or idealized type, sometimes marked by strange or unexpected incidents and developments; and "to romance" has come to mean "to make up a story that has no connection with reality."

For a proper understanding of these changes it is essential to know something of the history of the literary form to which, since the Middle Ages, the term has been applied. The account that follows is intended to elucidate historically some of the ways in which the word is used in English and in other European languages.

THE COMPONENT ELEMENTS

The romances of love, chivalry, and adventure produced in 12th-century France have analogues elsewhere, notably in what are sometimes known as the Greek romances — narrative works in prose by Greek writers from the 1st century BC to the 3rd century AD. The first known, the

Greek romances

fragmentary Ninus romance, in telling the story of the love of Ninus, mythical founder of Nineveh, anticipates the medieval *roman d'antiquite'*. A number of works by writers of the 2nd and 3rd centuries AD—Chariton, Xenophon of Ephesus, Heliodorus, Achilles Tatius, and Longus—introduce a theme that was to reappear in the *roman d'aventure:* that of faithful lovers parted by accident or design and reunited only after numerous adventures. Direct connection, however, can be proved only in the case of the tale of *Apollonius of Tyre*, presumably deriving from a lost Greek original but known through a 3rd- or 4th-century Latin version. This too is a story of separation, adventure, and reunion, and, like the others (except for Longus' pastoral *Daphnis and Chloë*), it has a quasi-historical setting. It became one of the most popular and widespread stories in European literature during the Middle Ages and later and provided Shakespeare with the theme of *Pericles.*

Romance style and subject matter. But the real debt of 12th-century romance to classical antiquity was incurred in a sphere outside that of subject matter. During the present century, scholars have laid ever-increasing emphasis on the impact of late classical antiquity upon the culture of medieval Europe, especially on that of medieval France. In particular, it is necessary to note the place that rhetoric (the systematic study of oratory) had assumed in the educational system of the late Roman Empire. Originally conceived as part of the training for public speaking, essential for the lawyer and politician, it had by this time become a literary exercise, the art of adorning or expanding a set theme: combined with grammar and enshrined in the educational system inherited by the Christian Church, rhetoric became an important factor in the birth of romance. Twelfth-century romance was, at the outset, the creation of "clerks"—professional writers who had been trained in grammar (that is to say, the study of the Latin language and the interpretation of Latin authors) and in rhetoric in the cathedral schools. They were skilled in the art of exposition, by which a subject matter was not only developed systematically but also given such meaning as the author thought appropriate. The "romance style" was, apparently, first used by the authors of three *romans d'antiquité*, all composed in the period 1150–65: *Roman de Thebes*, an adaptation of the epic *Thebaïs* by the late Latin poet Statius; *Roman d' Enéas,* adapted from Virgil's *Aeneid;* and *Roman de Troie,* a retelling by Benoît de Sainte-Maure of the tale of Troy, based not on Homer (who was not known in western Europe, where Greek was not normally read) but on 4th- and 5th-century Latin versions. In all three, style and subject matter are closely interconnected; elaborate set descriptions, in which the various features of what is described are gone through, item by item, and eulogized, result in the action's taking place in lavish surroundings, resplendent with gold, silver, marble, fine textiles, and precious stones. To these embellishments are added astonishing works of architecture and quaint technological marvels, that recall the Seven Wonders of the World and the reputed glories of Byzantium. *Troie* and *Enéas* have, moreover, a strong love interest, inspired by the Roman poet Ovid's conception of love as a restless malady. This concept produced the first portrayal in Western literature of the doubts, hesitations, and self-torment of young lovers, as exemplified in the Achilles–Polyxena story in *Troie* and in the Aeneas–Lavinia story in *Enéas.* Yet even more important is the way in which this new theme is introduced: the rhetorical devices appropriate to expounding an argument are here employed to allow a character in love to explore his own feelings, to describe his attitude to the loved one, and to explain whatever action he is about to take.

Earliest works in the "romance style"

Developing psychological awareness. As W.P. Ker, a pioneer in the study of medieval epic and romance, observed in his *Epic and Romance* (1897), the advent of romance is "something as momentous and as far-reaching as that to which the name Renaissance is generally applied." The Old French poets who composed the chansons de geste (as the Old French epics are called) had been content to tell a story; they were concerned with statement, not with motivation, and their characters could act without explicitly justifying their actions. Thus, in what is one of the earliest and certainly the finest of the chansons de geste, the *Chanson de Roland* (c. 1100), the hero's decision to fight on against odds—to let the rear guard of Charlemagne's army be destroyed by the Saracen hordes in the hopeless and heroic Battle of Roncesvalles rather than sound his horn to call back Charlemagne—is not treated as a matter for discussion and analysis: the anonymous poet seems to take it for granted that the reader is not primarily concerned with the reason why things happened as they did. The new techniques of elucidating and elaborating material, developed by romance writers in the 12th century, produced a method whereby actions, motives, states of mind, were scrutinized and debated. The story of how Troilus fell in love with Briseïs and how, when taken to the Grecian camp, she deserted him for Diomedes (as related, and presumably invented, by Benoît de Sainte-Maure in his *Roman de Troie)* is not one of marvellous adventures in some exotic fairyland setting: it is clearly a theme of considerable psychological interest, and it was for this reason that it attracted three of the greatest writers of all time: Boccaccio in his *Filostrato* (*c.* 1338), Chaucer in his *Troilus and Criseyde* (before 1385), and Shakespeare in his *Troilus and Cressida* (*c.* 1601–02). With the 12th-century pioneers of what came to be called romance, the beginnings of the analytical method found in the modern novel can easily be recognized.

The development of more sophisticated techniques

Sources and parallels. Where exactly medieval romance writers found their material when they were not simply copying classical or pseudo-classical models is still a highly controversial issue. Parallels to certain famous stories, such as that of Tristan and Iseult, have been found in regions as wide apart as Persia and Ireland: in the mid-11th-century Persian epic of *Wis and Ramin* and in the Old Irish *Diarmaid and Grainne;* but while in the latter case it is possible to argue in favour of a genetic link between the two traditions, the former is more likely to be a case of parallel development due, on the one hand, to the inner logic of the theme and, on the other, to certain similarities in the ideological and social background of the two works. Failure to maintain the essential distinction between source and parallel has greatly hindered the understanding of the true nature of medieval romance and has led to the production of a vast critical literature the relevance of which to the study of the genre is at best questionable.

The marvellous as a romance element. The marvellous is by no means an essential ingredient of "romance" in the sense in which it has been defined. Yet to most English readers the term romance does carry implications of the wonderful, the miraculous, the exaggerated, and the wholly ideal. Ker regarded much of the literature of the Middle Ages as "romantic" in this sense—the only types of narrative free from such "romanticizing" tendencies being the historical and family narrative, or Icelanders' sagas developed in classical Icelandic literature at the end of the 12th and in the early 13th century. The *Chanson de Roland* indulges freely *in* the fantastic and the unreal: hence Charlemagne's patriarchal age and preternatural strength (he is more than 200 years old when he conquers Spain); or the colossal numbers of those slain by the French; or, again, the monstrous races of men following the Saracen banners. Pious legends, saints' lives, and stories of such apocryphal adventures as those of the Irish St. Brendan (*c.* 486–578) who, as hero of a legend first written down in the 9th-century, *Navigatio Brendani,* and later widely translated and adapted, wanders among strange islands on his way to the earthly paradise—these likewise favour the marvellous. The great 12th-century *Roman d'Alexandre*, a *roman d'antiquite'* based on and developing the early Greek romance of Alexander the Great (the Alexander romance), was begun in the first years of the century by Alberic de Briançon and later continued by other poets. It introduces fantastic elements, more especially technological wonders and the marvels of India: the springs of rejuvenation, the flower-maidens growing in a forest, the cynocephali (dog-headed men), the bathyscaphe that takes Alexander to the bottom of the ocean, and the car in which he is drawn through the air by griffins on his celestial journey.

The setting. The fact that so many medieval romances are set in distant times and remote places is not an essential feature of romance but rather a reflection of its origins. As has been seen, the Old French word ***romanz*** early came to mean "historical work in the vernacular." All the ***romans d'antiquité*** have a historical or pseudohistorical theme, whether they evoke Greece, Troy or the legendary world of Alexander; but while making some attempt to give antiquity an exotic aspect by means of marvels or technological wonders, medieval writers were quite unable to create a convincing historical setting; and thus in all important matters of social life and organization they projected the western European world of the 12th century back into the past. Similarly, historical and contemporary geography were not kept separate. The result is often a confused jumble, as, for example, in the Anglo-Norman Hue de Rotelande's ***Protesilaus,*** in which the characters have Greek names; the action takes place in Burgundy, Crete, Calabria, and Apulia; and Theseus is described as "king of Denmark." This lavish use of exotic personal and geographical names and a certain irresponsibility about settings was still to be found in some of Shakespeare's romantic comedies: the "seacoast of Bohemia" in ***The Winter's Tale*** is thoroughly medieval in its antecedents. For in the medieval period, myth and folktale and straightforward fact were on an equal footing. Not that any marvel or preternatural happening taking place in secular (as opposed to biblical) history was necessarily to be believed: it was simply that the remote times and regions were convenient locations for picturesque and marvellous incidents. It is, indeed, at precisely this point that the transition begins from the concept of romance as "past history in the vernacular" to that of "a wholly fictitious story."

Errors of history and geography in early romance

THE MEDIEVAL VERSE ROMANCES

Arthurian romance. ***The matter of Britain.*** In his ***Historia regum Britanniae,*** Geoffrey of Monmouth "invented history" by drawing on classical authors, the Bible, and Celtic tradition to create the story of a British kingdom, to some extent paralleling that of Israel. He described the rise of the British people to glory in the reigns of Uther Pendragon and Arthur, then the decline and final destruction of the kingdom, with the exile of the British survivors and their last king, Cadwalader. Romances that have Arthur or some of his knights as main characters were classified as ***matikre de Bretagrte*** by Jehan Bodel (flourished 1200) in a well-known poem. There is in this "matter of Britain" a certain amount of material ultimately based on the belief, probably Celtic in origin, in an otherworld into which humans can penetrate, where they can challenge those who inhabit it, or enjoy the love of fairy women. Such themes appear in a highly rationalized form in the lays (*lais*) of the late 12th-century Marie de France, although she mentions Arthur and his queen only in one, the lay of ***Lanval.***

Chre'tien de Troyes. But it was Chrktien de Troyes (flourished 1165–80) who in five romances ***(Erec; Cligès; Lancelot, ou Le Chevalier de la charrette; Yvain ou Le Chevalier au lion;*** and ***Perceval ou li Conte du Graal)*** fashioned a new type of narrative based on the matter of Britain. The internal debate and self-analysis of the ***roman d'antiquite'*** is here used with great artistry. At times, what seems to matter most to the poet is not the plot but the thematic pattern he imposes upon it and the significance he succeeds in conveying, either in individual scenes in which the action is interpreted by the characters in long monologues or through the work as a whole. In addition to this, he attempts what he himself calls a ***conjointure—that*** is, the organization into a coherent whole of a series of episodes. The adventures begin and end at the court of King Arthur; but the marvels that bring together material from a number of sources are not always meant to be believed, especially as they are somehow dovetailed into the normal incidents of life at a feudal court. Whatever Chrétien's intentions may have been, he inaugurated what may be called a Latin tradition of romance—clear, hard, bright, adorned with rhetoric, in which neither the courtly sentiment nor the enchantments are seriously meant. Chrktien had only one faithful follower, the trouvère Raoul de Houdenc (flourished 1200–30), author of ***Méraugis de Portlesguez.*** He shared Chrétien's taste for love casuistry, rhetorical adornment, and fantastic adventure. For both of these authors elements of rhetoric and self-analysis remain important, although the dose of rhetoric varies from one romance to another. Even in Chrktien's ***Perceval ou Li Conte du Graal*** ("Perceval, or the Romance of the Grail")—the work in which the Grail appears for the first time in European literature—the stress is on narrative incident interspersed with predictions of future happenings and retrospective explanations. Arthurian romances of the period 1170–1250 are ***romans d'aventure,*** exploiting the element of the strange, the supernatural, and the magical in the Arthurian tradition. A number (for example, La ***Mule sans frein*** ["The Mule Without a Bridle"], ***c.*** 1200; ***L'Âtre périlleux*** ["The Perilous Churchyard"], ***c.*** 1250) have as their hero Arthur's nephew Gawain, who, in the earlier Arthurian verse romances, is a type of the ideal knight.

Introduction of the Grail theme into European literature

Love as a major theme in romance. The treatment of love varies greatly from one romance to another. It is helpful to distinguish sharply here between two kinds of theme: the one, whether borrowed from classical antiquity (such as the story of Hero and Leander, or that of Pyramus and Thisbe, taken from Ovid's ***Metamorphoses)*** or of much more recent origin, ending tragically; the other ending with marriage, reconciliation, or the reunion of separated lovers. It is noteworthy that "romance," as applied to a love affair in real life, has in modern English the connotation of a happy ending. This is also true of most Old French love romances in verse: the tragic ending is rare and is usually linked with the theme of the lover who, finding his or her partner dead, joins the beloved in death, either by suicide or from grief.

The Tristan story. The greatest tragic love story found as a romance theme is that of Tristan and Iseult. It was given the form in which it has become known to succeeding generations in about 1150–60 by an otherwise unknown Old French poet whose work, although lost, can be reconstructed in its essentials from surviving early versions based upon it. Probably closest in spirit to the original is the fragmentary version of c. 1170–90 by the Norman poet Béroul. From this it can be inferred that the archetypal poem told the story of an all-absorbing passion caused by a magic potion, a passion stronger than death, yet unable to triumph over the feudal order to which the heroes belong. The story ended with Iseult's death in the embrace of her dying lover and with the symbol of two trees growing from the graves of the lovers and intertwining their branches so closely that they could never be separated. Most later versions, including a courtly version by an Anglo-Norman poet known only as Thomas, attempt to resolve the tragic conflict in favour of the sovereignty of passion and to turn the magic potion into a mere symbol. Gottfried von Strassburg's German version, ***Tristan und Isolde (c. 1210),*** based on Thomas, is one of the great courtly romances of the Middle Ages; but although love is set up as the supreme value and as the object of the lovers' worship, the mellifluous and limpid verse translates the story into the idyllic mode. Another tragic and somewhat unreal story is that told in the anonymous ***Chastelaine de Vergi (c. 1250),*** one of the gems of medieval poetry, in which the heroine dies of grief because, under pressure, her lover has revealed their secret and adulterous love to the duke of Burgundy. The latter tells it to his own wife, who allows the heroine to think that her lover has betrayed her. The theme of the dead lover's heart served up by the jealous husband to the lady—tragic, sophisticated, and far-fetched—appears in the anonymous ***Chastelain de Couci (c. 1280),*** and again in ***Daz Herzmaere*** by the late 13th-century German poet Konrad von Wiirzburg. The theme of the outwitting of the jealous husband, common in the fabliaux (short verse tales containing realistic, even coarse detail, and written to amuse), is frequently found in 13th-century romance

Later versions of the Tristan and Iseult theme and in lighter lyric verse. It occurs both in the *Chastelain de Couci* and in the Provencal romance. *Flamenca (c.* 1234), in which it is treated comically.

The theme of separation anti reunion. But the theme that has left the deepest impress on romance is that of a happy resolution, after many trials and manifold dangers, of lovers' difficulties. As has been seen, this theme was derived from late classical Greek romance by way of *Apollonius of Tyre* and its numerous translations and variants. A somewhat similar theme, used for pious edification, is that of the legendary St. Eustace, reputedly a high officer under the Roman emperor Trajan, who lost his position, propel-ty, and family only to regain them after many tribulations, trials, and dangers. The St. Eustace theme appears in *Guillaume d'Angleterre*, a pious tale rather than a romance proper, which some have attributed to Chrétien de Troyes.

The Floire and Blanche-flor romance

A variant on the theme of separation and reunion is found in the romance of *Floire et Blancheflor* (*c.* 1170), in which Floire, son of the Saracen "king" of Spain, is parted by his parents from Blancheflor. daughter of a Christian slave of noble birth, who is sold to foreign slave dealers. He traces her to a tower where maidens destined for the sultan's harem are kept, and the two are reunited when he gains access to her there by hiding in a basket of flowers. This romance was translated into Middle High German, Middle Dutch, Norse, and Middle English (as *Floris and Blancheflur*, *c.* 1250) and in the early 13th century was imitated in *Aucassin et Nicolette*, which is a *chantefable* (a story told in alternating sections of sung verse and recited prose) thought by some critics to share a common source with *Floire et Blancheflor*. In it, the roles and nationality, or religion, of the main characters are reversed; Nicolette, a Saracen slave converted to Christianity, who proves to be daughter of the king of Carthage, disguises herself as a minstrel in order to return to Aucassin, son of Count Gavin of Beaucaire. Jean Renart's *L'Escoufle* (c. 1200–02) uses the theme of lovers who, accidentally separated while fleeing together from the emperor's court, are eventually reunited; and the highly esteemed and influential *Guillaume de Palerne* (*c.* 1200) combines the theme of escaping lovers with that of the "grateful animal" (here a werewolf, which later resumes human shape as a king's son) assisting the lovers in their successful flight. The popular *Partenopeus de Blois* (*c.* 1180), of which ten French manuscripts and many translated versions are known, resembles the Cupid and Psyche story told in the Roman writer Apuleius' *Golden Ass* (2nd century AD), although there is probably no direct connection. In the early 13th-century *Galeran de Bretagne*, Galeran loves Fresne, a foundling brought up in a convent; the correspondence between the two is discovered, and Fresue is sent away but appears in Galeran's land just in time to prevent him from marrying her twin sister, Fleurie.

The theme of a knight who undertakes adventures to prove to his lady that he is worthy of her love is represented by a variety of romances including the *Ipomedon* (1174–90) of Hue de Rotelande and the anonymous mid-13th-century Anglo-Norman *Gui de Warewic*. Finally, there are many examples of the "persecuted heroine" theme; in one variety a person having knowledge of some "corporal sign"—a birthmark or mole—on a lady wagers with her husband that he will seduce her and offer proof that he has done so (this is sometimes called the "Imogen theme" from its use in Shakespeare's *Cymbeline*). The deceit is finally exposed and the lady's honour vindicated. In the early 13th-century *Guillaume de Dôle by* Jean Renart, the birthmark is a rose; and in the *Roman de Violette*, written after 1225 by Gerbert de Montreuil. it is a violet. Philippe de Beaumanoir's *La Manekine* (*c.* 1270), Jean Maillart's *La Contesse d'Anjou* (1361), and Chaucer's *Man of Law's Tale* (after 1387) all treat the theme of the tribulations of a wife falsely accused and banished but, after many adventures, reunited with her husband.

The "Imogen theme"

THE MEDIEVAL PROSE ROMANCES

Arthurian themes. The Arthurian prose romances arose out of the attempt, made first by Robert de Borron in the verse romances *Joseph d'Arimathie, ou le Roman de l'estoire dou Graal* and *Merlin* (c. 1190–1200), to combine the fictional history of the Holy Grail with the chronicle of the reign of King Arthur. Robert gave his story an allegorical meaning, related to the person and work of Christ. A severe condemnation of secular chivalry and courtly love characterize the Grail branch of the prose Lancelot-Grail, or Vulgate, cycle as well as some parts of the post-Vulgate "romance of the Grail" (after 1225); in the one case, Lancelot (here representing fallen human nature) and, in the other, Balain (who strikes the Dolor.. ous Stroke) are contrasted with Galahad, a type of the Redeemer. The conflict between earthly chivalry and the demands of religion is absent from the *Perlesvaus* (after 1230?), in which the hero Perlesvaus (that is, Perceval) has Christological overtones and in which the task of knighthood is to uphold and advance Christianity. A 13th-century prose *Tristan* (*Tristan de Léonois*), fundamentally an adaptation of the Tristan story to an Arthurian setting, complicates the love theme of the original with the theme of a love rivalry between Tristan and the converted Saracen Palamède and represents the action as a conflict between the treacherous villain King Mark and the "good" knight Tristan.

The ideals of chivalry

In the 14th century, when chivalry enjoyed a new vogue as a social ideal and the great orders of secular chivalry were founded, the romance writers, to judge from what is known of the voluminous *Perceforest* (written *c.* 1330 and still unpublished in its entirety), evolved an acceptable compromise between the knight's duty to his king, to his lady, and to God. Chivalry as an exalted ideal of conduct finds its highest expression in the anonymous Middle English *Sir Gawayne and the Grene Knight* (*c.* 1370), whose fantastic beheading scene (presumably taken from a lost French prose romance source) is made to illustrate the fidelity to the pledged word, the trust in God, and the unshakable courage that should characterize the knight.

The special structure of the Lancelot-Grail romance

Structure of the prose romances. The Vulgate *Lancelot-Grail* cycle displays a peculiar technique of interweaving that enables the author (or authors) to bring together a large number of originally independent themes. The story of Lancelot. of Arthur's kingdom, and the coming of Galahad (Lancelot's son) are all interconnected by the device of episodes that diverge, subdivide, join, and separate again, so that the work is a kind of interlocking whole, devoid of unity in the modern sense but forming as impregnable a structure as any revolving around a single centre. One of its most important features is its capacity for absorbing contrasting themes, such as the story of Lancelot's love for Guinevere. Arthur's queen, and the Quest of the Grail; another feature is its ability to grow through continuations or elaborations of earlier themes insufficiently developed. The great proliferation of prose romances at the end of the Middle Ages would have been impossible without this peculiarity of structure. Unlike any work that is wholly true to the Aristotelian principle of indivisibility and isolation (or organic unity), the prose romances satisfy the first condition, but not the second: internal cohesion goes with a tendency to seek connections with other similar compositions and to absorb an increasingly vast number of new themes. Thus the prose *Tristan* brings together the stories of Tristan and Iseult, the rise and fall of Arthur's kingdom, and the Grail Quest. It early gave rise to an offshoot, the romance of *Palamède* (before 1240), which deals with the older generation of Arthur's knights. A similar example of "extension backward" is the *Perceforest*, which associates the beginnings of knighthood in Britain with both Brutus the Trojan (reputedly Aeneas' grandson and the legendary founder of Britain) and Alexander the Great and makes its hero, Perceforest, live long before the Christian era.

LATER DEVELOPMENTS

The Arthurian prose romances were influential in both Italy and Spain; and this favoured the development in these countries of works best described as *romans d'aventure*, with their constantly growing interest in tournaments, enchantments, single combat between knights, love intrigues, and rambling adventures. In Italy, early prose compilations of Old French epic material from the Charlemagne cycle were subsequently assimilated to the other great bodies of medieval French narrative fiction and infused with the spirit of Arthurian prose romance. The great Italian heroic and romantic epics, Matteo Boiardo's *Orlando innamorato* (1483) and Ludovico Ariosto's *Orlando furioso* (1516), are based on this fusion. The serious themes of the Holy Grail and death of Arthur left no mark in Italy. The romantic idealism of Boiardo and Ariosto exploits instead the worldly adventures and the love sentiment of Arthurian prose romance, recounted lightly and with a sophisticated humour.

The Spanish romance

In Spain the significant development is the appearance, as early as the 14th, or even the 13th, century, of a native prose romance, the *Amadís de Gaula.* Arthurian in spirit but not in setting and with a freely invented episodic content, this work, in the form given to it by Garci Rodriguez de Montalvo in its first known edition of 1508, captured the imagination of the polite society of western Europe by its blend of heroic and incredible feats of arms and tender sentiment and by its exaltation of an idealized and refined concept of chivalry. Quickly translated and adapted into French, Italian, Dutch, and English and followed by numerous sequels and imitations in Spanish and Portuguese, it remained influential for more than four centuries, greatly affecting the outlook and sensibility of western society. Cervantes parodied the fashion inspired by *Amadís* in *Don Quixote* (1605); but his admiration for the work itself caused him to introduce many of its features into his own masterpiece, so that the spirit and the character of chivalric romance may be said to have entered into the first great modern novel.

Parallels with sculpture and painting

More important still for the development of the novel form was the use made by romance writers of the technique of multiple thematic structure and "interweaving" earlier mentioned. Like the great examples of Romanesque ornamental art, both sculptural and pictorial, the cyclic romances of the late Middle Ages, while showing a strong sense of cohesion, bear no trace whatever of the classical concept of subordination to a single theme: an excellent proof, if proof were needed, of the limited relevance of this concept in literary aesthetics. Even those romances which, like the *Amadís* and its ancestor, the French prose *Lancelot*, had one great figure as the centre of action, cannot be said to have progressed in any way toward the notion of the unity of theme.

The spread and popularity of romance literature. This is as true of medieval romances as of their descendants, including the French and the English 18th-century novel and the pastoral romance, which, at the time of the Renaissance, revived the classical traditions of pastoral poetry and led to the appearance, in 1504, of the *Arcadiu* by the Italian poet Jacopo Sannazzaro and, in about 1559, of the *Diana* by the Spanish poet and novelist Jorge de Montemayor. Both works were widely influential in translation, and each has claims to be regarded as the first pastoral romance, but in spirit *Diana* is the true inheritor of the romance tradition, giving it, in alliance with the pastoral, a new impetus and direction.

The social milieu

Medieval romance began in the 12th century when clerks, working for aristocratic patrons, often ladies of royal birth such as Eleanor of Aquitaine and her daughters, Marie de Champagne and Matilda, wife of Henry the Lion, duke of Saxony, began to write for a leisured and refined society. Like the courtly lyric, romance was a vehicle of a new aristocratic culture which, based in France, spread to other parts of western Europe. Translations and adaptations of French romances appear early in German: the *Roman d'Enéas*, in a version written by Heinrich von Veldeke before 1186, and the archetypal Tristan romance in Eilhart von Oberge's *Tristant* of *c.* 1170–80. In England many French romances were adapted, sometimes very freely, into English verse and prose from the late 13th to the 15th century; but by far the most important English contribution to the development and popularization of romance was the adaptation of a number of French Arthurian romances completed by Sir Thomas Malory in 1469–70 and published in 1485 by William Caxton under the title of *Le Morte Darthur*. In the Scandinavian countries the connection with the Angevin rulers of England led to importation of French romances in the reign (1217–63) of Haakon of Norway.

The decline of romance. As has been seen, in the later Middle Ages the prose romances were influential in France, Italy, and Spain, as well as in England; and the advent of the printed book made them available to a still wider audience. But although they continued in vogue into the 16th century, with the spread of the ideals of the New Learning, the greater range and depth of vernacular literature, and the rise of the neoclassical critics, the essentially medieval image of the perfect knight was bound to change into that of the scholar-courtier, who, as presented by the Italian Baldassare Castiglione in his *Il Cortegiano* (published 1528), embodies the highest moral ideals of the Renaissance. The new Spanish romances continued to enjoy international popularity until well into the 17th century and in France gave rise to compendious sentimental romances with an adventurous, pastoral, or pseudo-historical colouring popular with Parisian *salon* society until *c.* 1660. But the French intellectual climate, especially after the beginning of the so-called classical period in the 1660s, was unfavourable to the success of romance as a "noble" genre. Before disappearing, however, the romances lent the French form of their name to such *romans* as Antoine Furetière's *Le Roman bourgeois* (1666) and Paul Scarron's *Le Romant comique* (1651–57). These preserved something of the outward form of romance but little of its spirit; and while they transmitted the name to the kind of narrative fiction that succeeded them, they were in no sense intermediaries between its old and its new connotations. The great critical issue dominating the thought of western Europe from about 1660 onward was that of "truth" in literature; and romance, as being "unnatural" and unreasonable, was condemned. Only in England and Germany did it find a home with poets and novelists. Thus, while Robert Boyle, the natural philosopher, in his *Occasional Discourses* (1666) was inveighing against gentlemen whose libraries contained nothing more substantial than "romances," Milton, in *Paradise Lost,* could still invoke "what resounds/In fable or romance of Uther's son . . ."

The search for "truth" in literature

The 18th-century romantic revival. The 18th century in both England and Germany saw a strong reaction against the rationalistic canons of French classicism—a reaction that found its positive conterpart in such romantic material as had survived from medieval times. The Gothic romances, of which Horace Walpole's *Castle of Otrunto* (1764; dated 1765) is the most famous, are perhaps of less importance than the ideas underlying the defense of romance by Richard Hurd in his *Letters on Chivalry and Romance* (1762). To Hurd, romance is not truth but a delightful and necessary holiday from common sense. This definition of romance (to which both Ariosto and Chrétien de Troyes would no doubt have subscribed) inspired on the one hand the romantic epic *Oberon* (1780) and on the other the historical romances of Sir Walter Scott. But influential though Scott's romantic novels may have been in every corner of Europe (including the Latin countries), it was the German and English Romantics who, with a richer theory of the imagination than Hurd's, were able to recapture something of the spirit and the structure of romance—the German Romantics by turning to their own medieval past; the English, by turning to the tradition perpetuated by Edmund Spenser and Shakespeare.

BIBLIOGRAPHY. Among older works the most notable are RICHARD HURD, *Letters on Chivalry and Romance* (1764); GEORGE ELLIS, *Specimens of Early English Metrical Romances,* 3 vol. (1805); and SIR WALTER SCOTT, "Essay on Romance" in the Supplement to the 1815–24 edition of the *Encyclopædia Britannica.* The academic study of romance as a form of imaginative narrative may be said to have begun in 1897 with the publication of W.P. KER, *Epic and Romance* (2nd ed. 1908, reprinted 1957), and of GEORGE SAINTSBURY, *The Flourishing of Romance and the Rise of Allegory.* It was soon continued in innumerable monographs and editions of texts, a full list of which would fill a volume. Most of the

critical works on romance fall into one of two major categories: studies of origins and sources and works on the nature and development of the genre. The former group includes such works as EDMOND FARAL, *Recherches sur les sources latines des contes et romans courtois du moyen âge* (1913); JESSIE L. WESTON, *From Ritual to Romance* (1920, reprinted 1957); ROGER S. LOOMIS, *Arthurian Tradition and Chrétien de Troyes* (1949); and JEAN MARX, *La Légende arthurienne et le Graal* (1952); the study of the genre, in addition to monographs on individual works and authors such as Chrétien de Troyes, Guillaume de Lorris, *Sir Gawain and the Green Knight,* and Sir Thomas Malory, is represented in recent years by FANNI BOGDANOW, *The Romance of the Grail* (1966); EUGENE VINAVER, *The Rise of Romance* (1971); by chapters on romance in ERICH AUERBACH, *Mimesis* (1946; 2nd ed., 1959; Eng. trans., 1953); and ROSEMOND TUVE, *Allegorical Imagery* (1966). J.D. BRUCE, *The Evolution of Arthurian Romance,* 2nd ed., *2* vol. (1928), at one time the standard work in this field, has now been largely superseded by R.S. LOOMIS (ed.), *Arthurian Literature in the Middle Ages: A Collaborative History* (1959). Since 1949 the International Arthurian Society has been publishing an annual *Bibliographical Bulletin* covering the whole range of Arthurian literature in all languages.

(E.Vi./F.Wh.)

Romance Languages

The Roman origins of Romance

The Romance languages, all derived from Latin within historical times, form a subgroup of the Italic branch of the Indo-European language family (see also ITALIC LANGUAGES). The major languages of the family include French, Italian, Spanish, Portuguese, and Romanian; among the Romance languages that now have less political or literary significance or both are the Occitan and Rhaetian dialects, Catalan, Sardinian, and Dalmatian (extinct), among others. Of all the so-called families of languages, the Romance group is perhaps the simplest to identify and the easiest to account for historically. Not only do Romance languages share a good proportion of basic vocabulary—still recognizably the same in spite of some phonological changes—and a number of similar grammatical forms, but they can be traced back, with but few breaks in continuity, to the language of the Roman Empire. So close is the similarity of each of the Romance languages to Latin as currently known from a rich literature and continuous religious and scholarly tradition that virtually no one doubts the relationship. For the layman, the testimony of history is even more convincing than the linguistic evidence; Roman occupation of Italy, the Iberian Peninsula, Gaul, and the Balkans accounts for the "Roman" character of the major Romance languages. Later colonial and commercial contacts with parts of the Americas, of Africa, and of Asia readily explain the French, Spanish, and Portuguese still spoken in those regions.

The name Romance indeed suggests the ultimate connection of these languages with Rome: the English word is derived from a French form of Latin Romanicus, used in the Middle Ages to designate a vernacular type of Latin speech (as distinct from the more learned form used by clerics) as well as literature written in the vernacular. The fact that the Romance languages share features not found in contemporary Latin textbooks suggests, however, that the version of Latin they continue is not identical with that of Classical Latin as known from literature. Nonetheless, although it is sometimes claimed that the other Italic languages (the Indo-European language group to which Latin belonged, spoken in Italy) did contribute features to Romance, it is fairly certain that it is specifically Latin itself, perhaps in a popular form, that is the precursor of the Romance languages.

Distribution of the Romance languages

Nearly 400,000,000 people today claim a Romance language as their mother tongue. To this number may be added the not-inconsiderable number of Romance creole speakers (a creole is a simplified or pidgin form of a language that has become the native language of a community) scattered around the world. French creoles are spoken in the West Indies (with perhaps 5,000,000 speakers), in North America (*e.g.,* Louisiana), and islands of the Indian Ocean (*e.g.,* Mauritius, Réunion, the Seychelles); Portuguese creoles in Africa, India, and Malaysia (perhaps 500,000 speakers); and Spanish creoles in the West Indies (about 200,000 speak Papiamento) and the Philippines. Many speakers use creole for informal purposes and the standard language for formal occasions. Romance languages are also used formally in some countries where one or more non-Romance languages are used by most speakers for everyday purposes. French, *e.g.,* is used alongside Arabic in Tunisia, Morocco, and Algeria; it is the official language of some 20 countries in West and Equatorial Africa and of the Malagasy Republic (Madagascar). Portuguese is the official language of Angola, Mozambique, and Portuguese Guinea; Italian, alongside English, is an official language in Somalia.

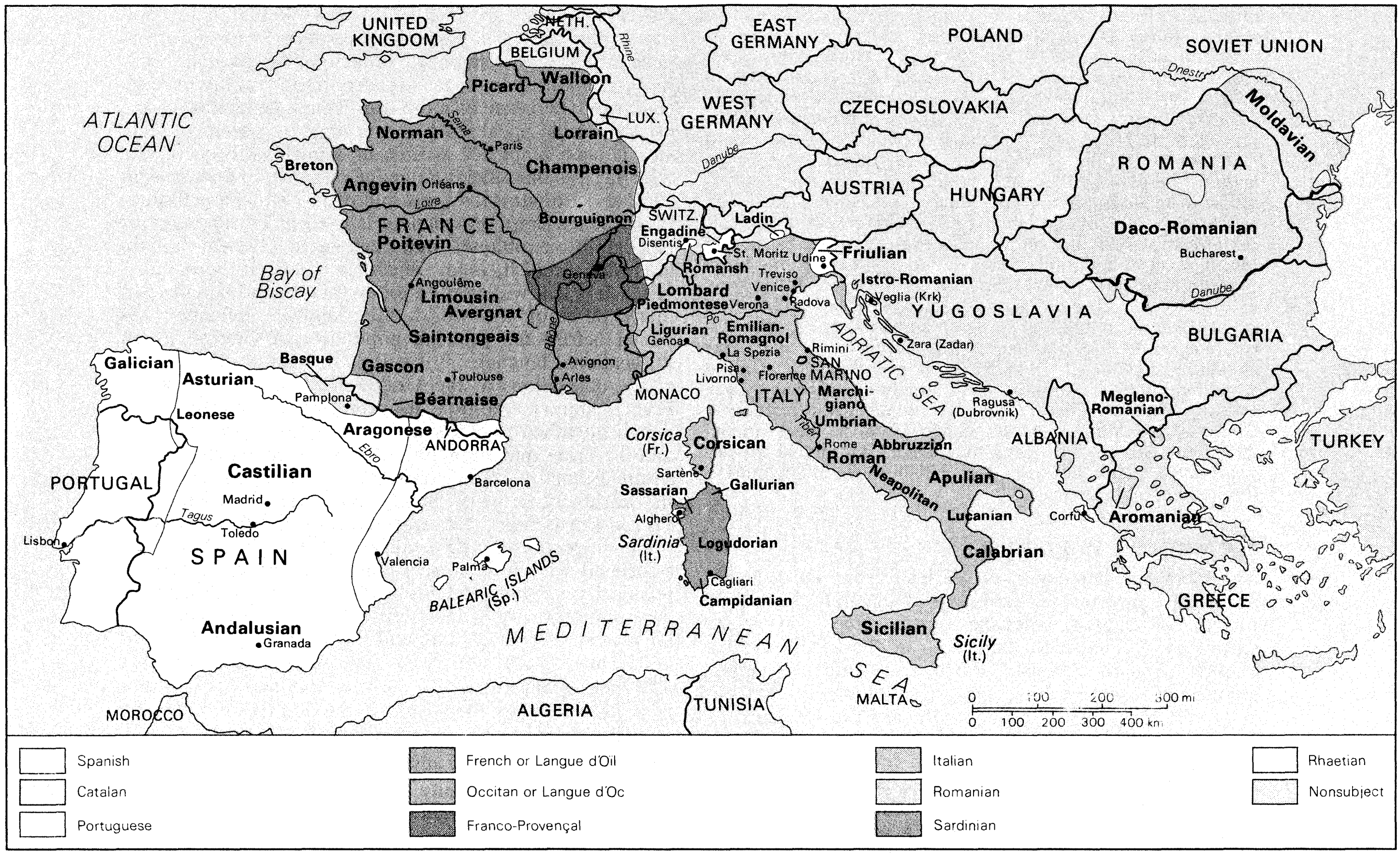
French is still widely used today as a second language in many parts of the world. Although its influence has waned before the growing popularity of English as an international language, it is still used by more than a third of the delegations at the United Nations; the wealth of French literary tradition, its precisely formulated grammar bequeathed by 17th- and 18th-century grammarians, and the pride that Frenchmen feel in their language may ensure French a lasting importance among languages of the world. By virtue of the vast territories in which Spanish and Portuguese hold sway, they will continue to be of prime importance. The beauty of the Italian language, associated with Italy's great cultural heritage, assures its popularity among students, even though territorially it has comparatively little extension. Some lesser Romance languages, such as Catalan and Romanian, retain their vitality, but others, with very few monolingual speakers left, such as Sardinian and the Rhaetian and Occitan dialects, are surely doomed to the extinction that has already overtaken a number of Romance tongues.

This article is divided into the following sections:

I. Languages of the family

Distinguishing languages and dialects

What constitutes a language, as distinct from a dialect, is a vexing question, and opinion varies on just how many Romance languages are spoken today: estimates range between five and 11. The political definition of a language—one that is accepted as standard by a nation or people—is the least ambiguous one; according to this definition, French, Spanish, Portuguese, Italian, and Romanian are certainly languages and possibly also Romansh (a national language of Switzerland since 1938 but probably related to other Rhaetian dialects spoken in



Distribution of Romance languages in Europe.

Italy) and Catalan (the official language of Andorra but also widely used in parts of Spain and France). On linguistic grounds Sardinian (not the language of an independent nation since the 14th century) and Occitan (the medieval Provençal) are usually regarded as languages rather than dialects, though in modern times Occitan has grown so near to French as to be intelligible to French speakers with comparative ease. The Rhaetian dialects of Italy (Ladin in the Dolomites and Friulian around Udine) are usually regarded as non-Italian. Sicilian is different enough from northern and central Italian dialects to be given separate status often, but in Italy all neighbouring dialects are mutually intelligible, with differences becoming more marked with geographical distance. Franco-Provençal (the name given to a group of dialects spoken around the Alpine region of France and Italy) is often also assumed to be a different language from both French and Occitan, though some think it is merely a transitional dialect. Only a few persons know it in France today, though it still survives in the Italian Valle d'Aosta (where French, rather than Italian, remains the language of culture).

Judeo-Spanish

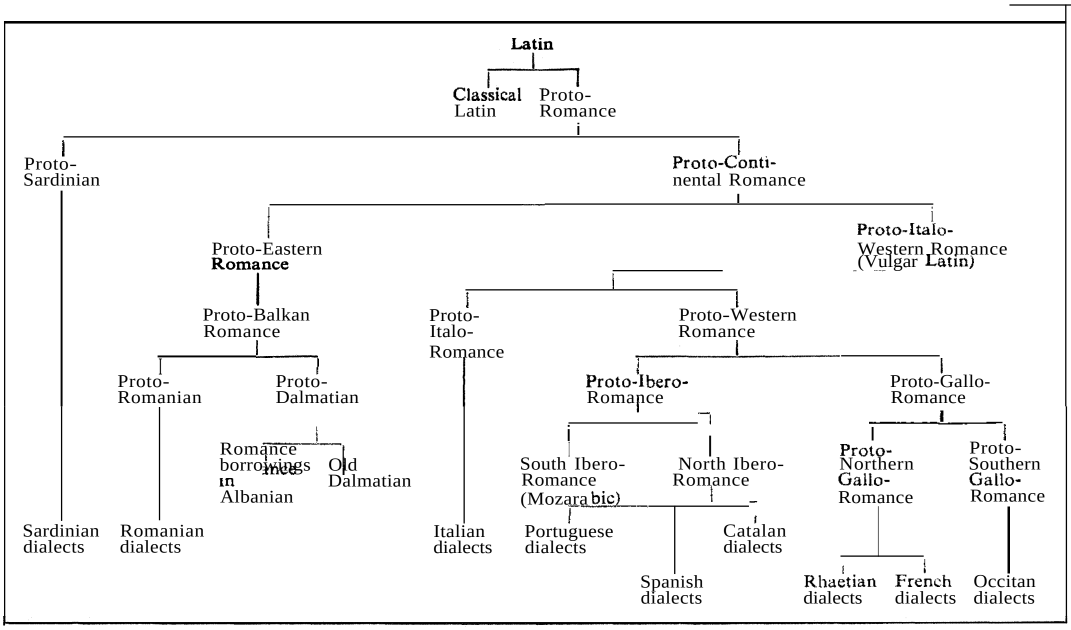
Judeo-Spanish is normally regarded not as an independent language but as an archaic form of Spanish preserving many features of the Castilian of the 15th century, when the Jews were expelled from Spain. There are possibly about 200,000 speakers, mostly originating in the Balkans and Asia Minor but, since World War II, scattered around the world; about 20,000 speakers now reside in Israel, and a significant number live in New York City and Buenos Aires.

Some linguists believe that creoles are often different languages from their metropolitan counterparts; Haitian, for instance, is said to be mutually unintelligible with French. Intelligibility varies so much with the speaker and the hearer, however, that it is difficult to formulate firm criteria on this basis.

The Dalmatian dialect

Many Romance dialects have virtually ceased to be spoken in the last century. Of these, Dalmatian is the most striking, its last known speaker, one Antonio Udina, having been blown up by a land mine in 1898. He was the main source of knowledge for his parents' dialect (that of the island of Veglia, or Krk), though he was hardly an ideal informant; Vegliot Dalmatian was not his native language, and he had learned it only from listening to his parents' private conversations. Moreover, he had not spoken the language for 20 years at the time he acted as an informant, and he was deaf and toothless as well. Most of the other evidence for Dalmatian derives from documents from Zara (modern Zadar) and Ragusa (modern Dubrovnik) dating from the 13th to 16th centuries. It is possible that, apart from isolated pockets, the language was then replaced by Croatian and, to a lesser extent, by Venetian (a dialect of Italian). It is certain, even from scanty evidence, that Dalmatian was a language in its own right, noticeably different from other Romance languages and presenting difficulties of classification.

On the Istrian Peninsula of the Yugoslav mainland close to the island of Veglia, another Romance language precariously survives (5,000 speakers); known as Istriot, it may be related to Vegliot, though some scholars dispute this and connect it with Rhaetian Friulan dialects or with Venetian dialects of Italian; others maintain that it is an independent language. There are no texts except those collected by linguists. A little further north in the same peninsula, another Romance dialect, Istrio-Romanian, is threatened with extinction (2,000 speakers). Usually classified as a Romanian dialect, it may have been carried to the Istrian Peninsula by Romanians taking refuge from the Turks in the 16th and 17th centuries and has undergone strong Croatian influence. There is evidence for its existence from a short list of words in a 1698 historical wofk, but it is otherwise unwritten. Another isolated Romanian dialect that may be nearing extinction is Megleno-Romanian, from a mountainous region of Macedonia, just west of the Vardar River, on the border between Yugoslavia and Greece. In 1914 there were 13,000 speakers, but many have emigrated to Asia Minor, Yugoslavia, and Romania, where small pockets survive. The only texts are those transcribed from oral traditions.



From R.A. Hall Jr., *Introductory Linguistics*; originally published by Chilton Books, now distributed by Rand McNally & Co.

Other Romance tongues earlier ceased to be spoken; there is evidence, for instance, of a form of Spanish spoken in Arab-occupied Spain until shortly after its liberation by the Spanish, accomplished at the end of the 15th century. Usually known as Mozarabic, from the Arabic word for an "Arabized person," or as *ʿajamī* "barbarian language," it was originally the spoken language of the urban bourgeoisie, who remained Christian while the peasantry generally converted to Islām, but it appears that many Arabs also came to use it, even though Arabic remained the only written language. Because most of the evidence, apart from a 15th-century glossary from Granada, is written in Arabic script (which uses no vowel signs), it is difficult to reconstruct the phonology of the language, but it appears to be a very conservative Spanish dialect. Much of modern information about Mozarabic comes from medical and botanical works that give Mozarabic terms alongside the Arabic. To this was added recently the discovery of Romance refrains (*kharjahs*) inserted in Arabic love ballads (*muwashshaḥs*) of the 11th and 12th centuries; study of these began only in 1946. For much of the Muslim period (beginning in 711), Christians were treated tolerantly and became culturally Arabized. Even after persecution by fanatical Muslim newcomers in the 12th century, the Mozarabs were often in conflict with Westernized "liberators" from the north. Their language died out soon after the Arabs were driven out of Spain at the end of the 15th century, though it is sometimes claimed that Mozarabic has left its mark on the dialects of southern Spain and Portugal.

Mozarabic

Other Romance languages may have developed in peripheral regions of the Roman Empire only to die out under pressure from neighbouring non-Italic languages. Often these extinct Romance dialects are known from words borrowed into surviving languages; Berber, for instance, bears witness to the long and brilliant Roman period in North Africa that was to end in the 7th century AD with Arab invasions, and British Celtic (especially Welsh) retains many traces of what appears to have been a conservative Romance dialect, completely eliminated by Anglo-Saxon in about the 5th century. Albanian has so many Romance words that some style it "semi-Romance," and farther north, in what was formerly the Roman province of Pannonia (modern northwestern Yugoslavia and western Hungary), Romance speech was probably not dead at the time of the Magyar invasion at the end of the 10th century. Thus, there is reason to believe that Romance dialects may have been spoken at one time over much of southeastern Europe. It is also evident that Romance languages have been retreating south before German for some time, and it is probable that Romance tongues were used in the whole of Switzerland and parts of Bavaria and Austria until about the 9th century. Some scholars maintain that the modern Rhaetian dialects of Switzerland and northern Italy are remnants of an earlier Germano-Romance speech form.

CLASSIFICATION METHODS AND PROBLEMS

Though it is quite clear which languages can be classified as Romance, on the basis primarily of lexical (vocabulary) and morphological (structural) similarities, the subgrouping of the languages within the family is less straightforward. Most classifications are, overtly or covertly, historico-geographical—so that Spanish, Portuguese, and Catalan are Ibero-Romance, French, Occitan, and Franco-Provençal are Gallo-Romance, and so on. Shared features in each subgroup that are not seen in other such groups are assumed to be ultimately traceable to languages spoken before Romanization. The first subdivision of the Romance area is usually into West and East Romance, with a dividing line drawn across Italy between La Spezia and Rimini. On the basis of a few heterogeneous phonetic features, one theory maintains that separation into dialects began early, with the Eastern dialect areas (including central and south Italy) developing popular features and the school-influenced Western speech areas maintaining more literary standards. Beyond this, the substrata (indigenous languages eventually displaced by Latin) and superstrata (languages later superimposed on Latin by conquerors) are held to have occasioned further subdivisions. Within such a schema there remain problem cases. (1) Is Catalan, for instance, Ibero-Romance or Gallo-Romance, given that its medieval literary language was close to Provençal? (2) Do the Rhaetian dialects group together, even though the dialects found in Italy are closer to Italian and the Swiss ones closer to French? Sardinian is generally regarded as linguistically separate, its isolation from the rest of the Roman Empire by incorporation into the Vandal kingdom in about 455 providing historical support for the thesis. The exact position of Dalmatian in any classification is open to dispute.

Systems for classifying the Romance languages

A family-tree classification is a commonly used method of classifying the Romance languages. If, however, historical treatment of one phonetic feature is taken as a classificatory criterion for construction of a tree, results differ. Classified according to the historical development of stressed vowels, French would be grouped with North Italian and Dalmatian but not with Occitan, while Central Italian would be isolated. Classifications that are not based on family trees usually involve ranking languages according to degree of differentiation rather than grouping them; thus, if the Romance languages are compared with Latin, it is seen that by most measures Sardinian and Italian are least differentiated and French most (though in vocabulary Romanian has changed most). By most nonhistorical measures, standard Italian is a "central" language (*i.e.*, it is quite close and often readily intelligible to all other Romance languages), whereas French and Romanian are peripheral (they lack similarity to other Romance languages) and require more effort if other Romance speakers are to understand them. Spanish and Portuguese are even today so close in most respects that they can be regarded from a linguistic point of view as dialects of the same language, even though structural criteria would assign them to different broad classes.

Extent of mutual intelligibility

In general, it is possible to maintain that all of the standard Romance languages are to some extent mutually intelligible (especially in their Latin-based written forms) and that they have become more so during the course of history, because of much borrowing from one another and remodelling on Latin, the religious language of most speakers. In 19th-century Romania, for instance, national pride prompted a turning toward other Romance cultures, especially the French, in order more clearly to differentiate Romania from neighbouring Slavic countries, with the consequence that Romanian has become "more Romance" in vocabulary if nowhere else. Local dialects in all the countries are less affected by such converging movements, but even here encroachment on the local dialect by the standard form of the language leads to the ironing out of dialect peculiarities, often ending with the loss of the local dialect, replaced by a regional variety of the standard dialect; this process is particularly evident in southern France, much less so in Italy.

MINOR LANGUAGES

Occitan. Occitan is the modern name given by linguists to the group of dialects spoken by some 12,000,000–14,000,000 people in the south of France (or about one-fourth of the whole French population). All Occitan speakers now use French as their official and cultural language, but their local dialects remain lively and, across most of the area, remarkably homogeneous. The name Occitan derives from the name of the area Occitanie (formed on the model of Aquitania). The medieval language is often called langue d'oc (from the word for "yes," compared with langue d'oïl, Northern French, and with the *si* languages, Spanish and Italian). In the area itself, the names Lemosi (Limousin) and Proensal (Provençal) were formerly used, but today these are often considered too localized to designate the whole range of dialects. Members of a vigorous literary movement in the Provence region, however, still prefer to call their language Provençal.

Occitan was rich in poetic literature in the Middle Ages until the north crushed political power in the south (1208–29). The standard language was, however, well established and did not really succumb before French until the 16th century, while only after the French Revolution did the French language penetrate into popular use in place of Occitan. In the mid-19th century, a literary Renaissance led by the Félibres (from an old word meaning "wisemen"), based on the dialect of the Arles–Avignon region, lent new lustre to Occitan, and a modern standard dialect was established. The most famous figure of this movement was a Nobel Prize-winning poet, Frédéric Mistral. Almost contemporaneously, a similar movement, based in Toulouse, arose and concentrated on problems of linguistic and orthographic standardization to provide a wider base for literary endeavour.

The Occitan dialects have changed comparatively little since the Middle Ages, though now French is influencing them more and more. Perhaps this influence has helped them to remain more or less mutually intelligible. The main dialect areas are Limousin, in the northwest comer of the Occitan area; Auvergnat, in the north central region of this area; northeastern Alpine-Provençal; and Languedocian and Provençal, on the west and east of the Mediterranean seacoast, respectively.

Gascon and Franco-Provençal

Gascon, in the southwest of France, is usually classified as an Occitan dialect, though to most other southerners it is today less readily comprehensible than Catalan. Some scholars claim that it has always been distinct from Occitan, because of the influence of a non-Celtic Aquitanian pre-Roman population. The Roman name of the region, Vasconia (from which the name Gascony derives), suggests the relationship of its original population with the non-Indo-European Basques. Although poets from this region used the Occitan literary language during the Middle Ages, there is evidence that their spoken language was noticeably different (the 14th century *Leys **d'amor*** calls it *lengatge estranh* "strange language"). Some of the region remained politically independent for a long period (the Kingdom of Béarn, which used its own standardized dialect as an official language, was not incorporated into France until 1620), and popular use of French is evident only from about 1700. Documents in the dialect are few, however; they date from about the 12th century.

Northeast of the Occitan region, along the French, Swiss, and Italian frontiers, is located a group of dialects that historically have shared most vowel developments with languages to the south and many consonant changes with those to the north. For the last 100 years claims have been made for the linguistic autonomy of these dialects, usually called Franco-Provençal; today it is estimated that somewhat fewer than 2,000,000 speakers use them (urban speakers are hard to find, and even in the countryside young speakers are few). Dialects are extremely diversified and heavily influenced by French, which has been used extensively in the area since the 13th century. Even during the Middle Ages there was no standard form of Franco-Provençal, though some 12th–13th-century documents exist. The dialect of Geneva (now extinct except in some rural communes) was the official language of the Swiss republic for some time, but otherwise none of the dialects has had official status. Some claim that a section of a manuscript, the so-called Alexander fragment, dating from the 1lth–12th century and apparently part of a lost poem, is Franco-Provençal in character, but others maintain that it, like other literary texts from the region, is mainly Provençal with some French features. Since the 16th century, there has been local dialect literature, notably in Savoy, Fribourg, and Geneva.

Catalan. Currently spoken by about 4,800,000 people in Spain and 200,000 in France (in Roussillon), as well as by 10,000 in Andorra and 12,000 in Alghero (Sardinia), Catalan has lost little of its former lustre, even though it is no longer an important national language (as it was between 1137 and 1749, as the official language of Aragon). Although in the Middle Ages there is no evidence of dialectalization, perhaps because of the standardizing influence of its official use in the Kingdom of Aragon, since the 16th century the dialects of Valencia and the Balearic Isles, especially, have tended to differentiate from the Central (Barcelona) dialect. Nevertheless, some degree of uniformity is preserved in the literary language, which continues to flourish in spite of the little encouragement received from Madrid since the Spanish Civil War. Although there were no publications of any sort in Catalan between 1939 and 1941, and only 12 titles were published in 1946, in 1968 the number of titles published had reached 520.

Historic use of Catalan

The earliest surviving written materials in Catalan date from the 12th century (a charter and six sermons), with poetry flourishing from the 13th century, before which time Catalan poets wrote in Provençal. The first true Catalan poet was Ramon Llull (1235–1315), and the language remained vigorous (its greatest poet was Ausiàs March, 1397–1459, a Valencian) until the union of the Aragonese and Castilian crowns in 1474 marked the beginning of its decline. After that, although mainly grammatical works appeared, the language was to wait for its renaissance until the late 19th century. In 1906 the first Catalan Language Congress attracted 3,000 participants, and in 1907 the Institut d'Estudis Catalans was founded. Yet not until 1944 was there a course in Catalan philology at the University of Barcelona; a chair of Catalan language and literature was not founded there until 1961.

It is much disputed whether Catalan is more closely related to Occitan or to the Hispanic languages. Medieval Catalan was so close to Lemosi, the literary dialect of Occitan in southern France, that it is thought by some to have been imported from beyond the Pyrenees in the resettlement of refugees from the Moors. In more modern times, Catalan has, however, grown closer to Aragonese and Castilian, so that its family-tree classification becomes less relevant. It mas occasionally called Llemosi by 19th-century Catalan revivalists, however, who wished to emphasize its independence from other Iberian tongues by stressing its relation to Occitan. Certainly, by most standards, Catalan merits the distinction of being deemed a language in its own right, and it shows little sign of decline.

Historical extent of Sardinian

Sardinian. Sardinian is currently spoken by more than 1,000,000 people, but it has many dialect differences, and there is virtually no literature, nor even a newspaper in the language (although satirical journals do appear from time to time). In earlier times the language was probably spoken in Corsica, where a Genoese dialect of Italian is now used (although French has been Corsica's official language for two centuries). Since the early 18th century Sardinia's destiny has been linked with that of the Italian mainland, and Italian is now the official language. From the 14th century till the 17th century, Catalan (at that time the official language of Aragon, which ruled Sardinia) was used extensively, especially for official purposes; a Catalan dialect is still spoken in Alghero. Castilian began to be used in official documents in 1600 but did not supplant Catalan in the south of the island until later in the 17th century. Sardinia was more or less independent from 1016, when Arab occupation was ended, until the arrival of the Aragonese in 1322, though much influenced by the Genoese and Pisans. The first documents in Sardinian are legal contracts dating from about 1080; in the north of the island Sardinian was used for such documents until the 17th century. The main dialect groupings are Logudorian (Logudorese), the central, most conservative dialect, which appears to have been used throughout the island in earlier times and which (in a northern form) provides the basis for a *sardo illustre* (a conventionalized literary language used mainly for folk-based verse); Campidanian (Campidanese), centred around Cagliari in the south, heavily influenced by Catalan and Italian; Sassarian (Sassarese) in the northwest; and Gallurian (Gallurese) in the northeast. It is sometimes said that these last two are not Sardinian dialects but rather Corsican. Gallurian in particular is related to the dialect of Sartène in Corsica, and it may have been imported into the Gallura region in the 17th and 18th centuries by refugees from vendetta feuds.

Sardinian is unintelligible to most Italians and, with its harsh consonants and hammered accent, gives an acoustic impression more similar to Spanish than Italian. It is clearly and energetically articulated but has always been regarded as barbarous by the soft-speaking Italians; Dante, for instance, said that Sardinians were like monkeys imitating men. It retains its vitality as a "home language," but dialect diversification is such that it has little chance of development to greater prominence. Perhaps the development of the island as a tourist centre, with better communications with the mainland, will lead to the eventual decline of the language.

Rhaetian. The Rhaetian, or Rhaeto-Romanic, dialects derive their conventional name from the ancient Raetics of the Adige area, who, according to classical authors, spoke an Etruscan dialect. In fact, there is nothing to connect Raetic with Rhaetian except geographical location, and some scholars would deny that the different Rhaetian dialects have much in common, though others claim that they are remnants of a once-widespread Germano-Romance tongue. Three isolated regions still use Rhaetian.

Romansh and other Swiss dialects of Rhaetian

In Switzerland, where in 1960 some 49,800 speakers (about 1 percent of the population) spoke one or another Rhaetian dialect, Romansh (Rumantsch), the standard dialect of Grisons canton, has been a "national" language, used for cantonal but not federal purposes, since 1938. The proportion of Rhaetian speakers in Grisons fell from 39.8 percent in 1880 to 26 percent in 1960, with a corresponding increase in the Italian-speaking population, but interest in Romansh remains keen, and there are five Romansh newspapers. The main Romansh dialects are usually known as Sursilvan (or Surselvan) and Sutsilvan (or Subsilvan), spoken on the western and eastern banks of the Rhine, respectively. Another important Swiss Rhaetian dialect, Engadine, is spoken in the Protestant Inn Valley, east of which there is now a German-speaking area that has encroached on former Romance territory since the 16th century. The dialects from the extreme east and west of the Swiss Rhaetian area are mutually intelligible only with difficulty, though each dialect is intelligible to its neighbour. Sursilvan (spoken around the town of Disentis) has one text dating from the beginning of the 12th century but then nothing else until the work of Gian Travers (1483–1563), a Protestant writer. The Upper Engadine dialect (spoken around Samedan and Sankt Moritz) is attested from the 16th century, notably with the Swiss Lutheran Jacob Bifrun's translation of the New Testament. Both dialects have had a flourishing local literature since the 19th century. In many ways the Swiss Rhaetian dialects resemble French, and speakers seem to feel more at home with French than with Italian.

Ladin and Friulian

In the Alto-Adige and Dolomites area of Italy, between 12,000 and 20,000 persons speak a language they call Ladin. Some Italian scholars have claimed that it is really an Italian (Veneto-Lombard) dialect. German is the other main language spoken in this now semi-autonomous region, much of which was Austrian until 1919. Though it is sometimes said that Ladin is threatened with extinction, it appears to retain its vitality among the mountain peasantry, as distinct from German-speaking hotel owners. Ladin newspapers are on sale in village shops, and speakers welcome visitors (there mainly to ski or climb) who show interest in their language, which is comprehensible without too much difficulty to a student of Romance languages. As it appears that these remote valleys were very sparsely populated until recently, it may even be that the number of speakers there has in fact grown. Since World War II, Ladin has been taught in primary schools in the Gardena and Badia valleys, in different conventionalized dialect forms. Although a Ladin document of the 14th century (from Val Venosta, to the west of the modern Ladin region) is known from references, the earliest written materials in Ladin currently possessed dates from the 18th century, a word list of the Badia dialect. In more recent times there have been a few literary and religious texts.

In Italy, north of Venice, stretching to the Yugoslav border on the east and to the Austrian border on the north, its western boundary almost reaching the River Piave, is the Friulian (Friulan, Frioulian) dialect area, centred around Udine, with 500,000 speakers. This dialect is much closer to Italian than Ladin or Romansh, and it is often claimed to be a Venetian dialect. Venetian proper has gained ground at the expense of Friulian both to the east and west since the 1800s. Friulian retains its vitality today in the well-populated, industrialized region, however, and supports a vigorous local literature; its most notable poet was Pieri Zorut (1792–1867). The first text in Friulian (apart from a doubtful 12th-century inscription) is a short one dating from about 1300, followed by numerous documents in prose, as well as some poems, up to the end of the 16th century, when a rich poetic tradition began.

Creoles. The French, Spanish, and Portuguese creoles, together with their metropolitan equivalents, share many things in common. Indeed, some scholars regard them as in some sense related, either in sharing an African grammatical base, with a superimposed Romance lexicon, or in historical derivation from a Portuguese pidgin lingua franca used by colonizers and slavers. with later addition of vocabulary from metropolitan languages, such as French and Spanish, with which they came into contact. Other scholars maintain that the creoles are continuators of French, Spanish, and Portuguese in the same way as these are themselves continuators of Latin but that, under the conditions that attended the slave trade, linguistic change was exceptionally rapid, so that the origins of the creoles are often hardly recognizable.

Origin of the term creole

"Creole" is a word first found in Spanish (*criollo;* 1590), meaning a Spaniard born in the colonies or his black household servant. It most probably originated in Portuguese, in which it is related to such words as *criança* "child" and criada "maid-servant," generally indicating a household dependent, although the word *crioulo* is not known until 1632. Today, "creole" has come to indicate a pidgin or trade language that has become the mother tongue of a population, often black; the conditions under which this has happened have included forcible transplantation and intermingling of people with mutually unintelligible native languages and imposition of the master's language, during the slave-trade era.

Of Romance creoles used today, French creoles are most widespread. In Haiti, for instance, there are 3,000,000–4,000,000 creole speakers, of whom only about 10 percent know French; the island of Santo Domingo (Hispaniola), of which the eastern half forms the Haitian Republic, was settled in 1665. The Lesser Antilles (Martinique, Guadeloupe, Dominica, St. Lucia, St. Kitts, etc.) were colonized in 1635, and many still use French creoles, even when change of ownership led to the imposition of English as the official language. French creoles are also used in Cayenne (French Guiana) and, though dying out, in Louisiana. In all, probably 5,000,000 speakers use French creoles in the Americas. On islands of the Indian Ocean, too, French creoles are spoken; in Mauritius (population 160,000), owned by France from 1715 to 1810, the creole retains its hold as a lingua franca even though English is the official language and though the majority of the population use Indian dialects as home languages. In the Seychelles (population 40,000), owned by France from 1768 until 1814, when they became British, and in Réunion (originally L'Île Bourbon; population 300,000), where French is still the official language, French creoles are still in use. Some French-creole speakers claim that creoles from other far-off regions are easily intelligible to them. Others contest this, however, pointing out that the creole used by educated speakers is often heavily larded with standard French on all but very informal occasions. Certainly, the linguist can easily discern similarities, especially in grammatical structure, that make the various French creoles seem more like each other than like standard French.

Portuguese creoles were purportedly once widely used in Asia, though probably more frequently as trade languages than as mother tongues. They survive today in Macau and Hong Kong and to some extent in Malaysia and Goa. In Africa a Portuguese creole is used by about 200,000 people in Portuguese Guinea, Senegal, the Cape Verde Islands, and some Gulf of Guinea islands (Annobón and Principe, where it is losing ground to Spanish, and São Tomé, a previously uninhabited island settled in 1485, where, of the 64,000 inhabitants, just over 28,000 speak creoles). In South America a Brazilian creole is still used in the interior, although at one time this language was more widespread in the country (spreading even to Surinam, where Portuguese Jews and their slaves fled from Brazil in the 17th century).

Papiamento

Papiamento, spoken by 200,000 people in the Dutch Caribbean islands of Aruba, Curaçao, and Bonaire, is today classed as a Spanish creole, though some claim that it was once a Portuguese creole that later acquired many new words from Spanish. A Spanish creole also survives precariously in the Philippines among descendants of mixed Spanish–Filipino stock.

On the whole, creoles are rarely used for literature, except satirical and comic pieces. Most speakers regard them as "bad" versions of the standard language and in formal situations try to improve their usage on the model of the standard, though they admit to feeling more relaxed speaking natural creole. Sometimes "purer" creole speech can be heard among speakers not as much exposed to the standard form of the language, as in West Indian islands where English is the official language, while a French creole is the home language.

MAJOR LANGUAGES

French. Probably the most internationally important of the Romance languages, French is used as the official language in 21 countries and as a co-official language in several more (including Algeria, Belgium, Canada, Luxembourg, Switzerland, Jersey). In France and Corsica about 47,000,000 use it as their first language; in Canada, 5,000,000; in Belgium, 4,000,000; in Switzerland (cantons of Neuchâtel, Vaud, Genève, Valais, Fribourg), more than 1,000,000; in Monaco, 20,500; in the Italian Valle d'Aosta, 100,000; and, in the United States (especially Maine and Louisiana), nearly 1,500,000. Moreover, about 5,000,000 Africans and 4,000,000 Indo-Chinese use it as their principal international language; many creole French speakers, too, use standard French in formal situations.

Standard French is based on the dialect of Paris (in the so-called Île de France with its Francien dialect), which assumed importance in about the second half of the 12th century; it was basically a north central dialect with some northern features. Before that, other dialects, especially Norman (which developed in Britain as Anglo-Norman, widely used until about the 14th century), and northern dialects, such as Picard, had more prestige, especially in the literary sphere. The Edict of Villers-Cotterêts (1539), however, established Francien as the only official language, as against both Latin and other dialects. From then on, standard French began to smother local dialects, which were officially discouraged until recent times, though the standard language did not spread to popular usage in all regions until well into the 19th century. Dialectal features, still admired and cherished by 16th-century writers, were ridiculed in the 17th and 18th centuries, when the grammar and vocabulary of the modern language were standardized and polished to an unprecedented degree.

Linguistic change in French

Linguistic change was more rapid and more drastic in northern France than it was in other European Romance regions, and influence from Latin was comparatively slight (though borrowing of Latin vocabulary has been great since the 14th century). The influence of the Germanic Frankish invaders is often held to account for exotic features in Old French, such as strong stress accent and abundant use of diphthongs and nasal vowels; but the change in about the 15th century to a more sober (even monotonous) intonation and loss of a stress accent on each word can hardly be attributed to influences from neighbouring languages. The popularity of French as a first foreign language, in spite of numerous pronunciation difficulties for nearly all foreigners, is perhaps as much the result of the precise codification of its grammar, effectuated especially in the 18th century, as of the brilliance of its literature at all periods. Thus. though Italian would be an easier language for them to master, most foreign speakers of Romance languages become acquainted with French at school and often retain an affection and admiration for the language.

The first document apparently written in French purports to date from 842; known as the Strasbourg Oaths, it is a Romance version of an oath sworn by two of Charlemagne's grandsons. Some claim that the text is thinly disguised Latin constructed after the event to look authentic, for political propaganda purposes; others suppose that its Latinizing tendencies reveal the struggle of the scribe with the problems of spelling French as it was spoken at the time. If the language of the Strasbourg Oaths is Northern French, it is difficult to decide what dialect it represents—some say that of Picard, others Franco-Provençal, and so on. The second existing text in Old French is a short sequence of the writings of St. Eulalia, precisely dated (AD 880–882) and localized (Valenciennes); it is definitely Picard in character. Two 10th-century texts (the Passion du Christ and the ***Vie*** de *St.* Ltger) seem to mingle Northern and Southern dialect features, while another (the "Jonas fragment") is obviously from the far north. After that the Norman dialect seems to dominate literature, though here it is probable that the language is better described as a standard language with elimination of gross dialect features. Two manuscript traditions---one from the far north and one from the west—seem to have developed, and it is possible that the intersection of the two produced the Francien dialect that was eventually to reign supreme.

Modern dialects of French

Modern dialects are classified mainly on a geographical basis, and most survive only in the peasant speech. Walloon, a dialect spoken mainly in Belgium, is something of an exception in that it has had a flourishing dialect literature since about 1600. Other dialects are grouped as follows:

Central: Francien, Orléanais, Bourbonnais, Champenois
Northern: Picard, Northern Norman
Eastern: Lorrain, Bourguignon (Burgundian), Franc-Comtois
Western: Norman, Gallo (around the Celtic Breton area), Angevin, Maine
Southwestern: Poitevin, Saintongeais, Angoumois.

Outside France, apart from the creoles, the French of Canada, originally probably of Northwestern dialect type, has developed the most individual features. Although in the 18th century Canadian French was regarded as exceptionally "pure" by metropolitan commentators, it began to diverge from Parisian French as English influences took over from the French after 1760. It is less clearly articulated, with less lip movement and with a more monotonous intonation than standard French; some change in consonantal sounds occurs (t, d shift to *ts, dz*, respectively, and *k* or *g* followed by *i* or *e* become palatalized [pronounced with the tongue touching the hard palate, or roof of the mouth]); nasal vowels tend to lose the nasal element; vocabulary and syntax are heavily anglicized. Though intellectuals turn toward France for cultural inspiration (some university-educated French Canadians may not even know English), the pronunciation and usage of standard French is sometimes derided by French Canadians; this may be because their English compatriots are taught Parisian at school. The French-speaking population of Canada is growing at a relatively fast rate, and at present 82 percent of the population of Quebec Province use French as their normal language. Even today, however, French is not as socially prestigious as English; the activities of the separatist movement are evidence of the feeling of grievance that many French Canadians still have.

Spanish. Spanish, the Romance language spoken as a first language by the most people in the world, is the official language of 18 American countries as well as that of Spain, and, though many South and Central Americans use native Indian languages as their first language, Spanish is spreading and achieving continuing educational progress. Estimated numbers of speakers are as follows (in order of numerical importance): Mexico, 35,000,000; Spain, 23,000,000; Argentina, 16,000,000; Colombia, 13,000,000; Dominican Republic and El Salvador, 3,000,000 each; the United States, Puerto Rico, and Uruguay, 2,500,000 each; Honduras and Bolivia, 2,000,000 each; Nicaragua and Guatemala, 1,500,000 each; Panama, Paraguay, Costa Rica, 1,000,000 each; the Philippines, 500,000. There are also a few hundred thousand Judeo-Spanish speakers and about 100,000 Spanish speakers in Africa.

The dialect spoken by nearly all these speakers is basically Castilian, and indeed Castellano is still the name used for the language in several American countries. In the norrh of Spain two other Spanish dialect groups (Asturo-Leonese and Aragonese) survive but seem doomed to extinction. The dialect of the Northwest (Galician, or Gallego) is properly a Portuguese dialect, and, on the east,,Catalan can be held to be a different language, as noted above. The now-unchallenged ascendancy of Castilian among Spanish dialects is the result of the particular circumstances of the Reconquista (the conquest of Moorish Spain by the Spanish, completed in 1492), with which the language went south. Having established itself in Spain, the Castilian dialect, possibly in its southern, or Andalusian, form, was then exported to the New World during the 16th century.

Standard Castilian is no longer the language of Old Castile, which was already regarded as rustic and archaic in the 15th century, but a modified form developed in Toledo in the 16th and 17th centuries and, more recently, in Madrid. American countries have developed their own standards, differing mainly in phonology (in which they often agree with the southern Spanish dialects) and in vocabulary (in which loanwords from English are more frequent), but differentiation is comparatively slight, and some Americans still regard true Castilian as their model, On the whole, American forms of Spanish are more musical and suave than the harsh Castilian of Madrid, but it is remarkable how little deformation, or creolization, of the language has occurred, even in the mouths of uneducated Indian speakers.

The standard Spanish dialect, Castilian

The first texts in Spanish consist of scattered words glossing two Latin texts of the 10th century, one from Rioja and the other from Castile; the language in the two documents shows few dialect differences. Another document, dating from about 980, seems to be Leonese in character. The Mozarabic verse forms known as *kharjahs* are the next-oldest surviving texts, but by the middle of the 12th century the famous epic poem El *cantar de mio Cid* ("The Song of the Cid") appeared in a language that is basically Castilian. Literary works in Leonese appear till the 14th century and in a conventionalized Aragonese till the 15th century, but Castilian was destined from the first to gain the upper hand, even making an impact on Portuguese, especially in the 15th and early 16th centuries.

Judeo-Spanish (Jewish-Spanish, Sefardi, Ladino) is the continuation of an archaic form of Castilian, reflecting the state of the language before 16th-century standardization. The expulsion of the Jews from the Iberian Peninsula in 1492 affected mainly the humbler classes, with the rich preferring "conversion," but the latter often later chose voluntary exile to settle in England and Holland, where their Sefardic tongue precariously survives as a religious language in a few communities. Earlier refugees fled to the Middle East and, once settled, continued to produce learned works in a literary archaic form of their language, Ladino, written in an adapted Hebrew script. The spoken dialects have differentiated considerably from Ladino, mainly by borrowing from Hebrew and local languages, and, after further dispersion during and after the World War II, these dialects are now threatened with extinction, though Ladino survives with a mainly religious function.

Portuguese. Portuguese owes its importance largely to its position as the language of Brazil, where between 60,000,000 and 70,000,000 people speak it. In Portugal itself there are about 10,000,000 speakers. The Galician (Gallego, Galego) dialect of northwestern Spain is historically a Portuguese dialect, though now much influenced by the standard Castilian Spanish; about 2,000,000 speakers use Galician as their home language. It is estimated that there are also about 1,500,000 Portuguese speakers in Africa (some of whom also use creole) and about 250,000 each in the United States and Goa, with a few thousand in other Portuguese possessions, such as Macau.

Portuguese dialect groups

There are four main Portuguese dialect groups, all mutually intelligible: (1) Northern, or Galician; (2) Central (Beira); (3) Southern (Estremenho, including Lisbon, Alemtejo, and Algarve); and (4) Insular, including the dialects of Madeira and Brazil. Standard Portuguese was developed in the 16th century, basically from the dialects spoken between Lisbon and Coimbra, to the north. Brazilian (Brasileiro) differs in several respects, in syntax as well as phonology and vocabulary, but many writers still use an academic metropolitan standard. A creolized form, once widespread in Brazil, seems now to be dying out. A Jewish Portuguese is attested in 18th-century Amsterdam (Holland) and Livorno (or Leghorn, in Italy), but virtually no trace of this remains today.

Portuguese speakers have little difficulty in understanding and speaking Spanish, in spite of considerable acoustic and grammatical differences between the two languages. In Portugal, however, they often show considerable resentment at being addressed in Spanish, and there are signs of social resistance to this neighbouring tongue, perhaps because Portugal has so frequently had to play a subordinate role to Spain in the course of its history. In the region of northwestern Spain that adjoins Portugal, the Galician dialects lack uniformity and are closer to Spanish. Even in Castile, where standard Spanish (Castilian) originated, Galician was the conventional language of the courtly lyric until about 1400, but it lost ground in the 15th century, and Castilian replaced Galician as the official language of Galicia in 1500. Dialect poetry in Galician has flourished from the 18th century, with an upsurge in the 19th century.

Before the reconquest of Moorish Spain, largely completed in the 13th century, Galician and Portuguese were indistinguishable. The first evidence for the language consists of scattered words in 9th–12th-century Latin texts; continuous documents date from about 1192, the date assigned to an extant property agreement between the children of a well-to-do family from the Minho Valley. Literature began to flourish especially during the 13th and 14th centuries, when the soft Gallego-Portuguese tongue was preferred by courtly lyric poets all over the Iberian peninsula except in the Catalan area. In the 16th century, Portugal's Golden Age, Galician and Portuguese grew further apart, with the consolidation of the standard Portuguese language.

Italian. Italian is currently spoken by over 50,000,000 people, of whom the vast majority live in peninsular Italy (including the Republic of San Marino), with about 5,000,000 more in Sicily and 1,500,000 Italian speakers in Sardinia. France, including Corsica, has about 1,000,000 Italian speakers and Switzerland about 500,000; there are, in addition, about 300,000 in Yugoslavia. For a large, if decreasing, proportion of these speakers, standard Italian is not the language of the home, where dialectal forms are used. Overseas (*e.g.,* in the United States, where it is estimated that there are 3,500,000 Italian speakers; in Argentina, with 1,500,000; and in Brazil, with 500,000), speakers sometimes do not know the standard language and use only dialect forms. A speaker of an Italian dialect, even one as superficially different as Sicilian, can with effort understand standard Italian, however, and can even teach it to himself by such means as listening to radio programs. For most Italians their first contact with the standard language comes in primary school, in which until recently it was the only dialect used; standard Italian is virtually the only dialect of culture in modern Italy, and with immigration from the south to the industrial north it is becoming more and more the language of intercommunication. Standard Italian is used as a cultural language in Malta and as a co-official language (with English) in Somalia. In Libya and Ethiopia it is now dying out of use.

Distribution of Italian speakers

Standard Italian began to be developed in the 13th and 14th centuries as a literary dialect. At first basically a Florentine dialect, stripped of local peculiarities, it has since acquired some characteristics of the dialect of Rome in particular and has always been heavily influenced by Latin. It overlies a wide variety of dialects, sometimes considered to represent a fundamental differentiation between northern and southern Italy that dates from Roman times. Today, however, these variant dialects form a continuum of intelligibility, although geographically distant dialects may be radically different. The northern dialects include what are often called the Gallo-Italian dialects (Piedmontese, Lombard, Ligurian, Emilian-Romagnol), in which some linguists discern the influence of a Celtic (Gaulish) substratum (*i.e.,* the traces of a language previously spoken in the region). The other northern group of dialects, spoken in northeastern Italy, is called Venetan (including Venetian, Veronese, Trevisan, and Paduan dialects, etc.). Istriot, a language spoken in Yugoslavia, is sometimes considered yet another northern Italian dialect, rather than an independent language. The Tuscan dialects (including those of Corsica) are often held to form a linguistic group of their own, while in the south and east three broad dialect areas are grouped loosely together: (1) the dialects of the Marche (Marchigiano), Umbria, and Rome; (2) Abruzzian, Apulian, Neapolitan, Campanian, and Lucanian; and (3) Calabrian, Otrantan, and Sicilian, believed by some to be influenced by the Greek once spoken there (which still survives in isolated pockets on the toe and heel of the peninsula).

Italian dialect regions

Outside Italy, Italian dialects are heavily influenced by contact with other languages (English in New York; Spanish in Buenos Aires). A pidgin Italian can still be heard in Addis Ababa but has little extension. Relics of a Jewish Italian survive within Italy; a colony of 6,000 Jews, who used a Venetan dialect as a home language in Corfu, was exterminated during World War II.

Early texts from Italy are written in dialect, because there was not yet, of course, any standard dialect. Possibly the very first text is a riddle from Verona, dating from perhaps the 8th century, but its interpretation is obscure and its language Latinized. More surely Italian are some 10th-century documents from Montecassino, after which there are three Central Italian texts of the 11th century. The first literary work of any length is the Tuscan Ritmo *Laurenziano* ("Laurentian Rhythm") from the end of the 12th century, followed soon by other compositions from the Marches and Montecassino. In the 13th century, lyric poetry was first written in a conventionalized Sicilian dialect that influenced later developments in central Italy.

In modem Italy, although dialects are still the primary spoken idiom, standard Italian is virtually the only written language and, with the spread of literary regional varieties of the standard language, may eventually replace the dialects. Neorealism, especially in the cinema, has introduced a limited use of dialect into cultural media, but the relevance of such a development is hotly debated.

Romanian. There are about 20,000,000 speakers of Romanian (or Rumanian), of whom about 16,000,000 live in the Socialist Republic of Romania, 2,500,000 in the U.S.S.R., and about 1,000,000 in Yugoslavia, Bulgaria, Greece, and Albania. There are about 75,000 Romanian speakers in the United States.

Standard Romanian

The standard language of Romania is based on a Walachian variety of so-called Daco-Romanian, the majority group of dialects; it was developed in the 17th century mainly by religious writers of the Orthodox Church and includes features from a number of dialects, though Bucharest usage now provides the model. Daco-Romanian is fairly homogeneous but shows greater dialectal diversity in the Transylvanian Alps, from which region the language may have spread to the plains. Moldavian, the variety of Romanian spoken in the U.S.S.R., is still written in a form of Cyrillic script, and some claim that it is a language in its own right, though most Western linguists see it definitely as a variant of Daco-Romanian. Other dialects of Romanian are barely mutually intelligible with the standard, and some can be counted as separate languages; these include Megleno-Romanian (Vlaši) and Istrio-Romanian, both already mentioned as nearly extinct. More vigorous is the Aromanian group of dialects scattered throughout Greece (150,000 speakers in 1932), Yugoslavia (100,000), Albania (65,000), and Bulgaria (40,000). Numbers have probably decreased considerably, but certainly before the war Aromanians were often prominent businessmen in their localities. The first known inscription in Aromanian, dated 1731, was found only in 1952 at Ardenita, in Albania; texts date from the end of the 18th century, and literary texts have been published in the 19th and 20th centuries (mostly in Bucharest).

The first known Daco-Romanian text is a letter dated 1521, though some manuscript translations of religious texts show Transylvanian dialect features and may be earlier. The vast majority of early texts are written in Cyrillic script, the Roman (Latin) alphabet having been adopted in 1859 at the time of the union of Walachia and Moldavia. Literature in Romanian began to flourish in the 19th century, when the emerging nation turned toward other Romance countries, especially France, for cultural inspiration. Today, in spite of efforts at industrialization, peasant life continues almost unchanged in many regions, and the linguistic standards of the capital have little impact; an autonomous Magyar-speaking (Hungarian) region in the middle of Romania adds further complication to the linguistic situation.

II. Historical survey of the Romance languages

LATIN AND THE PROTOLANGUAGE

Latin and the Italic languages

Latin is traditionally grouped with Faliscan among the Italic languages, of which the other main member is the Osco-Umbrian group. Oscan was the name given by the Romans to a group of dialects spoken by Samnite tribes to the south of Rome. It is well attested in inscriptions and texts for about five centuries before Christ and was used in official documents until c. 90–89 BC. The absence of great dialectal variations in the texts suggest that they are written in a standardized form. In early times, Umbrian was sooken northeast of Rome. to the east of the Etruscan region, possibly as far west as the Adriatic Sea at one period. It is attested mainly in one series of texts, the Tabulae Iguvinae, dated from 400 to 90 BC, and it is similar to Oscan. Probably Latin and Osco-Umbrian were not mutually intelligible; some claim they are not closely related genetically but that their common features arose from convergence as a result of contact.

The Roman dialect was originally one of a number of Latinian dialects, of which the most important was Faliscan, the language of Falerii, 32 miles (51 kilometres) north of Rome. The Faliscans were probably a Sabine tribe that early fell under Etruscan domination. The dialect is known mainly from short inscriptions dating from the 3rd and 2nd centuries BC and probably survived until well after the conquest of Falerii by the Romans in 241 BC. It shares one phonetic feature with Osco-Umbrian (medial f from Indo-European **bh* [the asterisk marks a hypothetical reconstructed form] when Latin has *b*—*e.g.*, Faliscan carefo = Latin carebo), but others are like Latin (*e.g.*, Faliscan cuando = Latin quando = Umbrian pan(*n*)*u*). Some Latin diphthongs, however, appear as simple vowels in Faliscan (*e.g.*, Latin ae = Faliscan *e*), and Latin final consonants are often absent (*e.g.*, Faliscan *cra* = Latin *cras*).

The earliest Latinian text is an inscription on a cloak pin (fibula) of the 6th century BC, from Palestrina (Praeneste); the inscription is definitely dialectal and seems to have Oscan features (*e.g.*, a reduplicated syllable in the perfect form — fhefaked = Latin fecit "he did, made"). Other Latinian inscriptions show marked differences from Roman Latin, for which there is, however, little evidence before the end of the 3rd century BC. What is certain is that the language changed so rapidly between the 5th century (the date of a mutilated inscription said to mark the tomb of Romulus and of the Twelve Tables, the contents of which are known from later evidence) and the 3rd century BC that older texts were no longer intelligible.

During this period the Romans subjugated their Latin neighbours (by 335 BC), and their language began to establish itself as a standard form, absorbing features from other dialects. The first author of any note was the comedian Plautus (254–184 BC), whosk language is thought to reflect a spoken idiom, some features of which appear to have survived into Romance.

Literary development in Latin

By 265 Rome had conquered Magna Graecia, in the south of the Italian peninsula, and had begun to absorb some of its Greek literary and cultural ideals. Poetic language was especially influenced by Greek until Latin poetry reached its zenith with Virgil. In the 1st century BC a literary prose was to be developed, with emphasis placed on rejection of vulgarity and rusticity and pride of place accorded to elegance and clarity. Grammatical rules were codified and tightened and vocabulary pruned, and the cult of the harmonious, balanced period held sway in rhetorical circles. With Cicero, Golden Age prose style attained its highest point; for the linguist, the distinction Cicero makes between the style of his letters and that of his speeches is especially interesting in that it provides evidence that even educated speech differed from written language. When Cicero uses the *sermo plebeius* ("plebeian speech"), his language is more elliptical, with shorter, less complex sentences and more colourful vocabulary (including plentiful diminutives). It seems obvious that truly popular language differed even more from the elaborate, sophisticated, classical literary idiom; there is evidence that archaic features, banned from literary style, survived in vulgar speech right through to the Romance stage of the language. It is sometimes claimed that the language of Roman historian and politician Sallust (86–35 BC) approximated popular usage, but it is more probable that his archaizing style derives more from conscious imitation of old Roman poetry. The Roman "judge of elegance" Petronius Arbiter (died AD 65/66), too, is often thought to imitate vulgar speech, but many of the odd features found in his "Cena Trimalchionis" ("Trimalchio's Dinner") may represent the broken Latin spoken by Greeks and such.

SOME CHARACTERISTICS OF CLASSICAL LATIN

Pronunciation. Evidence for pronunciation of the Latin of the classical era is often difficult to interpret. Orthography is conventionalized, and grammarians' comments lack clarity, so that to a considerable extent it is necessary to extrapolate from later developments in Romance in order to describe it. On the whole, linguists think that Latin probably sounded something like Italian, though areas of uncertainty exist.

Among these uncertainties, the most important concerns Latin intonation and accentuation. The way vowels developed in prehistoric Latin suggests that there was a heavy stress accent on the first syllable of each word, but in later times the accent fell on the penultimate syllable or, when this had "light" quantity, on the antepenultimate (much as in modern Italian). The nature of this accent is hotly disputed: contemporary grammarians seem to suggest it was a musical, tonal accent and not a stress accent. If this were so, the acoustic effect of Latin would be quite different from Romance and similar, perhaps, to modern West African languages or even Chinese. Some scholars claim, however, that Latin grammarians were merely slavishly imitating their Greek counterparts and that the fact that in Latin accent is linked with syllable vowel length makes it unlikely that such an accent was tonal. Probably it was a light stress accent that was normally accompanied by a rise in pitch; in later Latin evidence suggests that the stress became heavier.

The system of syllable quantity, connected with that of vowel length, must have given Classical Latin distinctive acoustic character. Broadly speaking, a "light" syllable ended in a short vowel and a "heavy" syllable in a long vowel (or diphthong) or a consonant. The distinction must have been reflected to some extent in late Latin or early Romance, for, even after the system of vowel length was lost, light, or "open," syllables often developed in a different way from heavy, or "closed," syllables.

Vowel length

Because the system of vowel length was lost after the classical period, it is not known with any certainty how vowels were pronounced at that period; but, because of later developments in Romance, the assumption is that the vowel-length distinctions were also associated with qualitative differences, in that short vowels were more open, or lax, than long vowels. Standard orthography did not distinguish between long and short vowels, although in early times various devices were tried to remedy this. At the end of the Roman Republic an "apex" (one form was like this: ') often was used to mark the long vowel, but this was replaced in imperial times by an acute accent (′). In Classical Latin the length system was an essential feature of verse, even popular verse, and mistakes in vowel length were regarded as barbarous. In later times, however, many poets were obviously unable to conform to the demands of classical prosody and were criticized for allowing accent to override length distinctions.

Besides the vowels *ā, ē,* i, *ō, ū* (long vowels) and *ă, ĕ, ĭ, ŏ, ŭ* (short vowels), educated speech at the classical period used a sound taken from Greek upsilon and pronounced rather like French *u* (the symbol [y] in the International Phonetic Alphabet — IPA) in words borrowed from Greek; in popular speech this was probably pronounced like Latin *ŭ*, though in later times *ī* sometimes substituted for it. A neutral vowel was probably used in some unaccented syllables and was written *u* or i (*optumus,* optimus "best"), but the latter rendering became standard. A long *ē*, from earlier ***ei***, had probably completely merged with i by the classical period. Classical pronunciation also used some diphthongs pronounced by educated Romans much as they are spelled, especially ae (earlier ai), pronounced perhaps as an open long e in rustic speech, au (rustic open long *ō*), and oe (earlier oi, late Latin *ē*).

The consonant system

The Classical Latin consonant system probably included a series of labial sounds (produced with the lips), p, b, *m*, *f*, and probably *w*; a dental or alveolar series (produced with the tongue against the front teeth or the alveolar ridge behind the upper front teeth), t, d, *n*, *s*, *l*, and possibly ɫ; a velar series (produced with the tongue approaching or contacting the velum or soft palate), *k*, *g*, and perhaps *ng;* and a labiovelar series pronounced with the lips rounded, k^w and g^w. The *k* sound was written c, and the k^w and g^w were written *qu* and *gu*, respectively.

Of these, k^w and *g''* were probably single labialized velar consonants, not clusters, as they do not make for a heavy syllable; g^w occurs only after *n*, so only guesses can be made about its single consonant status. The sound *ng* (as in English "sing"; represented in the International Phonetic Alphabet by [ŋ]]), written *ng* or *gn*, may not have had phonemic status (in spite of the pair *annus/agnus* "year"/"lamb," in which [ŋ] may be regarded as a positional variant of g). The Latin letter f probably represented by classical times a labiodental sound pronounced with the lower lip touching the upper front teeth like its English equivalent) but earlier it may have been a bilabial (pronounced with the two lips touching or approaching one another). The so-called consonantal i and *u* were probably not true consonants but frictionless semivowels; Romance evidence suggests that they later became a palatal fricative, [j] (pronounced with the tongue touching or approaching the hard palate and with incomplete closure) and a bilabial fricative, [β] (pronounced with vibration of the lips and incomplete closure), but there is no suggestion of this at the classical period. Some Romance scholars suggest that Latin *s* had a pronunciation like that of modern Castilian (with the tip, rather than the blade, raised behind the teeth, giving a lisping impression): in early Latin it was often weakened in final position, a feature that also characterizes eastern Romance languages. R was probably a tongue trill at the classical period, but there is earlier evidence that in some positions it may have been a fricative or a flap.

The nasal consonants were probably weakly articulated in some positions, especially medially before *s* and in final position; here probably there was mere nasalization of the preceding vowel.

In addition to the consonants shown, educated Roman speakers probably used a series of voiceless aspirated stops, written *ph, th, ch,* originally borrowed from Greek words but also occurring in native words (pulcher "beautiful," *lachrima* "tears," *triumphus* "triumph," etc.) from the end of the 2nd century BC.

Another nonvocalic sound, *h,* was pronounced only by educated speakers even in the classical period, and references to its loss in vulgar speech are frequent.

Doubled consonant sounds

Consonants written double in the classical period were probably so pronounced (a distinction was made, for instance, between anus "old woman" and *annus* "year"). When consonantal i appeared intervocalically, it was always doubled in speech. Earlier than the 2nd century BC consonant gemination (doubling of sounds) was not shown in orthography but was probably current in speech. Among the Romance languages, the eastern ones on the whole retained Latin double consonants, whereas in the west they were often simplified.

Morphology and **syntax.** Latin reduced the number of Indo-European noun cases from eight to six by incorporating the sociative-instrumental (indicating means or agency) and, apart from isolated forms, the locative (indicating place or place where) into the ablative case (originally indicating the relations of separation and source). The dual number was lost, and a fifth noun declension was developed from a heterogeneous collection of nouns (principally verbal abstracts in *-ie*). Of the other declensions, the Indo-European *ā*- and *ŏ* classes remained, with the introduction of new genitive singular forms in -ce and *-i*, while consonantal and *-i* stem nouns were amalgamated into a "third" declension, which also took in adjectives formerly of the *ū*-class (a "fourth" declension). Probably before the Romance period the number of cases was further reduced (there were two in Old French—nominative, used for the subject of a verb, and oblique, used for all other functions—and Romanian today has two, nominative–accusative, used for the subject and the direct object of a verb, and genitive–dative, used to indicate possession and the indirect object of a verb), and words of the fourth and fifth declension were absorbed into the other three or lost.

Among verb forms, the Indo-European aorist (indicating simple occurrence of an action without reference to duration or completion) and perfect (indicating an action or state completed at the time of utterance or at a time spoken of) combined, and the conjunctive (expressing ideas contrary to fact) and optative (expressing a wish or hope) merged to form the subjunctive mood. New tense forms that developed were the future in *-bō* and the imperfect in *-bam;* a passive in *-r*, also found in Celtic and Tocbarian, was also developed. New compound passive tenses were formed with the perfect participle and *esse* "to be" (*e.g.*, esf *oneratus* "he, she, it was burdened°)—such compound tenses were to develop further in Romance. In general, the morphology of the classical period was codified and fluctuating forms rigidly fixed. In syntax, too, earlier freedom was restricted; thus, the use of the accusative and infinitive in *oratio obliqua* ("indirect discourse") became obligatory, and fine discrimination in the use of the subjunctive was insisted on. When earlier writers might have used prepositional phrases, classical authors preferred bare nominal-case forms as terser and more exact. Complex sentences with subtle use of distinctive conjunctions were a feature of the classical language, and effective play was made with the possibilities offered by flexible word order.

New tense forms

In the postclassical era, Ciceronian style came to be regarded as laboured and boring, and an epigrammatic, compressed style was preferred by such writers as Seneca and Tacitus. Contemporaneously and a little later, florid, exuberant writing—often called African—came into fashion, exemplified especially by Apuleius (2nd century AD). Imitation of classical and postclassical models continued even into the 6th century, and there seems to have been continuity of literary tradition for some time after the fall of the Western Roman Empire.

The growth of the empire spread Roman culture over much of Europe and North Africa. In all areas, even the outposts, it was not only the rough language of the legions that penetrated but also, it seems, the fine subleties of Virgilian verse and Ciceronian prose. Recent research suggests that in Britain, for instance, Romanization was wider spread and more profound than hitherto suspected and that well-to-do Britons in the colonized region were thoroughly imbued with Roman values. How far these trickled down to the lower classes is difficult to tell; because Latin died out in Britain, it is often thought that it had been used only by the higher classes of the population, but some suggest that it was a result of wholesale slaughter of the Roman British. It is, however, more likely that the pattern of Anglo-Saxon settlements was not in conflict with the Romano-Celtic and that the latter were gradually absorbed into the new society.

Develop-mentof Romance from Latin

In the lands in which Romance is still spoken, it is of course certain that, sooner or later, Latin in some form was the normal language of most strata. Whether, however, the Romance languages continue rough peasant dialects of Latin (or even slave creoles) or the usage of more cultured urban communities is open to question. There are those who maintain that tbe Latin used in each area differentiated as soon as local populations adopted the conqueror's language for any purpose. According to this belief, dialects of Latin result from "interference" from indigenous languages (substrata), even though clear evidence for dialectal diversification cannot be found in extant texts. It is obvious that Latin usage must have differed over a wide area, but it can be questioned whether the differences were merely phonetic and lexical variations—regional accents and usage—not affecting mutual intelligibility or whether they were profound enough to form the basis of further differentiation when administrative unity was lost. The latter hypothesis would suggest a long period of bilingualism (up to about 500 years), as experience shows that linguistic interference between languages in contact rarely outlives the bilingual stage. Virtually nothing is known about the status of the indigenous languages during the imperial period, and only vague contemporary references can be found to linguistic differences within the empire. It seems odd that no one among the numerous Latin grammarians should have referred to well-known linguistic facts, but the absence of evidence is not sufficient to justify the assertion that there was no real diversification during the imperial era. Historical parallels are lacking—the British Empire did export English to widely different lands, but it lasted a comparatively short time, and its linguistic contribution was backed by modern communications media, besides being to some extent negated by nationalist feeling.

Latin and Romance coexisting

What is certain is that, even if popular usage within the empire showed great diversification, it was overlaid by a standard written language that preserved a good degree of uniformity until well after the administrative collapse of the empire. As far as the speakers were concerned, they apparently thought they were using Latin, though they were often conscious that their language was, through sheer ignorance, not quite as it should be. Not until about the 8th or 9th century—later in some parts—did it strike them that Classical Latin was perceptibly a different language, rather than merely a more polished, cultured version of their own.

Later Latin (3rd century AD onward) is often called Vulgar Latin—a confusing term in that it can designate the popular Latin of all periods and is sometimes also used for so-called Proto-Romance (roman commun), a theoretical construct based on consistent similarities among all or most Romance languages. All three Vulgar Latins in fact share common features but, given their different theoretical status, can hardly be called identical or even comparable. Written Vulgar Latin attained wide diffusion as the language of the Christian Church, officially adopted by the empire from the 4th century on. Its "vulgarisms" often called forth apologies from Christian authors, whose false humility seems akin to pride in that they did not succumb to the frivolities of pagan literary style.

Aside from the numerous inscriptions from all over the empire, there is no shortage of texts in Vulgar Latin. One of the first is the so-called Appendix Probi (3rd–4th centuries AD; "Appendix of Probus"), which lists correct and incorrect forms of 227 words, probably as an orthographical aid to scribes, but as a result illustrates some phonological changes that may have already occurred in the spoken language (*e.g.*, loss of unstressed penultimate syllables and loss of final m). The Vulgate, St. Jerome's translation of the Bible (AD 385–404), and the works of St. Augustine (AD 354–430) are among Christian works in Vulgar Latin. Particularly amusing and linguistically instructive also is the so-called Peregrinatio Etheriae ("Journey of Etheria"), written probably in the 4th century by a nun, describing her visit to the Holy Land. Medical and grammatical works also abound from the 4th to the 7th centuries (among the writers were the provincials Cosentius, from Gaul; Virgilius Maro, from southern Gaul; and St. Isidore of Seville, from Spain).

New vocabulary

Some of the characteristics of Vulgar Latin recall popular features of classical and preclassical times and foreshadow Romance developments. In vocabulary, especially, many of the sober classical words are rejected in favour of more colourful popular terms, especially derivatives and diminutives: thus, *portare* "to carry" (French porter, Italian *portare,* etc.) is preferred to ferre; *cantare* "to sing again and again" (French chanter, Spanish and Portuguese *cantar,* etc.) to *canere*; *vetulus* "little old man" (Romanian vechi, Italian vecchio, French *vieux,* etc.) to *vetus.* In grammar, classical synthetic constructions are often replaced by analytic; thus, the use of prepositions often makes case endings superfluous. *Ad regem* for regi "to the king," for instance, or anomalous morphological forms are simplified and rationalized (*e.g.,* plus *sanus* for *sanior* "healthier"). Shorter, less complex sentences are preferred, and word order tends to become less flexible.

The most copious evidence for Vulgar Latin is in the realm of phonology, though interpretation of the evidence is often open to dispute, consisting as it does of the confused descriptions of grammarians and the misspellings of bewildered scribes. Much of the evidence points to a strengthening of stress accent during the late period, leading to the shortening and swallowing of unaccented syllables: thus, viridem "green" becomes virdem (verde in several Romance languages); vinea "vine" becomes vinia (French vigne, Spanish *viña* ["vineyard"], etc.). It is often thought that Classical Latin had a tonal, not a stress, accent; though this is uncertain, any stress on accented syllables was probably light and less acoustically perceptible than an accompanying rise in pitch. There is some scant evidence that a stress accent was used in popular and dialectal preclassical speech (*e.g.,* vinia in a 3rd-century-BC epitaph, Oscan minstreis for Latin *minister* "attendant"), but the first undisputed testimony is to be found in 3rd-century-AD texts.

Among other phonological features of Vulgar Latin, probably the most striking is the loss of the system of long and short vowels. On the whole, long vowels became tense and short vowels lax, resulting in a wholesale change in the rhythm of the language. In the texts there is evidence of the confusion of *ĭ* and *ē* and of *ŭ* and *ō* that has occurred in the western Romance languages. A similar collapse of *ō* and *ŭ* seems to have occurred in Oscan, but it is unlikely that this is connected with later developments. It is to be remembered that even popular Latin verse used measures of vowel length, and there is no evidence to suggest that vowel-length distinctions were lost in vulgar preclassical speech.

An archaic feature that does recur in Vulgar Latin is the loss of word-final -m, of which virtually no trace remains in Romance. It is possible, however, that the written letter of Classical Latin was no more than an orthographical convention for a nasal twang: in scanning Latin verse, the *-m* is always run in (elided) before a vocalic initial. Reduction of the diphthongs ae (to e) and au (to o) seems also to be a popular and dialectal feature reflected in Vulgar Latin texts; in the latter case, however, the Romance languages do not support the hypothesis that the diphthong was reduced early, for it remains in Old Provençal and in Romanian and, probably, in early Old French.

Borrowing from Latin into non-Romance languages

The prestige of Rome was such that Latin borrowings are to be found in virtually all European languages, as well as in the Berber languages of North Africa, which preserve a number of words, mainly agricultural terms, lost elsewhere. Basque has borrowed a good number of words, mainly from administrative, commercial, and military spheres, though it is difficult in some cases to determine whether the terms were later borrowings from Spanish, rather than from Latin. This is not a problem in the case of the 800 Latin words found in three British Celtic languages (Welsh, Cornish, and Breton)—words drawn from a wide sphere of activities. In the Germanic languages, borrowed Latin words principally involve trade and often reflect archaic forms. The very large number of Latin words in Albanian form part of the basic vocabulary of the language (including kinship terms) and cover such spheres as religion, although there is doubt about whether some of them were later borrowings from Romanian. In other cases Latin words in Albanian have survived in no other part of the former Roman Empire. Greek and Slavic languages have comparatively few Latin words, many of them administrative or commercial in character.

Latin has had a continuous influence on the Romance languages and their neighbours in its capacity as a language of religion and culture. With Christianity, Latin penetrated to new lands, and it was perhaps the cultivation of Latin in a "pure" form in Ireland, whence it was exported to England, that paved the way for an 8th-century reform of the language by Charlemagne. Conscious that current Latin usage was falling short of classical standards, Charlemagne invited Alcuin of York, a scholar and grammarian, to his court at Aix-la-Chapelle (Aachen), where he remained from 782 to 796, inspiring and guiding an intellectual renaissance. It was perhaps as a result of the revival of so-called purer Latin that vernacular texts began to appear, for it now became obvious that the vernacular and Latin were not the same language. Thus, in 813, just before Charlemagne's death, the Council of Tours decreed that sermons should be delivered in *rusticam Romanam* linguam ("in the rustic Roman language") to make them intelligible to the congregation.

Latin as the language of religion and education

Latin has remained the official language of the Roman Catholic Church and as such has been in constant use by most Romance speakers; it is only very recently that church services have begun to be conducted in vernacular. As the language of science and scholarship, Latin held sway until the 16th century, when, under the influence of the Reformation, nascent nationalism, and the invention of the printing press, it began to be replaced by modern languages. Nevertheless, in the west, along with Greek, the Latin language has remained a mark of the educated man throughout the centuries, although since World War II the popularity of classical languages in schools has declined, and a generation of scholars who know no Latin, except for the numerous terms borrowed by all European languages, will soon be seen.

THE EMERGENCE AND DEVELOPMENT OF THE ROMANCE LANGUAGES

Earliest period. The question of when Latin ended and Romance began, which has occupied scholars in the past, is largely a problem created by terminology. In some senses, today's Romance languages are regional varieties of one uniform set of speech patterns that resembles the Vulgar Latin of attested texts fairly closely—indeed, the analyses of generative phonologists make the modem "underlying forms" (as distinct from their phonetic representation in speech) look almost identical with the reconstructed ancestor of the Romance languages, Proto-Romance. On the other hand, sveakers are conscious that today they are speaking a "different language" from their neighbours, even though they may understand a good deal of their neighbours' discourse. Perhaps the speaker's consciousness is the best measure of divergence; when, one may ask, did Romance speakers realize that they were not using Latin in their everyday speech? Some scholars suggest that the realization must have dated from about the 5th century, when barbarians were streaming into the Roman Empire and, supposedly, hindering communication. Others prefer to rely on positive textual evidence, indicative of efforts to make up a written form of Romance distinct from Latin. Such evidence begins to appear only in the 9th century, first in northern France and then in Spain and Italy. The reforms of Charlemagne, re-establishing more classical standards in written Latin, may have been at once cause and result of the development of conventional written forms for vernacular Romance. Perhaps it was also the emergence of a new type of social organization, feudalism, that had linguistic effects as a result of the splitting of the open society of Roman tradition into small closed territorial units.

Romance glosses to Latin texts

From the 7th century onward, consciousness of linguistic change was strong enough to prompt scribes to gloss little-known words in earlier Latin texts with more familiar terms. Though the glosses often reflect Romance forms, however, they are usually given in a Latinate form, and one gains the impression of a few superficial adjustments to archaic but fundamentally comprehensible texts. The best known set of glosses—to the Vulgate Bible of St. Jerome—formerly belonged to the abbey of Reichenau, on an island in Lake Constance, Germany, and probably dates from the 8th century. The vocabulary of the Reichenau glosses appears to be French in flavour (*e.g., arenam* "sand" glossed by sabulo, French sable; vespertiliones "bats," by calvas *sorices*, French *chauve-souris*), and some words of Frankish origin appear (*e.g.*, scabrones "beetles" is glossed by wapces "wasps," *respectant* "they look about" by rewardant). The glosses provide some evidence of morphological simplification (*e.g.*, saniore "healthier" is glossed by plus sano "more healthy" and cecinit "he sang" by cantavit), but for the most part only lexical items are regarded as meriting comment. Another well-known glossary, known as the Kassel (or Cassel) glosses, probably dates from the very early 9th century. It gives Latin equivalents of German (Bavarian) words and phrases and provides evidence of lexical and phonetic differentiation within Latin that permits scholars to localize the work as probably French or Rhaetian (*e.g., mantun* "chin," as compared with modern French *menton*). Although orthographically eccentric, however, the text is obviously meant to represent Latin, not a Romance tongue; when phrases rather than isolated words are glossed, the Latin is often very close to classical models.

Beginnings of Romance literature

Later in the 9th century (with the Strasbourg Oaths, possibly, and more clearly in the Eulalia poem), deliberate attempts were made to write vernacular Romance, though the resources of the Latin alphabet were not wholly adequate to the task. That northern French texts were the first to appear is not surprising, for in that region Latin had changed more radically than elsewhere. By the 10th century the need to couch legal documents in more readily comprehensible vernacular, rather than Latin, was felt in other regions. Vernacular literature did not really get under way, however, until around the 12th century, when the arts flourished throughout western Europe. Raeto-Romance and Romanian, however, had to wait for the Reformation period to take on literary form.

Late-medieval period to the Renaissance. There was a good deal of cross-fertilization between Romance literary languages during the period of development of medieval poetry; the example of the Provençal lyric especially left its mark on all vernacular literatures, and borrowing of lexical items from one language to another was abundant. The 13th century saw some shift of linguistic influence from southern to northern France and from Sicily to Tuscany, toward the politically and economically more powerful regions. Portuguese and Catalan developed flourishing literatures somewhat later, taking over some of the traditions of the badly battered southern French region and dominating the literary scene of the Iberian Peninsula. French was fast losing its hold in England, which, a century earlier, had boasted a rich Anglo-Norman literature, and within France the central Parisian dialect began to dominate. In Italy, the Florentine dialect was showing signs of rising to prominence and providing the base for a literary standard.

Influence of Classical Latin on Romance

The rediscovery of classical literature and art, first in Italy and then in other Romance regions, had some considerable effect on the languages in the shape of extensive borrowing from Latin and Greek and, often, conscious attempts to model grammatical constructions in the vernacular on Classical Latin. The Italian standard language, in particular, owes much to the influence of Latin, which it resembles more closely than do the spoken dialects. French, except in the 16th century, was influenced grammatically less by Latin, but from the 14th century onward the habit of preferring words with a quasi-Latin shape to inherited forms became well established, so that much of the French vocabulary has a "learned" appearance. The trickle of Latinisms into Spanish became a flood in the 15th century, and, though Spanish has been more reluctant than French to reject old words, they today form a considerable proportion of the lexicon.

Standardization of the Romance languages in the 17th and 18th centuries. It was in Italy first that the "question of the language" became a matter of hot dispute. Dante himself made an important contribution to the debate on what should constitute a *volgare illustre* (an "illustrious popular speech") capable of rivalling Latin for literary and scholarly purposes. Controversy did not reach its peak, however, until the 16th century. In the Spain of 1492 the completion of the reconquest of Spain from the Arabs and the discovery of America were matched linguistically by the appearance of Antonio de Nebrija's

Gramática Castellana ("Grammar of the Castilian Language"), which argues the need for an ennobled language fit for imperial exportation. In France during the 16th century, with the Renaissance backed by the Reformation and the advent of printing, French really took over the remaining functions of Latin—scholarly, scientific, and religious—and efforts were made to put together a worthy national language from dialect and Latin sources. The choice of standard was not made definitively, however, until the late 17th century, when, with political power and social influence centred exclusively at the royal court, the only acceptable usage became that of the court. It would seem that social acceptance and advancement were inextricably bound up with correct behaviour, especially linguistic behaviour, so that the well-to-do bourgeoisie set out to ape the speech habits of their "betters"—hence the popularity of works describing *le bon* usage "good usage." The influence of French, resplendent with the achievements of French dramatic poet Racine and of Louis XIV, was destined to remain dominant within the Romance languages; the Golden Age of Spain and Portugal had already passed, and Italy was going through a period of comparative stagnation.

"Correct" language

The French grammarians of the 18th century had lasting effect on all the Romance standards, concerned as they were with maintaining "purity," eliminating "vulgarity," and strictly codifying usage, often more in accord with logical than linguistic considerations. The belief that correct language is not a birthright but a tool to be carefully fashioned and skillfully handled, that conscious effort was required to allow it to mirror thought with the minimum of distortion, is one that has persisted in Romance and that still has important effects on educational practice. To many English speakers it seems ludicrous that the criterion of competence in a language should be strict adherence to grammar-book rules rather than native-like performance, but in Romance countries a foreigner is often frowned upon if he permits himself the "negligence" of native usage, rather than the more stilted correct expression. Educated Romance speakers often speak very formally, with flowing, complex sentences and precise vocabulary, in contrast with the casual, slangy expression of the less educated (who openly envy the speech habits of their "betters"). The passionate interest shown in subtleties of language usage (including regular articles in the better class newspapers) is something that characterizes all Romance speakers, though perhaps only the French take it to excess.

Modern developments. The Romantics of the early 19th century were eager to break the stranglehold of intellectual, aristocratic language, but their attempts to introduce more colourful expressions did not bring them nearer to popular usage, for their efforts were mainly directed toward enriching the vocabulary while leaving grammar intact. Sentimental idealization of peasant existence aroused interest in dialectal usage, and egalitarian sentiments provoked some groups of speakers to proclaim the worthiness of their own mother tongues to rival more politically important languages. Occitan, Catalan, Rhaeto-Romance, and, indeed, Romanian were to develop literatures under the impact of such ideas, which took political form in the demands of regional separatists.

Popular usage in literature

The introduction into literature of conventionalized popular usage is mainly a 20th-century trend, but in the Romance languages it remains more limited than, for example, in English. Other, more literary attempts to break out of the straitjacket of standard usage are connected with such artistic movements as Symbolism and Surrealism, in which syntax, as well as vocabulary, suffers onslaught, and sentence construction tends to cut loose from logical ties. Yet, even within these movements, fine, elegant style continues to be appreciated by many, and it is the traditional forms that have wider appeal. The use of a weighty bureaucratic style in nonliterary writings is also evident, often characterized by an excessive use of Latinism and by the use of verbal nouns, but educational systems continue to place emphasis on plain, classic style. It is difficult to imagine that the long-continuing tradition of interest in "correctness" in language will die out in the Romance countries; recent educational reforms of France and Italy seem rather to emphasize the value of such a tradition, and organizations and individuals, no matter how revolutionary in political outlook, rarely work to counteract the cultural values so long regarded as paramount in their homelands.

III. Characteristics of the Romance languages

As a group, the Romance languages share many characteristics besides that which defines the family (*i.e.*, the presence of a significant proportion of lexical cognates). In comparison with Germanic languages, for instance, they seem musical and mellifluous—probably because of the relatively greater importance of vowels than consonants. On the whole, the vowels are clear and bell-like and articulation energetic and precise, though Portuguese and Romanian convey a more muted acoustic impression. Foreigners often think that Romance speech is particularly rapid and voluble, no doubt because individual words receive only light stress (or, in French, no stress), and elision, the running of words into each other within stress groups, is common. Romanian is something of an exception in that speech tempo is comparatively slow. Intonation patterns, surface manifestations of nonlexical meaning, such as interrogation, exclamation, scorn, surprise, and so forth, seem to some to denote excitability and emotional expressiveness in the speakers. Northern French is comparatively sober, with typically about a one-octave range in intonation, but Italian seems to be sung, with sinuous pitch movement over two octaves, and Castilian jumps jerkily and up and down over about an octave and a third.

Retention of the synthetic character of Latin

Grammatically, the modern languages have retained to a greater or lesser extent some of the synthetic character of Latin, principally in the verb, but in Romanian also in the noun. French, since about the 14th century, has undergone most radical changes in grammatical typology, so that much greater reliance is placed on word order and intonation to convey sentence meaning than on morphological form. Other languages allow a little more flexibility of word order but far less than in Classical Latin.

Dominant purist grammarians have always opposed influence from foreign languages and reproved their fellows for sullying their language with lavish borrowing (at present primarily from English), but they have never been able to stem the flood of neologisms. French vocabulary, particularly, has always been receptive to change and has been as quick to lose old words as to adopt new. Codification of grammar, on the other hand, has had a permanent effect on the stability of the standard languages, even feeding back into spoken usage via the education system. Acceptance of the most minor changes follows long debate and deliberation and requires governmental edicts that decree what can be marked as correct in all-important examinations. Curiously enough, this rigidity and consequent self-confidence have resulted in greater teachability, so that standards of correctness of, for instance, French among Africans or Spanish among American Indians are remarkably high. The moves toward codification were, indeed, originally linked to a desire to give the languages international importance, and language teaching is, in the Romance ethos, indissolubly linked to the diffusion of cultural and moral values.

LINGUISTIC TYPOLOGY OF THE ROMANCE LANGUAGES

As stated previously, the most "central" Romance language is standard Italian, which has retained and even re-adopted many Latin characteristics. In some ways its morphology lacks the elegance and efficiency of Castilian, which has most ruthlessly eliminated anomalies during the modem period; there are signs in Italian of historical inertia, a harking back to a glorious past, that has hindered popular development. Romanian remains closest in grammatical type to Latin, though its noun-declenison system, based on the definite article placed after the noun, and its frequent use of the subjunctive mood may owe much to its Balkan neighbours (or to an earlier linguistic substratum). Its vocabulary has incorporated so many Slavic and Turkish words, however, that it often

Causes of the divergences in French

appears less Romance than the rest. French, by any standard, has diverged most—radical phonetic changes that transformed the outward appearance of the language must have preceded the earliest surviving (9th-century) texts. Such changes are usually ascribed to Celtic and Frankish influence. Another wave of change, with loss of word accent and of many morphological markers, probably dates from around the 15th century, but it is difficult to find external motivation for these phenomena. Occitan and Catalan are conservative in character; the long persistence of Roman schools in South Gaul is often seen as the cause of stability there. Spanish and Portuguese are close enough to lead some scholars to assign their shared characteristics to Iberian substratum and Moorish superstratum influence. Castilian's forceful character and receptivity to grammatical innovation contrast sharply with Portuguese softness and its inertia in retaining morphological oddities, however. One might conceivably see the differences as connected with climatic and geographical conditions, though just how would be difficult to discern. Rhaeto-Romance and Dalmatian peculiarities can most easily be connected with the impact of other languages (mainly German, Italian, and Serbo-Croatian), while Sardinian is often regarded as an extremely conservative, peasant language, some dialects of which have been penetrated by features from Italian and Spanish.

PHONOLOGY

Some important phonological developments, such as the loss of the system of contrasting vowel lengths and the strengthening of the stress accent, must have occurred during the Vulgar Latin period, while some degree of unity still existed among the various Romance dialects. Certain other changes shared by the Western Romance languages, especially the collapse of *ē* and i, might have postdated the linguistic separation of Sardinia and parts of southern Italy from the other areas, while the distinct development of 6 and *ŭ* in Romanian and Vegliot suggests a split between Eastern and Western Romance at a later date.

Unaccented vowels

Vowels. Everywhere, unaccented vowels have had a different history from accented, and in some languages they have so weakened as to disappear altogether in certain positions. At the end of a word, for instance, even -a, the most sonorous of the vowels, has weakened to a neutral vowel in Romanian, Portuguese, and some Catalan and Rhaetian dialects—in some French dialects it is still pronounced as a neutral vowel sound (such as the second vowel in English "alphabet"), but it has been lost completely in the standard language. Final *-o*, from Latin *-o* or *ŭ*, was lost very early in French, Occitan, Catalan, and Rhaetian and remains only before an article following the word in Romanian; in Portuguese it is closed to a *u* sound (such as the *u* in English "lunar"). Final *-e* is even more evanescent, regularly remaining as a full vowel only in parts of central and southern Italy and Sardinia.

Under the main stress accent of the word, Latin vowels have often become diphthongs in Romance, perhaps as a result of lengthening under heavy stress or as a consequence of the raising influence of following high vowels (a process known as breaking, similar in action to German umlaut). The vowels most affected are the "open" *e* sound (as in "met"), from Latin *ĕ*, and to a lesser extent the "open" o sound (similar to the aw sound in "law" in many American English dialects and to the o in British English "ingot"), from Latin 6, while high close vowels i and *u* are virtually untouched. Transformation of short *e* to a diphthong (usually a ye sound, as in "yet") is so common that some believe it occurred during the Vulgar Latin period. The conditions of this process (and similar ones) vary, however; in some languages (notably French and Italian) it happens only in open syllables (*i.e.*, those ending in a vowel in Vulgar Latin), whereas Romanian, Vegliot, Spanish, and perhaps Rhaetian show similar developments in all accented syllables. Portuguese possibly did not join in the diphthong-forming process at all, though, as in Occitan, Catalan, Sardinian, and some Italian dialects the short e- and *o*- sounds may at one time have developed into diphthongs under the influence of a following high vowel (i or u), later to be reduced once more to a single vowel. Table 1 illustrates treatment of stressed Latin *ĕ* and 6 in different languages.

Table 1: Occurrence of Diphthongs Replacing Stressed Short Vowels in Romance Languages

	pĕde "foot"	*hhrba* "grass, herb"	*mŏrit* "he dies"	*mŏrtem* "death"
Sardinian	*pe*	*erva*	*móridi*	*morte*
Portuguese	*pe*	*herva*	*morre*	*morte*
Catalan	*peu*	*herba*	*mor*	*mort*
Occitan	*pe*	*erba*	*mor*	*mort*
French	*pied*	*herba*	*meurt* (Old French *muert*)	*mort*
Italian	*piede*	*erba*	*muore*	*morte*
Romanian	—	*iarbă*	*moare*	*moarte*
Spanish	*pié*	*hierba*	*muere*	*muerte*
Rhaetian (Sursilvan)	*pei*	*jarva*	*miere*	*mort*
Rhaetian (Friulian)	*pid*	—	—	*muart*
Vegliot	*pi*	*járba*	—	*muart*

Reflexes of Latin *ō* (and *ŭ*) and *ē* (i) became diphthongs ou and ei in Northern French at an early period (after the 5th but before the 9th century); the 12th-century phonetic results *eu* and oi provided the present-day spellings, though the sounds thus represented have changed considerably since (compare *fleur* "flower," from flour, from Pore). The greater extension of spontaneous diphthong formation in French than in other Romance languages (including perhaps also reflexes of a--compare mer "sea," from **maer* [?], from Latin mare) is often attributed to the effects of the heavy stress presumably used by the Frankish superstratum.

Effects of nasal consonants on vowels

In nearly all Romance languages a following nasal consonant has caused peculiar development in a preceding vowel. In most cases the effect is limited to a raising or closing influence, but in two major languages, French and Portuguese, phonological nasalization has taken place (*i.e.*, a series of vowels distinguished by the presence of nasal resonance has developed). Here, as well as in some other dialects (especially Chilean, Caribbean, and Andalusian Spanish, in the Romanian spoken in Albania, and in northern Occitan), nasal vowels are distinct from their oral counterparts and not mere variants (*i.e.*, they are phonemic). Thus, they serve to differentiate one meaningful form from another: *e.g.*, French pin "pine," pronounced *pɛ̃* (ɛ stands for a short *e* sound, and ˜ marks nasalization) versus *paix* "peace," pronounced *pɛ*; Portuguese *lã* "wool" versus la "there"; Andalusian *cantã* "they sing" versus *canta* "he sings." Occasionally, nasalization of a vowel is caused by a preceding consonant (*e.g.*, Portuguese *mãe* "mother," from *matre*), but this is comparatively rare.

Nasalization in both French and Portuguese was probably noticeable by the 10th century, though it may not have become phonemic until much later. Some claim that even today nasal vowel resonance is merely a surface manifestation of a latent underlying nasal consonant. It would appear that in both languages nasal vowels were more frequent in the Middle Ages than today; in about the 16th century in France, denasalization took place when the nasal consonant was intervocalic, and the n sound was retained—in, for example, French bon "good [masculine]" (pronounced *bõ*) and bonne "good [feminine]" (pronounced bon or [bon]). In Portuguese the consonant did not always reappear after denasalization (compare boa "good [feminine]," from *bõa*, from bona), though between i and a or o the palatal nasal consonant (close to ny in "canyon") is inserted (vinho "wine," from vio, from vinu).

Nasalization has sometimes, though without much conviction, been attributed to Celtic substratum influence. A better case can be made for the effect of such influence in the French *u* sound, [y], pronounced like German *ü* or Greek upsilon, though ignorance of Gaulish and certain chronological and geographical discrepancies make it difficult to argue in detail. The French *u* sound is also found in most Occitan dialects (in which it may be a recent introduction from French), in Rhaetian, and in parts of Portugal and Italy; elsewhere it is sometimes a characteristic of affected speech.

French uvular *r*

Consonants. Another French pronunciation that is often imitated by socially pretentious speakers is that of the Parisian uvular *r* (produced by vibration of the uvula, an appendage at the back of the mouth), which was not accepted in standard French until after the Revolution, though probably used by the Parisian bourgeoisie from the 17th century. It probably developed from the Latin double *-rr-*, differentiated from single *-r-*, which in Middle French tended to be pronounced with local friction, almost as a ***th*** or *z* sound (compare ***chaise*** "chair" and ***chaire*** "chair—throne, pulpit"). In most dialects of Provence today the distinction between the two *r* sounds is still made (though Occitanian dialects in general are adopting the French pronunciation). Brazilian Portuguese uses a similar contrasting pair of ***r*** sounds, with the usual trilled *r* represented in orthography by "*r*" and a velar, or "rough," *r* represented by "rr": Brazilian *caro* "dear" and ***carro*** "cart." Elsewhere only Puerto Rican Spanish and a few North Italian and Romanian dialects use the velar ***r*** regularly, though it is heard sporadically nearly everywhere.

One phonological development that is thought by many to be indicative of a very early split between the Eastern and Western Romance areas concerns the treatment of consonants between vowels. To the north and west of a line drawn between La Spezia and Rimini, in Italy, most dialects voiced Latin voiceless consonants between vowels and simplified geminates (doubled consonants); southern and eastern dialects to a greater extent retain the Latin voiced–voiceless–geminate system. The dividing line appears also to run through Sardinia, so that northern dialects are "Western" and southern ones "Eastern." Table 2 shows the treatment of intervocalic ***p*** and ***t***.

Table 2: Development of Latin Intervocalic p and t in Romance Languages

	ripa "bank"	***rota*** "wheet"
Vegliot	*raıpa*	—
Romanian	*rîpă*	***roata***
Italian	***ripa***	***ruota***
Logudorian	***rrba***	***roda***
Occitan	*riba*	***roda***
Catalan	***riba***	***roda***
Spanish	***riba***	***rueda***
Portuguese	***rrba***	***roda***
French	***rive***	***roue*** (Modern French) ***ruede*** (Old French)
Rhaetian	***riva***	***roda, ruede***

Some believe that the voicing of voiceless sounds is connected with a similar, though not identical, process known as lenition in Celtic. Lengthening and subsequent development into diphthongs of accented vowels may be linked to the reduction of Latin doubled consonants to single consonants, as some recent theories suggest.

One noticeable difference between Latin and all the Romance languages is that the consonantal systems of the latter include a number of palatal and palato-alveolar consonants, which did not exist in Latin. (Palatal consonants are formed with the tongue touching the hard palate; palato-alveolar sounds are made with the tongue touching the region of the alveolar ridge or the palate.) One consequence of the strengthening of the stress accent in the later Latin period was that unstressed *ĭ* and *ĕ* following consonants became shortened to a nonsyllabic palatal y sound (called ***jod***). The effects of this new sound on preceding consonants are varied, but in many cases these have been pronounced with the tongue raised more toward or against the roof of the mouth, or palate (a process classified under the general heading assimilation), sometimes ending up eventually as a dental fricative (such as ***z*** or ***th***) or affricate (such as ***ch***) and perhaps modifying the preceding vowel. That this process began early is suggested by the not-infrequent confusion of *-tĭ-* and *-cĭ-* in orthography, sometimes represented even as ***tz*** in inscriptions. This palatal shift in pronunciation led to developments such as French ***rouge***, Portuguese *ruivo*, Catalan ***roig***, and Italian *rosso* from Latin *rubeum* "red" and French *feuille*, Portuguese *folha*, Italian *foglia*, and Sardinian ***fodza*** from Latin ***folia*** "leaf."

Shifts involving palatal consonants

Another source of palatal consonants in Romance has been back (velar) consonants when immediately followed by a front sound: the velar consonant has often moved forward in the mouth, sometimes eventually to dental or alveolar position but often settling on a palatal or palato-alveolar position. This process, too, probably began early, first affecting velar consonants *k* and *g* preceding front vowels ***e*** and ***i***. That it had not occurred at the classical period is shown by its absence in early loanwords into other languages (Berber, Basque, Celtic, Germanic, Albanian, and Greek). As central Sardinian dialects retain velar pronunciation in the environment of front vowels, it may be assumed that palatalization postdated the separation of the island from the rest of the empire. Vegliot evidence is difficult to interpret, as *ē* does not seem to have provoked palatalization, whereas *ĕ*, *ĭ*, and ***u*** did so. It was this sound change that resulted in the pronunciation of "soft" ***c*** before ***e*** and ***i*** (in most Romance languages this is an ***s*** or ***ts*** sound; in Italian and Rhaetian it is a ***ch*** sound). Before ***a***, ***o***, and ***u*** the ***c*** retained its "hard" pronunciation (that is, a ***k*** sound). In Classical Latin, before the sound change occurred, all ***c*** sounds were "hard." Hence, Latin ***centum*** ("kentum") gave rise to Italian ***cento*** ("chento"), Portuguese ***cento*** ("sento"), and Spanish ***ciento*** ("siento" or, in Castilian, "thiento").

In north central France, Latin ***a*** must have advanced to a front position, with the result that it, too, palatalized preceding ***k*** and ***g*** sounds. The results give the palato-alveolar sounds of ***sh*** and *zh* (written in the International Phonetic Alphabet as [ʃ] and [ʒ], respectively), via [tʃ], the ***ch*** sound in "church," and [dʒ], the *j* sound in "jam"; *e.g.*, French ***chanter*** "to sing" developed from Latin *cantare*, ***joie*** "joy" from *gaudia*. West Rhaetian dialects show a similar development (compare Sursilvan *tgaun*, Engadine ***chaun***, French *chien*, from Latin *canem* "dog"), as do Franco-Provençal and Northern Occitan dialects, but Picard and some Norman dialects do not (Picard ***canter***, with an unpalatalized ***c***, from Latin *cantare*; ***kier*** "dear," from *carum*). The change is assumed to have taken place at a later period than the palatalization of k when followed by ***e*** or i, which did not affect Frankish words. These, on the other hand, succumbed to the type of palatalization in which k changed to ***ch*** [tʃ] and then to ***sh*** [ʃ] (**skina* > *échine* "backbone").

Romanian consonant developments

In Romanian, velar consonants were moved forward under the influence of a following i and ***e***, and dental consonants were moved back to a palatal position under the same influence; *e.g.*, *ţară* from *terram* "earth"; ***şi*** "and" from ***sic*** "thus." Labial consonants are also affected in some dialects: ***k'ept*** from ***piept*** from ***pectum*** "chest"; *jin* from *vin* from *vinum* "wine." Romanian also has, in final position, a series of "soft" consonants, reminiscent of the Slavic sounds. These are transparently derived from earlier "hard" consonants followed by *i*, performing certain important morphological functions: *lupi* [lup'] "wolves" / *lup* [lup] "wolf"; *cînţi* [kints'] "thou singest" / *cînt* [kint] "I sing."

Palatalization of consonants in Romance was effected not only by following front vowels but also by juxtaposed front consonants, especially when a velar (such as Latin ***c*** or ***g***) was next to a dental (such as ***t***, ***s***, ***n***) or a lateral (***1*** sound) in medial position, sometimes as a consequence of the loss of an unaccented vowel during the Vulgar Latin period. Results of this process vary from language to language. Table **3** gives examples of these changes.

It will be noted that in Romanian a labial consonant has been substituted for the velar in the Latin clusters *-ct-*, *-x-* [ks], and *-gn-*. Perhaps there was first assimilation of the velar to the dental—as in Italian *-tt-* from Latin *-ct-* and

Table 3: Results of Palatalization of Consonant Clusters

	noctem "night"	*coxam* "hip"	*piscem* "fish"	*pugnum* "fist"	*oc'lum* "eye"
Venliot	*nwaf*	—	*pask*	—	*vaklu*
Romanian	*noapte*	*coapsă*	*peşte*	*pumn*	*ochi*
Sardinian	*notte*	*koša*	*piske*	*pundzu*	*okru*
Italian	*notte*	*coscia*	*pesce*	*pugno*	*occhio*
Occitan	*nôit, nuech*	*cuoissa*	*peis*	*ponh*	*uelh*
Catalan	*nit*	*cuixa*	*peix*	*puny*	*ull*
Spanish	*noche*	*cojo*	*pez*	*puño*	*ojo*
Portuguese	*noite*	*coxa*	*peix*	*punho*	*olho*
Rhaetian					
Sursilvan	*notg*	*queissa*	*pesch*	*pugn*	*egl*
Engadine	*not*	—	—	*puoñ*	—
Friulian	*ñot*	—	*pes*	—	—
French	*nuit*	*cuisse*	*(poisson)*	*poing*	*oeil*

Sardinian *-nn-* from Latin *-gn-* (***linna*** from ***Iigna*** "line")—followed by differentiation of the first element of the geminate. It is notable that Latin *1* regularly becomes *jod* after another consonant in Italian (*piacere* from *placere* "to please"; *fiore* from ***flore*** "flower"; ***chiave*** from ***clave*** "key"; ***ghianda*** from *glanda* "acorn") and after velars in Romanian (*plăcea, floare,* but **cheie** [kjej], *ghindă* [gjinda]). In Spanish and Portuguese a following *l* in Latin often palatalizes labial consonants ***(p, f)*** as well as velars, in initial as well as medial position; *e.g.*, Latin ***planum*** becomes Spanish ***llano*** "plain," Portuguese *chão;* Latin *afflare* becomes Spanish *hallar* "to find," Portuguese ***achar.***

GRAMMAR

Item for item, the Romance languages all appear grammatically close to Latin and to each other: superficial resemblances in individual expressions may, however, mask differences of content and construction that are difficult to describe. The most obvious difference between Latin and Romance is in the comparative autonomy of morphemic units, especially words. In Romance, Latin inflectional endings have been much reduced, and more reliance is placed on syntactic construction to convey sentence meaning; that is, Romance languages are more "analytic" than the predominantly "synthetic" Latin. A corollary of this is that word order is less flexible in Romance, as it has become the principal means of showing relationship between words in the sentence.

Reduction of inflectional endings

Forms of nouns and adjectives. The inflectional endings have been lost most in nouns and adjectives. The Classical Latin five-case declensional system has everywhere been replaced (with a couple of doubtful exceptions) by a two-gender system, in which normally masculine gender is marked by survivors of the second ***(-us)*** declension endings of Latin (Italian ***cavallo,*** Portuguese ***cavalu,*** Romanian *calul,* Sardinian ***kaddu,*** Rhaetian ***cavagl,*** from Latin *caballus* "horse"), and feminine is marked by first ***(-a)*** declension endings (Italian ***capra,*** Spanish ***cabra,*** Rhaetian ***caura,*** Romanian *capră,* from Latin ***capra*** "goat"). Cognates of third-declension Latin noun forms are incorporated into the same system, but their gender is marked by changes in the article or accompanying adjective (agreement or accord) rather than by overt markers in the word itself (for example, masculine Italian ***il monte,*** Catalan ***es mumt,*** from Latin ***mons, montem*** "mountain"; feminine Italian ***la notte,*** Catalan ***sa nit,*** from Latin ***nox, noctem*** "night"). In modern French, although gender is marked in the written language, however inconsistently, by the presence or absence of final ***-e,*** any overt morphological markers the spoken language may have are more complex in character, and more reliance is placed on syntactic agreement; thus, ***chatte*** "she-cat" is distinguished from ***chat*** "cat" by the presence or absence of the final consonant sound *-t* in pronunciation, but ***(le) tour*** "tour, trick" and ***(la) tour*** "tower" have identical phonetic shapes though they belong to different gender classes.

All the Romance languages continue to mark plurality in nouns and adjectives morphologically, though in modern spoken French this is not done consistently. In Western Romance the sign of the plural is usually *-s,* derived from the Latin accusative plural flection: Spanish ***caballos, cabras, montes;*** Occitan ***cavals, cabras,*** *mons;* Catalan ***cavalls, cabres,*** *muntes;* Sardinian ***kaddos, krabas, montes;*** Old French ***chevals, chkvres, monts.*** In Italian and Romanian, however, plurality is shown by a final ***-i*** (which in Romanian "softens" the preceding consonant) or, in the case of some feminine nouns, by a final ***-e:*** Romanian ***cai, capre, munţi, nopţi;*** Italian ***cavalli, capre, monti, notti.*** These endings may derive from Latin nominative plural first- and second-declension endings ***-ae*** and -**i**, or they may represent a somewhat irregular development of the *-s,* favoured elsewhere.

The Latin nominal case system has disappeared in all modern languages except Romanian, in which the inflected article distinguishes the nominative and accusative from the genitive and dative (see Table 4). Thus, when other Romance languages would use a preposition to indicate a certain relationship between words, Romanian resembles Latin in using an inflected form (*e.g.*, Latin ***matris*** "the mother's: Romanian ***mamei,*** French ***de la mkre,*** Italian *della madre*).

Loss of case system

Table 4: Declensional System of Romanian

	singular	plural
Masculine "son"		
Nominative–accusative	*un fiu, fiul*	*fii, fiii*
Genitive–dative	*unui fiu, fiului*	*unor fii, fiilor*
Feminine "mother"		
Nominative–accusative	*o mamă, mama*	*mame, mamele*
Genitive–dative	*unei mame, mamei*	*unor mame, mamelor*

In Old French and Old Provençal some remnants of a case system remained, in that the masculine nominative (subject of the verb) was distinguished from the other cases (collectively called oblique). Today such grammatical information is conveyed by word order in most Romance languages, as in English, with the subject normally preceding the verb: French ***Pierre appelle Paul*** "Peter calls Paul"; Portuguese ***Pedro chama Paulo;*** Italian ***Piero*** *chiama* ***Paulo.*** Some Romance languages pick out the object of the verb, if it is a person, by an additional particle: Spanish ***Pedro llama a Pablo;*** Romanian ***Petru cheamii pe Pavel.*** Several Italian dialects, as well as Sardinian and occasionally Engadine and Portuguese dialects, have similar constructions: Calabrian ***Chiamu a Petru*** "I call Peter"; Elba *Ò visto* ***a*** *ttuo* ***babbo*** "I saw your grandpa"; Engadine ***Amk a vos inimihs*** "Love your enemies." It is notable that the Italian-based lingua franca used by Mediterranean sailors since the 16th century also picks out the personal object (*e.g.,* ***Mi mirato per ti*** "*I* saw you").

The definite and indefinite articles were unknown in Latin but developed everywhere in Romance, usually from the Latin demonstrative *ille* "that" (though in a few parts from reflexive ***ipse*** "himself") and the numeral *unus* "one." The articles seem to have played some part, during the older stages of the languages, in distinguishing subject from object; the article is more often used where a Latin nominative would have occurred than in other cases, perhaps to give prominence to the topic of the sentence. Today the use of the article has so extended that such distinction is no longer possible; in French, for instance, a common noun is always accompanied by a determiner such as an article, demonstrative, or possessive, so forms remaining from the earlier stage, such as ***avoir faim*** "to be hungry," are often regarded as idiomatic and inexplicable in terms of modern structure.

The system of verbs. In the passage from Latin to Romance, verbal inflection has survived much more than noun declension. Although the four regular Latin conjugations have been virtually reduced to two, with only the ***-a-*** class remaining truly productive, other features of the verb seem almost unchanged. In most languages, for instance, the person markers are directly traceable to Latin origins (*i.e.,* to Latin ***-6, -s, -t, -mus, -tis, -nt***). Modern spoken French is the only major language in which the personal endings no longer serve the same function as Latin. Today, person is marked in French principally by pronouns derived mainly from the Latin emphatic nominative forms of the personal pronoun: *J'aime* [ʒɛm] "I love," *tu aimes* [tyɛm] "you love" from *(ego) amo, (tu) amas*. The creoles have taken this process even further, in that their verb forms are usually invariable but are prefixed by elements indicating person, tense, aspect, etc., as in many West African languages: Louisiana French [motegt] "I was having" from *mon* [mo] *étais* [te] *gagner* [gẽ]; and similarly [ilagt] "he will have."

Verb conjugations

In the metropolitan languages, verbal modalities are shown, as in Latin, by inflection. Some Latin verb endings, such as that of the *-r* passive or of the future, have disappeared; others, such as the pluperfect indicative and subjunctive, have survived in a few languages with modified function. But most languages today have reflexes of the present, perfect, and imperfect indicatives and of one or more subjunctive tenses. The imperfect indicative, a Latin innovation, survives almost intact, though the evolution of its form, not to mention its function, presents problems. The -ī- stem form in Latin *-iēba-* is thought to have coalesced early with the *-P-* stem *-ēba-* form, but a few languages (notably Italian, Friulian, and some Spanish and Portuguese dialects) today have reflexes of an *-ība-* form that might have survived from popular Latin. The Latin *-āba-* form survives almost everywhere, though in most French dialects its older reflexes, *-eve* and *-oue*, have been replaced in modern times by forms derived from Latin *-ēba-*. These latter are thought to be widespread but are puzzling phonologically as they have very often irregularly lost their *-b-* (Spanish, Portuguese, etc., *-ía*, French *-ais*).

The Latin perfect of the type *amdvit* "he has loved" is known by all the literary languages but is rare in speech in French, Italian, and Romanian, in which it has been replaced by a new compound past made up of the verb for "to have" and a past participle. The latter structure is known to some extent in all Romance languages, often being used to express a more recent past than the preterite *amāvit* form, which also indicates action in the past (without reference to duration or repetition): Romanian *au cîntat*, Italian *ho cantato*, French *j'ai chantk,* Spanish *he cantado*, Old Portuguese *hei cantado*, Engadine *ha chantd, hè chantò*, Sardinian *kantau appo,* from Latin *habeo cantatum "I* have sung." In Modern Portuguese the preferred auxiliary is *ter* "to have, to hold" rather than *haver,* producing forms such as *tenho cantado*, while modem Catalan more commonly uses the verb for "to go" plus the infinitive, giving *vaig cantar* rather than the pan-Romance type *he' cantat*.

Representation of the future tense

The disappearance of the Latin future has been remedied in most Romance languages by the development of new forms of periphrastic origin. Many of these forms use some reflex of *habēre* "to have" joined to an infinitive. From Latin *cantāre habēo* "'I will sing" are derived Italian *canter&,* Spanish, Catalan *cantare'*, Portuguese *cantarei*, French *je chanterai*, Rhaetian *c(h)antero, c(h)antera*, Occitan *cantarai; habēo cantāre* gives southern Italian *aggio cantd* (similar forms are seen in earlier Spanish, Portuguese, and northern Italian). Latin *habēo ad cantāre* produces Sardinian *ap a kantare*, and *habēo de cantare* gives Portuguese *hei-de cantar* (more popular than *cantarei)*.

A periphrastic future of the type known in English "I'm going to sing" enjoys popularity in Romance, mainly to indicate a less distant future event than the more formal future tense (*e.g.*, French *je vais chanter,* Spanish *voy a cantar*). Other periphrases used in Romance are "I will (wish to) sing," as in Romanian *voi cinta;* "I must sing," as in Sardinian *deppo kantare;* "I'm coming to sing," Sursilvan *jeu vegnel a cantar;* and *"I* have that I should sing," as in popular Romanian *am sa cînt*. Notably, Dalmatian does not seem to know periphrastic Romance futures but uses a form *kantuora* (perhaps from Latin *cantāverō*) as both future and conditional.

The Romance conditional, or "future in the past," a form not found in Latin, is in many languages related to the new future. In the Western languages it is composed of the future stem (or infinitive) plus a past-tense marker related to reflexes of *habēre*. In some cases an imperfect form is used, in others a perfect form; examples are French *je chanterais* "I would sing," Spanish, Portuguese, Occitan, and Catalan *cantaria,* and Italian *canterei, -ebbe*, etc. In Romanian the conditional marker can either precede or follow the infinitive and may be derived from the imperfect of *vrea* "to wish": for example, *aşi cinta, ar cînta*, etc., or (more literary) *cîntare-aş, cîntare-ar*, etc.

Word order

Word order is the means most used by modern Romance languages to show the grammatical relationship between words; statistically the most frequent order in statements is subject–verb–noun object. In many of the Romance languages, interrogation can be shown by inversion of the subject and verb, placing the verb, as the element on which the interrogation falls, at the beginning of the sentence (Spanish *¿Vino el hombre?,* Italian *È venuto l'uomo?* "Has the man come?"). In such examples, however, it is the intonation (represented in writing by the question mark) rather than the word order alone that marks the question. Inversion, without interrogative intonation, is not infrequent in emphatic assertions. Unambiguous question markers—such as the Latin particles *-ne, nonne*, and *num*—are lacking in most Romance standards; popular speech, though relying everywhere principally upon intonation, often has developed new particles to reinforce interrogation. Romanian has *oare, şi (Oare a venit?* "Has he come?" *Si te ai culcat?* "Have you been to bed?"); Italian uses dialectal *ce, che,* or *o* (Vulgar Tuscan *Che è venuto?* "Has he come?"; *O come si chiame?* "What is he called?"); Sardinian has *a (A mosse kkŭstu kăne?* "Does this dog bite?'); and French and Limousine have ti (generalized from such forms as *a-t-il?;* French *Je suis-ti bête?* Limousine *Sieu-ti nesci?* "Am *I* stupid?"). In modern standard French great use is made of *est-ce que* as an interrogative particle: *Est-ce qu'il est venu?* "Has he come?" *Comment est-ce qu'il s'appelle?* "What is his name?"

Negation in Latin was expressed by a range of special items *(non, nemo, nihil, nullus, nunquam,* etc.). Although some of the others survive in Romance, continuators of *non* have taken over the main burden of negative expression and are regularly prefixed to the verb. Nuances within negation are usually expressed by the adjunction of other items. In France, both north and south, and in northern Italy and some of the Swiss Rhaetian areas, the *non* particle has been so weakened phonetically that it no longer can express unambiguously the important distinction between negative and positive; hence, formerly positive adjuncts have acquired its negative meaning.

French *personne / une personne* signifies "no one / a person"; *pas / un pas* means "not / a step"; and *plus* can mean "more / no more." In popular speech the *non* particle is frequently omitted altogether in areas that use these additional forms (*e.g.*, French *Je (ne) le vois pas;* Occitan *Lou vese pas* for *Noun lou vese* "I don't see it").

Reduction of the subjunctive

Morphologically, the verb system survived comparatively intact from Latin to Romance; if the schoolbooks, heavily influenced by Latin grammar, are right, the ways in which the verb forms are used are not so very different from Latin either. The most obvious change has been the reduction of uses as well as of forms of the subjunctive, with, at the extreme, modern French treating them as automatically determined variants to be used obligatorily after certain phrases and conjunctions and virtually eliminating tense differences within the subjunctive mood. When the subjunctive retains a function in Romance—that is, in contexts in which it can contrast with the indicative—it has developed emotive overtones, especially suggesting doubt, unreality, or some sort of hypothetical futurity. It is used especially in subordinate clauses dependent on verbal expressions of command and exhortation, emotion, or doubt: Romanian *voi sa vină* "I want him to come"; Engadine *Mieu bap voul ch'eau lavura* "My father wants me to work"; French *Je doute qu'il vienne* "I doubt that he's coming"; Portuguese *Duvido que seja feliz* "I doubt that he is happy"; Italian *Temo che sia tarde* "I'm afraid it's late"; Spanish *Temo que él lo diga* "I'm afraid he'll say it." The subjunctive also regularly follows subordinating conjunctions that project action forward into the future, such as “until,” “before,” "in order that": French *avant que vous soyez venu* "before you came"; Spanish *hasta que sea feliz* "until he is happy"; Italian *perch2 potessi fare in tempo* "so that I might do it in time"; Portuguese *antes que eu o veja* "before I see it"; Catalan *abans que vingui* "before he comes."

On the whole, however, the Romance languages use the subjunctive less than Latin, with recession particularly, when no doubt is implied, in indirect speech and in temporal and concessive clauses (in French, use of the subjunctive after concessive conjunctions such as *bien que* and *quoique* "although" was imposed by 18th-century grammarians). The infinitive is often used in subordinate constructions when Latin would have used a subjunctive; *e.g.*, French *dites-lui de s'en aller*, for *dites-lui qu'il s'en aille* "tell him to go away." Romanian, on the other hand, has even extended the use of the subjunctive in such constructions, perhaps reflecting a substratum influence that is felt, too, in other Balkan languages. Greek influence is sometimes credited with similar constructions (usually using the indicative rather than the subjunctive) found in northeast Sicily, northern Calabria, and the Salentine Peninsula.

Conditional clauses

One area of syntax in which the Romance languages vary widely in the extent to which they retain and in the manner in which they replace the Latin subjunctive is that of past-tense hypothetical conditional clauses. The Latin formula *si habuissem dedissem* "if I had had it, I would have given it," though challenged by a type using the indicative tense since Ciceronian times, has sporadically survived into Romance, especially in the older stages of the languages and in scattered parts of southern Italy *(Se potessi, facessi* "If I could, I would do it"), Rhaetian (Sursilvan *Jeu vegness, sche jeu vess peda* "I'd come, if I had time") and Romanian *('dacă agi avea destui bani, agi cumpăra-o* "If I had enough money, I'd buy it").

In most languages, however, a new conditional form replaces the subjunctive in "if" clauses. Thus, in Spanish, Portuguese, and most Italian dialects, sentences of this type are seen: Spanish *si* yo *tuviese bastante dinero, lo compraria;* Italian *se avesse abbastanza danaro, lo comprerei;* Portuguese *se tivesse bastante dinheiro compraríao* (“if I had enough money, I'd buy it"). Spoken Catalan usually prefers a similar construction *(si estudiessis ho sabries* "if you studied, you would know it"). Another construction that replaces the subjunctive by the imperfect indicative in the "if" clause is, however, considered more correct in Catalan and is normal in French as well as in Corsica and Sardinia: Catalan *si estudiaves ho sabries* ("if you studied, you would know"); French *si j'avais assez d'argent, je l'achèterais* ("if I had enough money, I'd buy it"); Logudorian *si denia abba deo dia buffare* ("if I had water, I'd drink”). Other constructions using the imperfect indicative or the conditional in both clauses are found mainly in substandard styles—both types are common in French, the former in Tuscany, southeastern Italy, and Spain and the latter in much of southern Italy.

Forming new words

Word formation. Romance methods of forming new words from native sources are in part inherited from Latin (the morphological device of adding a suffix and that of prefixing an element that modifies the original meaning) and in part later developments (mainly that of combining two or more free forms to make compound words and of changing or extending the syntactic distribution of an already existing word).

Derivation by means of suffixes is the most popular and widespread device; verbs in particular must be morphologically marked as members of a conjugation, of which those corresponding to Latin *-āre* form by far the most frequent and indeed in modern times virtually the only productive class (thus Latin *plantāre* "to plant," Italian *plantare,* Engadine *plaunter*, French *planter,* Catalan *plantar,* from *planta* "plant"). Infixes, inserted between the verbal root and the conjugation marker, are common. Sometimes they continue Latin infixes, such as the frequentative (compare *jactāre* for *jacere* "to throw," Italian *gettare*, French *jeter*, Catalan *getar*, etc.); sometimes they add semantically to the root meaning (compare pejorative Italian *lavoracchiare* “to slack off” from *lavorare* "to work," French *criailler* "to bawl" from *crier* “to cry"). The Greek verbal infix *-iz* (as in English “ize”) is particularly popular in Romance today *(e.g., latinisare, automatiser)*.

Among noun suffixes, diminutives are frequent and, except perhaps in French, still productive. Romanian uses *-aş (degetaş* "little finger", but the other languages prefer derivatives of Latin *-ittus* (especially in Spanish: *arbolito* "little tree," *señorita* “Miss, young lady," etc.; but also French *sachet* "little sack," Italian *foglietta* “little leaf," etc.) or of Latin *-īnus* (preferred in Italian: *tavolina* "little table, desk," *signorina* "young lady"; and Portuguese: *copinho* "little drinking glass," *senhorinha* "young lady"). The Latin *-ōne* suffix has, conversely, acquired augmentative meaning in several languages (Romanian *căloiu*, Italian *cavallone,* Spanish *caballón* "large horse").

Other frequent suffixes sometimes have a "learned" modern form alongside the older "popular" one; *e.g.*, Latin *-atione:* Italian *-agione / -azione*, French *-aison / -ation*, Spanish *-azón / ación*, Portuguese *-azão / ação;* also Romanian *-ăciune*, and Occitan *azó*.

Suffixes that remain extremely productive include the Latin verbal adjectival *-bilis* (not found in Romanian): Italian *bastevole* "enough," French *admirable,* Spanish *ainable* "pleasing"; and verbal nominal *-mentum:* French *abonnement* "subscription," Spanish *cobijamiento* "lodging," Italian *abboccamento* "interview, parley," Romanian *acoperămînt* "cover."

Prefixing of modifying elements remains frequent in all languages (Italian *autostrada* “highway,” Spanish *contraveneno* "antidote," French *photocopie* "photocopy"), although some older prefixes may hardly be recognized as such today. The "repetitive" verbal prefix *re-* remains particularly active (Romanian *răsări* "rebound," Italian *ricattare* "to recover," French *racheter* "to buy back." etc.).

Compound words, though less frequent than in the Germanic languages, are not uncommon (*e.g.*, French *cheflieu* "principal town," Italian *primavera* "spring," Spanish *lavamanos* "wash basin").

Adverb formation

Originally a compounding process, the most common method of forming adverbs from adjectives (suffixing of Latin *mente* "mind") has become in most languages a morphological process, although Spanish and Portuguese retain traces of the earlier stage in phrases such as *severa e* (y) *cruelmente* "severely and cruelly."

Among the syntactic means that most Romance languages use to extend vocabulary is the potent device, unavailable to Latin, of juxtaposing to any part of speech an article or other determiner and using it as a noun (*e.g.*, Italian *il perch2* "the reason," Spanish *lo útil* "utility, something useful," French *un je ne sais quoi* "an I-don't-know-what"). In French and Spanish, verbal infinitives are frequently so treated (*le devoir* "duty," *el poder* “power,” etc.); Romanian also uses infinitives as verbal nouns, but they are differentiated formally by retaining the full form (*e.g.*, *cintare* "singing"), compared with the shortened verbal form *(cinta)*. In earlier stages of most Romance languages the verbal root (most often as it appears in the 3rd-person singular present indicative) could be used as a noun, a process known as back-formation (compare Romanian *laudă* "praise," Italian *domanda* "question," French *approche* "approach," *désir* "desire," Spanish *baila* "dance," Portuguese *muda* "change").

Just as former adjectival forms are frequently used as substantives, so are nouns used with adjectival function; there seem to be few restrictions on this use, though practice varies as to whether agreement should be made (French *les frères ennemis* "enemy brothers," with agreement; *une femme médecin* "a woman doctor," without agreement). Past-participial forms normally act as adjectives, as in English.

Romance makes use of gender classification to extend and modify its vocabulary, especially by relating the gender markers to sex differences (*e.g.*, Romanian *nepot, nepoată;* Occitan *nebut, nebudo,* Spanish *nieto, nieta,* Portuguese *neto, neta*, Catalan net, *neta* “nephew, niece," with Italian invariable *nipote*, and French lexically differentiated *neveu, nièce*). Modern French makes particularly fruitful use of gender differences (originally via ellipse); thus, *le* (vin de) *champagne* (the drink) / la *Champagne;* La Nor*mandie* (the province) / *le Normandie* (the ship).

VOCABULARY

Importance of Latin in vocabulary formation

The basic vocabularies (the most frequently used lexical items) of all the Romance languages are in the main directly inherited from Latin. This applies equally to "function" words, such as *de* "of, from" (Romanian de, Italian di, Rhaetian *da*, French de, Spanish *de*, Portuguese *de*), as to common lexical items, such as *facere* "to do" or aqua "water" (Romanian face, *apă*, Italian fare, *acqua,* Logudorian *fágere,* abba, Engadine fer, ova, French *faire,* eau, Catalan *fer,* aygua, Spanish *hacer*, agua, Portuguese fazer, *água*). In some cases different Romance languages inherit words perhaps from different strata of Roman society. Thus, for "lamb," forms derived from Latin *agnus* remain in southern Italy and Galician (*año*), but forms derived from diminutive *agnellus* prevail in Romanian (*miel*), Italian (*agnello*), French (*ag*ncau), Rhaetian (Engadine *agné*, Friulian *añel*), Occitan (*anhel*), and Catalan (anyell), with Sardinian and some Calabrian dialects using another form derived from Latin agnone (Logudorian *andzone*). Spanish and Portuguese, however, prefer a derivative of a different word, *chorda* (cordero, *cordeiro*), referring perhaps to the birth process; this word is also found in Occitan and Catalan.

Some words shared by the majority of the Romance languages are not of Latin origin but were probably borrowed from other languages before Latin unity was disrupted. These include especially words of Celtic origin, such as Latin *carrum* "cart," Romanian car, Italian carro, Logudorian karru, Rhaetian k'ar, French *char*, Occitan and Catalan cur, Spanish and Portuguese carro.

In Christian Latin a great many Greek ecclesiastical terms were borrowed, which survived in most Romance languages. For example, the Greek word *episkopos* (literally, "overseer") was borrowed into Latin as *episcopus* "bishop," which gave rise to Vegliot *pasku*, Logudorian *pískamu*, Italian vescovo, Engadine *ovaisch*, Friulian *veskul*, French *evêque*, Occitan avesque, Catalan bisbe, Spanish obispo, and Portuguese bispo.

Germanic words did not penetrate into Latin very frequently before the separation of the various Romance languages from Latin, so that few of them have more than limited extension. Only one Germanic word is known for certain to be found in both Eastern and Western Romance—*sapōne* "soap," recorded in Pliny and occurring as Romanian *săpun*, Vegliot *sapaun*, Italian *sapone*, Logudorian *sabone*, Engadine *savum*, French *savon,* Occitan and Catalan *sabó,* Spanish *jabón*, and Portuguese *sabão.*

"Learned" words

Many Latin words are widespread throughout the Romance languages even though they do not date back directly to the imperial period; these are the “learned” words that have freely entered the languages at virtually every period, borrowed from Latin used as a scholarly language. Because of this later borrowing, such words as capital, *natura*, *adulterium*, and discipulus appear in Romance virtually unchanged from Latin, as they do in other European languages; Romance Latinisms, however, are quite normally used in contexts in which similar words would sound stilted and pedantic in English (*e.g.*, French *supprimer* "suppress" but often used to mean "to do away with").

However similar the Romance vocabularies are to each other, considerable differences nevertheless exist. Some of these may be traced back to imperial times, when provinces may have developed their own vocabulary preferences. For instance, for "oak" Eastern Romance seems to have preferred Latin quercus (Logudorian kerb ~sọuthern Italian *quercia,* etc.), whereas the West preferred the alternative *robur* (Italian *rovere*, Occitan and Catalan *roure*, Spanish and Portuguese roble, Old French louvre—modern French *chêne* is of Celtic origin, while Romanian *stejar* is perhaps of Balkan origin). In some cases the conservative peripheral areas have retained a word that was displaced in more central regions; thus, for "beautiful," *formosus* is preferred in Romanian (*frumos*), Spanish (*hermoso*), and Portuguese (*formoso*), whereas *bellus* is more popular in Vegliot (bial), Italian (*bello*), Rhaetian (*bal*, *biel*), French (beau), Occitan (bel), and Catalan (bell).

Borrowing from substratum and super-stratum languages

When Romance borrowed vocabulary from the substratum, differentiation must have taken place early (certainly before the indigenous languages died out). Thus, Spanish *vega*, Portuguese *veiga* "wooded ground by a river" (probably from a non-Indo-European Iberian language, compare Basque ibaiko “riverbank”), French *charrue* "plow," borne "boundary stone" from Celtic, and Romanian *barză* "stork" (perhaps from Dacian, compare Albanian bar) probably were used during Roman times in some form. The debt of Romance vocabulary to substrata languages is probably great but difficult to estimate with any certainty. When there is no known source form or cognate for a word, scholars often suggest an Iberian, Dacian, Ligurian, or Gaulish origin, but, as little is known of these languages, some such theories are mere speculation.

After the influx of barbarian invaders, Romance vocabularies differentiated further as each borrowed from its own superstratum (language superimposed upon Romance). French, for instance, is estimated to have taken some 700 words from Frankish (a Germanic language), not all of which have survived but some of which have passed via French into other Romance languages. Many of these were concerned with agriculture (*jardin* "garden," *houe* “hoe,” ble' "wheat," *gerbe* "sheaf," etc.) or with war (*guerre* "war," *héaume* "helmet") or social organization (*sénéchal* "seneschal," chambellan "chamberlain," *maréchal* "marshal," *baron* "baron"). The occupation of much of north Italy by speakers of Langobardic (also a Germanic language) left less of a mark on Italian vocabulary, though dialects retain more words (estimated at about 300) than the standard language. Standard Italian borrowed little in the way of administrative or military terms but accepted a number of words from rural life (*melma* "mud," *zecca* "sheep tick," *stamberga* "hut," etc.). The Visigoths, who occupied Iberia, were more Romanized than the other Germanic invaders and indeed had abandoned their Germanic tongue by the 7th century AD. Thus, borrowings from Visigothic into Spanish and Portuguese are less frequent, though still not inconsiderable; some (such as estaca "stake," brotar "to bud") are common to all the Iberian Peninsular languages.

Slavic infiltration into the Balkans led Romanian to adopt a very large number of Slavic words, some in the basic part of the vocabulary. At exactly what stage in history they were borrowed is uncertain, for the earliest Romanian texts: of the 16th century AD, are saturated with Slavic terms from different dialectal sources, though South Slavic predominates. Possibly the borrowings occurred in the 9th century, when the Hunnish Bulgarians, who had adopted Slavic speech, established a powerful state and embraced Christianity, and Slavic pressures were already very strong. Among common Romanian words of Slavic origin one may mention a *trăi* "to live," *hrană* "food," *ceas* “hour,” *bogat* "rich," *prieten* "friend," *munci* “to work." The Magyars (modern Hungarians) also lent a smaller number of words to their Romanian neighbours (*e.g.*, *oraş* "town").

Words borrowed from Arabic

Islāmic invaders into Europe from the 8th century had considerable effect on the vocabulary of the Western Romance languages, even though occupation was confined to southern regions. With its superior cultural and agricultural skills, the Arab world had much to teach Europe of the Dark Ages. Words entered via two routes: Sicily and Spain, and usually their form gives clues about their provenance—if the Arabic definite article (al) has coalesced with the root, the word is from Moorish Spain (thus Spanish *algodón* "cotton," Portuguese *algodão,* Old French *auqueton* via Spain, but Italian *cotone*, French *coton* via Sicily). The Arabs introduced into Europe many exotic plants and fruits and with them their names, such as oranges (Spanish *naranja*), lemons (Spanish *limón*), and artichokes (Spanish alcachofa, Italian *carciofo*). In some cases the Iberian Peninsula has adopted the Arabic word for such plants, while other languages prefer words of other origin—“rice” is arroz in Spanish and Portuguese, *arròs* in Catalan, but Italian and French prefer a Greek word (riso, *riz*), as do Vegliot (rize), Rhaetian (Friulian *ris*), and Romanian (orez). Apart from the numerous Arabic words known throughout Romance (especially "arithmetic," "algebra," and the like), many are peculiar to the Hispanic languages, such as administrative terms such as Spanish alcalde "mayor" or *alguacil* "senior police officer," commercial terms such as *almacén* "warehouse, department store," as well as everyday words such as ahorrar "to save," alboroto "noise."

Many of the words individual languages borrowed from other sources or fashioned themselves from native sources did not remain private property for long. Interchange among the Western languages has been common since the earliest times and especially from the 16th century. Perhaps French has been the greatest supplier of words throughout the ages, often displacing native words. But French, too, has borrowed heavily from the other languages, especially when they have been purveyors of new objects (such as *patate*, *banane*, tabac, introduced into Europe by Spanish and Portuguese explorations) or of special cultural values (Italian musical and architectural terms, as well as words to do with banking). Borrowing into minor languages from prestigious neighbours has, naturally, been prolific. Passage of words in the other direction is rare and usually employed for comic or other emotive effect (though Occitan in its heyday supplied a good many words of all sorts—even, it is said, amour "love" to French).

Anglicisms

Borrowings from non-Romance languages are less frequent and often frowned on by purists, but far from negligible. Any contact in specialized spheres has produced a crop of loanwords, especially since the 17th century, when French in particular began to borrow a fair number from its Germanic neighbours. In recent times, the influx of anglicisms has become a flood, resisted to the death by some purists. Many are, however, ephemeral or specialized, and none affects the basic vocabulary in which Latin-inherited words continue to predominate.

ORTHOGRAPHY

Today the Romance languages are all written in the Latin alphabet, with certain modifications, though until the mid-19th century Romanian was normally written in Cyrillic (still used in Moldavia), and, in the Middle Ages, Arabic script was used for some Spanish dialects.

As soon as scribes first made attempts to write in vernacular Romance, they found the resources of the Latin alphabet inadequate to represent the non-Latin sounds of their spoken language. One device used to overcome these difficulties was to add the letter *h* to another, to indicate a deviant pronunciation: thus, *ch* might represent the ch sound in Spanish (*e.g.*, *muchacho* "boy") or the *sh* (earlier ch) sound in French (*e.g.*, chef "chief"). *C* would normally be used for the k sound (before a, o, *u*) or an *s* or th sound (before *i, e*). In Italian, conversely, *ch* serves to distinguish the k sound, followed by e, from the ch sound (compare *che* [key] “that, who" with *c'è* [chey] "there is"). H was also sometimes added to *n* and *l* to indicate a palatal pronunciation (similar to the ny in English "canyon" and *li* in "scallion"), as today in Portuguese *vinho* "wine" and *filho* "son." Another device frequently used to stretch the capacity of the Latin alphabet was to distinguish the letters i and *j*, u and v, which were originally each single letters i (with variant form j) and v (with variant form *u*, and in Latin pronounced u or w). In Romance, v and *j* came to represent consonants, while u and i retained their vowel values.

The palatal consonants *n* and 1 are also often depicted by doubled letters or other combinations of letters: the palatal n as *nn* (or its scribal variant A), *gn,* nj, or *in;* palatal l as *ll, gl, lj,* il, or *yl*, as well as the combinations nh and *lh*, already mentioned. The Latin letter *x*, an abbreviation for *ks*, was also put to other uses in Romance; in Portuguese, Catalan, Sicilian, and Old Spanish it represents an sh sound, in modem Spanish a strong h sound, more commonly spelled with a *j*, and in northern Italian dialects the *z* sound. Other letters pressed into use for new consonantal sounds were *z* (used in Italy for *ts* and *dz* sounds, Germanic k and w, and the Visigothic *ç* for ts and sometimes *s*, as today in French and Portuguese).

Vowels were less of a problem for early Romance scribes-diphthongs were simply shown as vowel combinations such as ie, uo. Later, the diaeresis (¨) was sometimes introduced to distinguish diphthongs from adjoining vowels that were to be pronounced separately. Non-Latin vowels are rarely clearly distinguished: French *u* (pronounced like German *ü*), for instance, was written *u* and not consistently distinguished from Latin u (pronounced as in "lunar" and, in modern French, written *ou*). Nasal vowels in French are marked by a following n or m; in Portuguese a tilde (~) is often used for final nasal vowels and diphthongs (*ã, ãi*). Use of diacritics was not consistent until modern times; thus, so-called long and short *e*, still not always distinguished in Italian, are shown as *é* and *è* or *ê* (e.g., *élève* "student") in French (since the 18th century) and as e and *é* in Portuguese (since about 1930). Romanian established the use of *î* and *ă* only in the 20th century.

Diacritical Markings

In most of the languages with a long history of writing, the original attempts by scribes at phonological transcription were followed by an "etymological" period in which Latinized spelling gained ground. Castilian was least subject to this fashion, and, because its phonology has changed comparatively little since the Middle Ages (when its spellings became more or less fixed), it has few orthographical problems today. Standard Italian retains a fairly etymological orthography that covers up various minor regional differences of pronunciation; small reforms have been made through the centuries (in the 17th century, for instance, the use of h--except in *ho,* ha, hanno—was discontinued; in the 20th century, *î* and *j* for *ii*, as in studii, have virtually disappeared), but chaos still reigns in the use of accents. Romanian suffered from etymologizing orthography in the 19th century, but successive edicts of the Romanian Academy, of which the most important was dated 1932, have established a more or less phonetic spelling (the notable exception being the depiction of final "soft" consonants by a following i). Modern Catalan, like other “minor” languages, has had the aid of expert linguists in the establishment of its orthography. A standard was proposed by Antonio Maria Alcover Sureda, a Catalan priest, philologist, and writer from Majorca, in 1913, which is accepted, with small variations, by most writers.

Only two of the Romance languages, French and Portuguese, have had major orthographical problems, mainly resulting from the radical transformations that have affected their phonology since the Latin period. Portuguese has attempted to overcome its difficulties by a series of governmental reforms during the 20th century, but, in spite of official agreements between Portugal and Brazil in 1931 and 1945, there is still little consistency in usage, with Brazilian writers, especially, remaining more conservative (*i.e.*, etymological). In France, in spite of vociferous demands for reform since the 16th century, only minor changes have been accepted (usually originally from unofficial sources, such as printers), so that French orthography today reflects 12th-century phonology, overlaid by the etymologizing of Middle French legal scribes. Battles still rage between the reformers, who deplore the absurdly large proportion of school time devoted to teaching spelling, and the defenders of tradition, who point out that the phonological character of French, with no consistent phonetic markers for the word, make it unsuitable for phonetic transcription and that written French has its own structure, not identical with that of the spoken language.

BIBLIOGRAPHY. I. IORDAN and J. ORR, ***An*** *Zntroduction to* Romance *Linguistics*, revised reprint of the 1937 edition, with an additional essay, "Thirty Years On" by REBECCA **POSNER** (1970), describes the work done in Romance during the 19th

and 20th centuries and provides an extensive bibliography of all works written both on the family as a whole, and on individual languages. On the Romance languages in general, two books written in English will prove useful: the more philologically oriented The Romance Languages by W.D. ELCOCK (1960); and the more popular, linguistically oriented The Romance Languages: *A* Linguistic Introduction by REBECCA POSNER (1966). Outstanding is Y. MALKIEL, *Essays* on Linguistic Themes (1968), which includes several articles on the Romance languages and Romance linguistics. On the individual languages, among works in English particularly to be recommended, are L.R. PALMER, The Latin Language (1954); BRUNO MIGLIORINI, with T. GWYNFOR GRIFFITH, *Storia della* lingua *italiana* (1960, Eng. trans., The Italian Language, 1966); A.E. EWERT, The French Language, 2nd ed. (1961); and WILLIAM ENTWISTLE, The Spanish Language (1936).

(Re.P.)

Romania

Romania, a country of southeastern Europe (formally the Socialist Republic of Romania), derives much of its ethnic and cultural character from its position astride major continental migration routes. But emphasis should also be placed on three distinct elements basic to the physical geography. The vast arc of the Carpathian Mountains and their extension, the Transylvanian Alps, crosses the country from north to south, encircling the Transylvanian Plateau to create a huge natural amphitheatre. In terms of drainage, the country is indisputably Danubian, for the lower course of this great river runs eastward across the lowlands of the southern portion of the country, emptying into the Black Sea by way of a delta that is one of the finest natural endowments of the continent. Finally, a small eastern portion of the country, because of its Black Sea littoral, exhibits maritime characteristics.

Since the late 19th century, Romania has undergone an economic and social transformation marked by accelerating urbanism and a drop in the traditional predominance of agriculture. Romania's population of more than 21,-000,000 in 1975 made it nearly as populous as Canada, ninth in population among European countries, and sixth among the 14 nations of the Commuhist bloc. At 91,699 square miles (237,500 square kilometres), its area is comparable to that of Laos or Uganda. Its boundaries total 1,959 miles (3,153 kilometres), with the Soviet Union on the east and north, Hungary on the west, Yugoslavia on the southwest, Bulgaria on the south, and the Black Sea on the east. The national capital is Bucharest (Bucureşti), with a population exceeding 1,700,000 by 1976.

This article is concerned with the contemporary nation. For information on related subjects, see BALKANS, HISTORY OF THE; BLACK SEA; BUCHAREST; CARPATHIAN MOUNTAINS; and DANUBE RIVER.

THE LAND

There is a certain symmetry in the physical structure of Romania. The country forms a complex geographical unit centred on the Transylvanian Plateau, around which the peaks of the Carpathians and their associated subranges and structural platforms form a series of crescents. Beyond this zone, the extensive plains of the south and east of the country, their potential increased by the Danube and its tributaries, form a fertile outer crescent extending to the frontiers. There is great diversity in the topography, geology, climate, hydrology, flora, and fauna. For millennia this natural environment has borne the imprint of a human population, ever renewed by migratory movements but nevertheless having roots deep in the country's past.

Relief divisions. The *Carpathians.* The relief of Romania is dominated by the Carpathian Mountains, which can be divided into three major sections: the Eastern Carpathians (Carpafii Orientali), the Southern Carpathians (also known as the Transylvanian Alps and called in Romanian the Carpaţii Meridionali), and the Western Carpathians (Carpaţii Occidentali).

The Eastern Carpathians extend from the Soviet frontier to the Prahova River Valley and reach their maximum height in the Rodna Mountains (Munfii Rodnei), with Pietrosul rising to 7,556 feet (2,303 metres). They are made up of a series of parallel crests that are oriented in a more or less north–south direction. Within these mountains is a central core made up of hard, crystalline rocks and with a bold and rugged relief. Rivers have cut narrow gorges here (known locally as *chei*)—in, for example, Cheile Bistrifei and Bicazului—and these offer some magnificent scenery. This portion of the Carpathians is bounded on the eastern side by a zone of softer flysch. For some 250 miles on the western fringe the volcanic ranges Oag (Munfii Oagului) and Harghita, with a concentration of volcanic necks and cones, some with craters still preserved, lend character to the landscape. St. Ana Lake—the only crater lake in Romania—is also found here. The volcanic crescent provides rich mineral resources (notably copper, lead, and zinc) as well as the mineral-water springs on which are founded such health resorts as Vatra Dornei, Borsec, Tugnad, and Malnaş. The Carpathian range proper is made up in large part of easily weathered limestones and conglomerates, which again provide some striking scenery. The Maramureş, Giurgiu, Ciuc, and Birsei depressions further break up the mountainous relief.

Volcanic scenery

The Southern Carpathians, or Transylvanian Alps, lie between the Prahova River Valley on the east and the structurally formed Timiş and Cerna river valleys to the west. They are mainly composed of hard crystalline and volcanic rocks, which give the region the massive character that differentiates it from the other divisions of the Carpathians. The highest points in Romania are reached in the peaks of Moldoveanu (8,346 feet- [2,544 metres]) and Negoiu (8,317 feet [2,535 metres]), both in the Făgăraş Massif (Munfii Făgărasului), which, together with the Bucegi, Paring, and Retezat–Godeanu massifs, forms the major subdivision of the region. The last named contains a national park of more than 49,000 acres (20,000 hectares), which, besides offering spectacular mountain scenery, provides an important refuge for the chamois (Rupicapra rupicapra) and other animals. Ancient erosion platforms, another distinguishing feature of the area, have been utilized as pastures since the dawn of European history. At the highest levels, beautiful glacial lakes testify to the last Ice Age. The southern slopes of this region offer special interest: the waters of the Bistrifa, Cerna, and other rivers have carved deep valleys in the soft limestone rock, and the region also contains the Polovragi Cave and the Muierii Grotto. Communication is possible through the high passes of Bran, Novaci-Şugag, and Vîlcan, at altitudes of up to 7,400 feet, but the scenic Olt, Jiu, and Danube river valleys carry the main roads and railways through the mountains. At the Porţile de Fier (Iron Gate) on the Danube, a joint Romanian–Yugoslavian navigation and power project has harnessed the fast-flowing waters of the gorge; its power station has a capacity exceeding 2,000,000 kilowatts, and navigation facilities have been greatly improved. Finally, as in the Eastern Carpathians, there are important lowland depressions within the mountains (notably Brezoi, Haţeg, and Petrogani), and agriculture and industry are concentrated in them.

The Western Carpathians extend for about 220 miles (350 kilometres) between the Danube and Someş rivers. Unlike the other divisions of the Carpathians, these do not form a continuous range but a cluster of massifs around a north–south axis. Separating the massifs is a series of deeply penetrating structural depressions. Historically, these have functioned as easily defended "gates," as is reflected in their names: the Iron Gate of Transylvania (at Bistra); the Eastern Gate, or Poarta Orientaliă (at Timiş-Cerna); and, most famous, the Iron Gate on the Danube.

Historic "gates"

Among the massifs themselves, the Banat and Poiana Ruscă mountains contain a rich variety of mineral resources and are the site of two of the country's three largest metallurgical complexes, at Regifa and Hunedoara. The marble of Ruschiţa is well known. To the north lie the Apuseni Mountains, centred on the Bihor Massif, from which emerge fingerlike protrusions of lower relief. On the east, the Bihor Mountains merge into the limestone tableland of Cetăţile Ponorului, where the erosive action of water along joints in the rocks has created a fine example of the rugged karst type of scenery. To the west lie the parallel mountain ranges of Zărand, Codru-Moma (also called Munţii Codrului), and Pădurea Craiului; on the south, along the Mureş River, the Metaliferi and Trască mountains contain a great variety of metallic and other ores, with traces of ancient Roman mine workings still visible.

The Western Carpathians generally are less forested than other parts of the range, and human settlement reaches to the highest altitudes. The population maintains many traditional features in architecture, costumes, and social mores, and the old market centres, or *nedei,* are still important. The Tîrgul de Fete (Maidens' Fair) is still held every year in late July on Găina Mountain and is attended by peasants from the Arieş and Crişul Alb valleys. Mining, livestock raising, and agriculture are the main economic activities, the last named being characterized by terrace cultivation on the mountain slopes, a survival from Roman times.

The Subcarpathians. The great arc of the Carpathians is accompanied by an outer fringe of rolling terrain known as the Subcarpathians and extending from the Moldova River in the north to the Motru River in the southwest. It is from two to 19 miles wide and reaches heights ranging between 1,300 and 3,300 feet (400 and 1,000 metres). The topography and the milder climate of this region favour vegetation (including such Mediterranean elements as the edible chestnut) and aid agriculture; the region specializes in cereals and fruits, and its wines—notably those of Odobegti and the Călugărească Valley—have a European reputation. The area is densely populated, and there are serious problems of economic development in remoter areas where there is little scope for further agricultural expansion.

Carpathian "double towns"

The tablelands. Tablelands are another important element in the physical geography of Romania. The largest is in Transylvania, with large deposits of methane gas and salt, first exploited for a chemical industry in the 1930s. The salt lakes have given rise to the health resorts of Ocna Sibiului and Sovata. The region as a whole is well populated, with a good transport system. A belt of towns has grown up on the margins, and these often parallel another outer fringe of towns commanding the main trans-Carpathian passes. Examples of such "double towns" are Suceava and Bistrifa; Făgăraş and Cîmpulung; Sibiu and Rîmnicu Vîlcea; Alba Iulia and Arad; and Cluj and Oradea.

In the east, between the outer fringe of the Subcarpathians and the Prut River, lies the Moldavian Plateau (Podigul Moldovei), with an average height of 1,600 to 2.000 feet. In contrast to Transylvania, which experienced considerable urban development during the Dacian and Roman periods, Moldavia did not begin to develop towns till the Middle Ages, when the old Moldavian capitals of Iagi and Suceava had close commercial connections with the towns of Transylvania and derived benefit from trade passing between the Baltic and Black Sea ports. Finally, the Dobruja tableland, an ancient, eroded, rock mass in the southeast, has an average altitude of 820 feet and reaches a maximum of 1,532 feet (467 metres) in the Pricopan Hills (Dealul Pricopanului).

The plains. Plains cover a third of Romania, reaching their fullest development in the south and west. Their economic importance has increased very greatly since the early 19th century. In the southern part of Romania is the lower Danube Plain, which can be divided into the Romanian Plain (Cimpia Romînă), to the east of the Olt River, and the Oltenian Plateau (Podişul Olteniei), to the west. The whole region is covered by deposits of loess, on which rich, black chernozem soils have developed, providing a strong base for agriculture. The Danube floodplain is important economically, and along the entire stretch of the river, from Calafat in the west to Galafi in the east, former marshlands have been diked and drained to increase food production. Willow and poplar woods border the river, which is important for fishing but much more so for commerce. Ten river port towns (including Drobeta–Turnu Severin, Turnu Măgurele, Giurgiu, Brăila, and Galafi) complement the rural settlements. There are good rail connections with the main lines, including the two that cross the Danube, at Cernavodă (linking Bucharest with the Black Sea port of Constanfa) and Giurgiu (connecting Romania with Bulgaria).

Economic importance of the Danube Delta

The Danube Delta. On the northern edge of the Dobruja region, adjoining the Soviet Union, the great, swampy triangle of the Danube Delta is a unique physiographic region covering some 1,950 square miles, of which 1,750 square miles (4,530 square kilometres) are in Romania. The delta occupies the site of an ancient bay, which in prehistoric times became wholly or partially isolated from the sea by the Letea sandbanks. The delta contributes about half of Romania's fish production from home waters, fishing off the Danube mouth contributing 90 percent of the sturgeon catch (and subsequent caviar production) as well as 80 percent of Danube herring catch. The plant and animal life of the delta region is unique in Europe, with many rare species. The area is also a stopping place for migratory birds. A great number of birds, including pelicans, swans, wild geese, ibis, and flamingos, as well as wild pig and lynx, are protected by law, and a large part of the region has been declared a nature reserve, with hunting and fishing prohibited. The whole delta is of great interest to scientists, conservationists, and a growing number of tourists from other countries. Two dozen or more settlements are scattered over the region, but many are exposed to serious flood risks and only two (the ports of Sulina and Tulcea) have attained urban status.

The Black Sea coast. The Black Sea coastal strip has its own special environment, including a temperate climate with continental aspects and good sand beaches. Lakes—among which Taşaul, Siutghiol, Agigea, Techirghiol, and Mangalia are the most significant—further enhance the attractions of the region. Several of them contain deposits of mud and sulfurous hot springs believed to have therapeutic properties. The development of recreational facilities dates back to the turn of the century, and a series of new health and tourist centres has sprung up in recent decades. The towns of Năvodari, Mamaia, and Eforia are entirely new creations, while the older settlements of Mangalia and Techirghiol have undergone extensive redevelopment. It seems likely that the whole 34-mile stretch between Năvodari and Mangalia will become within a few years a single resort area.

Climate. Romania's geographic situation in the southeastern portion of the European continent gives it a climate that is transitional between temperate regions and the harsher extremes of the continental interior. In the centre and west, humid Atlantic climatic characteristics prevail; in the southeast the continental influences of the Russian Plain make themselves felt; and in the extreme southeast there are even milder sub-Mediterranean influences. This overall pattern, however, is substantially modified by relief, and there are many examples of climatic zones induced by altitudinal changes.

The effect of relief on climate

The average annual temperature is 52° F (11° C) in the south and 45° F (7" C) in the north, although, as noted, there is much variation according to altitude and related factors. Extreme temperatures range from 111° F (44" C) in the Bărăgan region to —36" F (—38" C) in the Braşov Depression. Average annual rainfall amounts to 26 inches (660 millimetres), but in the Carpathians it reaches about 55 inches (1,400 millimetres) and in the Dobruja region it is only about 16 inches (400 millimetres). Humid winds from the northwest are commonest, but often the drier winds from the northeast are'strongest. A hot southwesterly wind, the *austru,* blows over western Romania, particularly in summer. In winter, cold and dense air masses encircle the eastern portions of the nation, with the cold northeasterly known as the *crivăţ* blowing in from the Russian Plain, while oceanic air masses from the Azores, in the west, bring rain and mitigate the severity of the cold.

Drainage. The rivers of Romania are virtually all tributary to the Danube, which forms the southern frontier from Moldova Nouă to Călăraşi. Nearly 40 percent of the total Danubian discharge into the Black Sea is, in fact, provided by Romanian rivers. The final discharge takes place through three arms—the Chilia (67 percent of the flow), Sulina (9 percent), and Sfintu Gheorghe (24 percent)—that add to the scenic attractions of the delta region. The most significant of the Romanian tributary rivers are the Prut, Mureş, Olt, Siret, Ialomifa, and Someş. The rivers have considerable hydroelectric potential, although there are great seasonal fluctuations in the discharge and few natural lakes to regulate the flow. The total surface-water potential of the tributary rivers exceeds 1,400,000,000,000 cubic feet (40,000,000,000 cubic metres) annually, although this figure is dwarfed by the volume discharged at the Danube mouth, which is more than five times as large. Subsoil waters have been estimated at an annual volume of some 250,000,000,000 cubic feet (7,000,000,000 cubic metres). These overall figures, like those for many aspects of the Romanian environment, mask the fact that water resources are not uniformly distributed over the country and may vary not only from year to year but within the same year.

Hydro-electricity

The total theoretical hydroelectric potential of Romania —given optimum technological conditions—hasbeen calculated at some 70,000,000,000 kilowatt-hours in an average year. About half this amount might well be harnessed, given the present state of technology, but for economic reasons a figure of 24,000,000,000 kilowatt-hours represents a more feasible goal. Geographically, the hydroelectric reserves of Romania are concentrated along the Danube and in the valleys of rivers emerging from the mountain core of the country. Other hydrographic resources include the more than 2,500 lakes, ranging from the glacial lakes of the mountains to those of the plains and the marshes of the Danube Delta region. The main effort since the 1940s, however, has been on the Argeg, Bistrifa, Lotru, Olt, and Someş as well as on the Danube at the Iron Gate.

Plant and animal life. Forests, which cover almost 27 percent of Romania, are an important component of the vegetation cover, particularly in the mountains. Up to about 2,600 feet (800 metres) oaks predominate, followed by beeches between 2,600 and 4,600 feet and conifers between 4,600 and 5,900 feet (1,400 and 1,800 metres). At the highest levels, alpine and subalpine pastures are found. In the tableland and plains regions, the natural vegetation has to a large degree been obliterated by centuries of human settlement and agriculture.

The rich and varied Romanian animal life includes some rare species, notably the chamois, which is found on the alpine heights of the Carpathians. Forest animals include brown bear, red deer, wolf, fox, wild pig, lynx, marten, and various songbirds. The lower course of the Danube, and particularly the delta, is rich in animal, bird, and fish life. Among the last named the most valuable is undoubtedly the sturgeon, yielding caviar.

THE PEOPLE

Settlement patterns. The natural environment of Romania has offered favourable conditions for human settlement. The accessibility of the region to the movements of peoples across the Eurasian landmass has also meant that the region has absorbed cultural influences from many nations and peoples, and this, too, is reflected in the contemporary patterns of Romanian life.

The population is fairly uniformly distributed from the shores of the Black Sea, across the plains, and up to the mountain foothills. The mountain areas themselves are inhabited by peoples whose origins are found in very early European history and who have, in many respects, been little changed by contemporary events. The villages scattered among the mountains up to altitudes of more than 5,000 feet, the sheepfolds, the newer holiday resorts, and the large number of roads all combine to give the human geography of the Carpathian Mountains a distinctive character. Pastures on the ancient erosional platforms among the mountain peaks can be found up to the highest levels, and cultivation is possible up to 3,900 to 4,300 feet. Further, the ancient commercial trade between the old market towns on either side of the Carpathians lends support to the view that the mountains have served as much as a link as a barrier in the country's development. In the many lowland areas scattered among the mountains there has been a long continuity of settlement, as may be seen in very old place-names and distinct regional consciousness.

The Carpathian heritage

Ethnic origins. *Ancient times.* Historical and archaeological evidence and linguistic survivals seem to confirm that the present territory of Romania had a fully developed population, with a high degree of economic, cultural, and even political development, long before the Roman armies crossed the Danube into what became known as the province of Dacia. Roman influence was profound, creating a civilization that managed to maintain its identity during the great folk migrations that followed the collapse of the empire. The Dacian–Roman population of the region led a life in which farming and particularly transhumance played an important part. They lived in small settlements, sometimes retiring to places providing better shelter in the troubled centuries. Life was not entirely transhumant, however, and a primitive agriculture was practiced on the upland terraces and in the more secluded river valleys. Thus the ethnic core of contemporary Romania developed in the remoter regions, although settlement did take place on the more exposed plains. The Romanian language—the morher tongue of almost 90 percent of the people, one of the Romance languages, with Latin grammatical structure—was formed between the 7th and 10th centuries. The first mention of Walachs (Volokhs, Vlachs), the name given to the Romanian people by their neighbours, appears in the 9th century.

Origin of the language

The feudal period. As more settled conditions gradually came to prevail on the continent, more complex economic activities—embracing agriculture, fruit growing, viticulture, and livestock raising—developed in both the mountains and plains. The densest settlement was in the Subcarpathians, where the population could exchange products with persons living on both sides of the mountain range. Population density increased as the feudal economy developed. In the 13th century, the existing largely Romanian population was augmented by colonists brought to Transylvania, particularly into the Carpathians, and including Saxons, Szeklers, and Teutonic Knights. The proliferation of mining also brought in foreign elements. The Hungarian monarchy wished to consolidate the defenses of eastern Transylvania against raiders from the steppes, but during the same period, with the mountain crests marking a political frontier, there appeared two independent Romanian feudal states: Walaehia (called in Romania Ţara Românească, literally "Romanian Land") and Moldavia (Romanian Moldova), both on the southern and eastern slcpes of the Carpathians. Initially, the core areas of these states were centred in the foothills of the Carpathians, at Curtea de Argeş, Cimpulung, and Baia; only later, as the Romanian lands on the plains were gradually consoiidated, were the major settlements transferred from the mountains, first to Tîrgovişte and Suceava and later to Bucharest and Iaşi. This marked an important stage in the development of the human geography of Romania.

The modern period. The 19th century saw important political advances, notably the unification of the two Romanian principalities (1859) and the attainment of independence under a Hohenzollern monarch (1878). The monarchy survived until 1947, when the Soviet Union, whose forces were maintaining the first Russian occupation (1944–58) in the country's independent history, demanded a complete Communist takeover. Although the Romanian Communist Party seeks credit for recent economic achievements, it should not be overlooked that for the first 70 years of independence the monarchy presided over the modernization of the country, including the expansion of education and the building of administrative and transport systems to sustain industrialization.

The country increased in size, first by the annexation of southern Dobruja in 1913 and second by the acquisition of Bessarabia, Bukovina, and Transylvania in 1918–19. As a consequence of World War II, southern Dobruja was lost to Bulgaria, and Bessarabia and northern Bu-

MAP INDEX

Political subdivisions

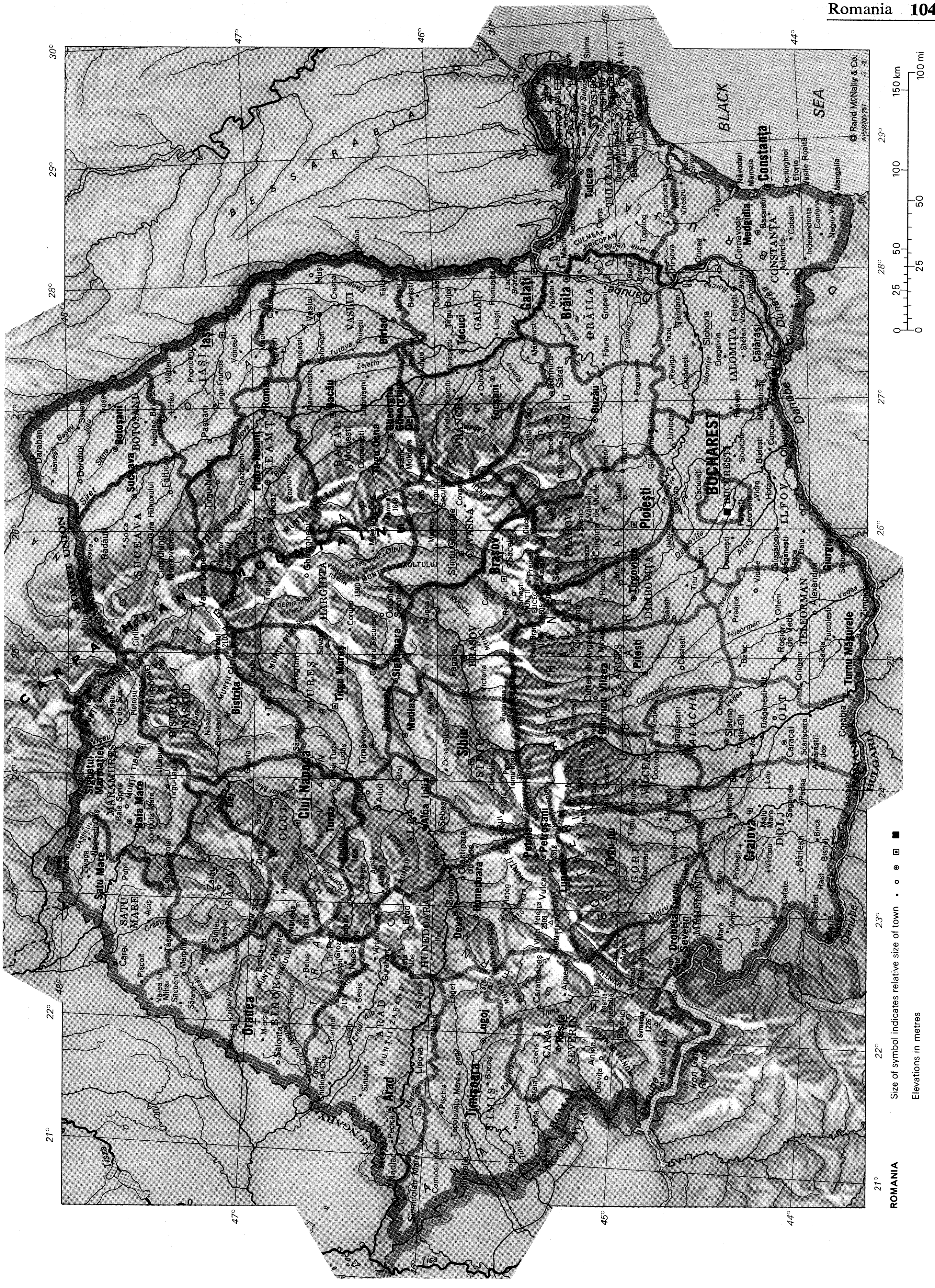
Alba	46·15n	23·30e
Arad	46·20n	21·40e
Argeș	45·00n	24·45e
Baciu	46·30n	26·45e
Bihor	47·00n	22·15e
Bistrița-Năsăud	47·15n	24·30e
Botoqani	47·50n	26·45e
Brăila	45·00n	27·40e
Bragov	45·45n	25·15e
București	44·20n	26·10e
Buzău	45·15n	26·40e
Caraș-Severin	45·15n	22.00e
Cluj	46·45n	23·45e
Constanfa	44·20n	28·20e
Covasna	46·00n	26·00e
Dîmbovița	45·00n	25·30e
Dolj	44·15n	23·45e
Galați	45·45n	27·45e
Gorj	45·00n	23·20e
Harghita	46·35n	25.30e
Hunedoara	45·45n	23·00e
Ialomița	44·30n	27·20e
Iași	47·15n	27·15e
Ilfov	44·30n	26·15e
Maramureg	47·40n	23·40e
Mehedinti	44·30n	22·50e
Mureș	46·35n	24·40e
Neamț	47·00n	26·30e
Ôit	44·20n	24·30e
Prahova	45·15n	26·00e
Sălaj	47·15n	23·15e
Satu Mare	47·40n	22·50e
Sibiu	46·00n	24·15e
Suceava	47·30n	25·45e
Teleorman	44·00n	25·15e
Timiș	45·40n	21·20e
Tulcea	45·00n	29·00e
Vaslui	46·30n	27·45e
Vilcea	45·19n	24·00e
Vrancea	45·45n	27·00e

Cities and towns

Abrud	46·16n	23·04e
Acîș	47·33n	22·47e
Adamclisi	44·05n	27·57e
Adjud	46·04n	27·11e
Agnita	45·59n	24·38e
Aiud	46·19n	23·44e
Alba Iulia	46·04n	23·35e
Aleșd	47·04n	22·25e
Alexandria	43·59n	25·20e
Amărăștii de Jos	43·57n	24·10e
Anina	45·05n	21·51e
Arad	46·11n	21·20e
Armeniș	45·12n	22·19e
Azuga	45·27n	25·33e
Babadag	44·54n	28·43e
Bacău	46·34n	26·55e
Baia de Aramă	45·00n	22·49e
Baia Mare	47·40n	23·35e
Baia Sprie	47·40n	23·42e
Băile Govora	45·05n	24·06e
Bâile Herculane	44·54n	22·25e
Băile Olănești	45·11n	24·16e
Băilești	44·02n	23·21e
Balaci	44·21n	24·55e
Bălcești	44·37n	23·56e
Balș	44·21n	24·06e
Băneasa	44·04n	27·42e
Basarabi	44·10n	28·24e
Beceni	45·23n	26·46e
Bechet	43·46n	23·58e
Beclean	47·11n	24·10e
Beiuș	46·40n	22·21e
Berești	46·06n	27·53e
Bicaz	46·55n	26·04e
Bîrca	43·58n	23·37e
Bîrlad	46·14n	27·40e
Bistreț	43·54n	23·30e
Bistrița	47·08n	24·30e
Blaj	46·11n	23·55e
Borș	47·07n	21·49e
Borșa	46·56n	23·40e
Botoșani	47·45n	26·40e
Bozovici	44·55n	21·59e
Brad	46·08n	22·47e
Brăila	45·16n	27·58e
Brașov	45·39n	25·37e
Bratca	46·56n	22·37e
Breaza	45·11n	25·40e
Bucharest	44·26n	26·06e
Budești	44·14n	26·28e
Buhuși	46·43n	26·41e
Burila Mare	44·27n	22·34e
Buzău	45·09n	26·49e
Buziaș	45·39n	21·36e
Căciulați	44·38n	26·10e
Calafat	43·59n	22·56e
Călărași	44·11n	27·20e
Călimănești	45·14n	24·20e
Călugăreni	44·07n	26·01e
Caracal	44·07n	24·21e
Caransebeș	45·25n	22·13e
Carei	47·41n	22·28e
Casimcea	44·43n	28·23e
Căzănești	44·37n	27·01e
Cehu Siivaniei	47·25n	23·11e
Cermei	46·33n	21·51e
Cerna	45·04n	28·18e
Cernavodă	44·21n	28·01e
Cetate	44·06n	23·03e
Chișineu-Criș	46·31n	21·31e
Cîmpeni	46·22n	23·03e
Cîmpia Turzii	46·33n	23·54e
Cîmpina	45·08n	25·44e
Cimpulung	45·16n	25·03e
Cirnpulung Moldovenesc	47·31n	25·34e
Ciocănești	44·12n	27·04e
Cirlibaba	47·35n	25·07e
Cisnădie	45·43n	24·09e
Cluj-Napoca	46·47n	23·36e
Cobadin	44·04n	28·13e
Codăești	46·52n	27·46e
Codlea	45·42n	25·27e
Colonești	46·34n	27·18e
Comana	43·54n	28·19e
Comănești	46·25n	26·26e
Comloșu Mare	45·54n	20·38e
Constanța	44·11n	28·39e
Corabia	43·46n	24·30e
Corbu	44·29n	24·43e
Corund	46·28n	25·11e
Corzu	44·28n	23·10e
Costești	44·40n	24·53e
Covasna	45.51n	26·11e
Craiova	44·19n	23·48e
Crasna	45·36n	26·08e
Crîngeni	44.01n	24·47e
Cristuru Secuiesc	46·17n	25·02e
Crucea	44·32n	28·14e
Curcani	44·12n	26·35e
Curtea de Argeș	45·08n	24·41e
Curtici	46.21n	21·18e
Daia	44·03n	25·59e
Dămienești	46·44n	26·59e
Darabani	48·00n	26·23e
Deda	46·57n	24·53e
Dej	47·09n	23·43e
Deta	45·24n	21·13e
Deva	45·53n	22·55e
Dobroteasa	44·47n	24·23e
Doctor Petru Groza	46·32n	22·28e
Domnești	44·25n	25·27e
Dorohoi	47·57n	26·24e
Dragalina	44·26n	27·20e
Drăgănești-Olt	44·10n	24·32e
Drăgănești-Vlașca	44·06n	25·36e
Drăgășani	44·40n	24·16e
Drobeta-Turnu-Severin	44·18n	22·39e
Dumbrăveni	46·14n	24·35e
Dunavățu-de-Sus	44·59n	29·13e
Eforia	44·00n	28·19e
Ezeriș	45·24n	21·53e
Făgăraș	45.51n	24·58e
Făget	45·51n	22·10e
Fălciu	46·18n	28·08e
Fălticeni	47·28n	26·18e
Făurei	45·06n	27·09e
Fetești	44·23n	27·50e
Fieni	45·08n	25·25e
Filiași	44·33n	23·31e
Focgani	45.41n	27·11e
Foeni	45·30n	20·53e
Frumușița	45·40n	28·04e
Furculești	43·52n	25·09e
Găești	44·43n	25·19e
Galați	45·26n	28·03e
Gătaia	45·26n	21·26e
Gheorghe Gheorghiu-Dej	46·14n	26·22e
Gheorgheni	46·43n	25·36e
Gherla	47·02n	23·55e
Girbovu	44·44n	23·21e
Giurgiu	43·53n	25·57e
Glodeanu-Siliștea	44·50n	26·48e
Gorgota	44·47n	26·05e
Gorgova	45·11n	29·10e
Gropeni	45·04n	27·53e
Gruia	44·16n	22·42e
Gurahouț	46·16n	22·21e
Gura Hurnorului	47·33n	25·48e
Hațeg	45·37n	22·57e
Hîrlău	47·25n	26·54e
Hirșova	44·41n	27·57e
Holod	46·47n	22·08e
Horezu	45·09n	24·01e
Hotarele	44·10n	26·22e
Huedin	46·52n	23·02e
Hunedoara	45·45n	22·54e
Huși	46·40n	28·04e
Iași	47·10n	27·35e
Iazu	44·44n	27·25e
Ibănești	48·04n	26·22e
Ilia	45·56n	22·39e
Independența	43·58n	28·05e
Ineu	46·26n	21·49e
Isaccea	45·16n	28·28e
Ișalnița	44·24n	23·44e
Jebel	45·33n	21·14e
Jimbolia	45·47n	20·43e
Lăpuș	47·30n	24·01e
Leu	44·11n	24·00e
Lichitișeni	46·23n	27·17e
Liești	45·38n	27·32e
Lipova	46·05n	21·40e
Livada	47·52n	23·07e
Luduș	46·29n	24·05e
Lugoj	45·41n	21·54e
Lupeni	45·22n	23·13e
Măcin	45·15n	28·08e
Malnaș	46.01n	25·50e
Malu Mare	44·15n	23·51e
Mamaia	44·17n	28·37e
Mangalia	43·48n	28·35e
Mărășești	45·52n	27·14e
Marghita	47.21n	22·21e
Mărtinești	45·30n	27·18e
Medgidia	44·15n	28·16e
Medias	46·10n	24·21e
Mehadia	44·55n	22·22e
Miercurea-Ciuc	46·22n	25·42e
Miersig	46·53n	21·51e
Mihăești	45·07n	25·00e
Mihai Viteazu	44·39n	28·41e
Mînăstirea	44·13n	26·54e
Mironeasa	46·58n	27·25e
Mizil	45·00n	26·26e
Moinești	46·28n	26·29e
Moldova Nouă	44·44n	21·40e
Moreni	45·00n	25·39e
Murgeni	46·12n	28·01e
Nădlac	46·10n	20·45e
Năsăud	47·17n	24·24e
Năvodari	44·19n	28·36e
Negrești	46·50n	27·27e
Negrești-Oaș	47·52n	23·25e
Negru Vodă	43·49n	28·12e
Nicolae Bălcescu	47·34n	26·52e
Nucet	46·28n	22·35e
Oancea	45·55n	28·06e
Ocna Mureș	46·23n	23·51e
Ocna Sibiului	45·53n	24·03e
Odobești	45·45n	27·04e
Odorheiu Secuiesc	46·18n	25·18e
Olteni	44·10n	25·18e
Oltenița	44·05n	26·39e
Oradea	47·03n	21·57e
Orăștioara de Jos	45·46n	23·11e
Oravița	45·02n	21·41e
Orșova	44·42n	22·24e
Osica de Jos	44·15n	24·17e
Ostrov	44·06n	27·22e
Padea	44.01n	23·52e
Panciu	45·55n	27·05e
Pagcani	47·15n	26·44e
Pătîrlagele	45·19n	26·22e
Pecica	46·10n	21·05e
Periprava	45·24n	29·32e
Petrila	45·27n	23·25e
Petroșani	45·25n	23·22e
Piatra-Neamț	46·56n	26·22e
Piatra-Olt	44·24n	24·08e
Pișchia	45·55n	21·20e
Pișcolt	47·35n	22·18e
Pitești	44·52n	24·52e
Ploiești	44·56n	26·02e
Podu Turcului	46·12n	27·23e
Pogoanele	44·54n	27·00e
Pomi	47·42n	23·19e
Popești	47·14n	22·25e
Popești-Leordeni	44·23n	26·10e
Popricani	47·18n	27·31e
Preajba	44·24n	25·20e
Predeal	45·30n	25·35e
Predești	44.21n	23·36e
Pucioasa	45·34n	25·26e
Puiegti	46·25n	27·33e
Pungești	46·42n	27·20e
Răcari	44·38n	25·45e
Rădăuți	47.51n	25·55e
Rădinești	44·48n	23·46e
Rast	43·53n	23·17e
Răsvani	44·25n	26·53e
Rizboeni	47·05n	26·32e
Reghin	46·47n	24·32e
Reșița	45·17n	21·53e
Reviga	44·42n	27·06e
Rimnicu Sărat	45·23n	27·03e
Rimnicu Vilcea	45·06n	24·22e
Rîșnov	45·36n	25·28e
Roman	46·55n	26·56e
Rogiori de Vede	44·07n	25·00e
Rovinari	44·55n	23·11e
Roznov	46·50n	26·31e
Rupea	46·02n	25·13e
Săcele	45·37n	25·42e
Săcueni	47·21n	22·06e
Sagu	46·03n	21·17e
Sălard	47·13n	22·03e
Salcia	43·57n	24·56e
Salonta	46·48n	21·40e
Sărmașu	46·46n	24·11e
Satu Mare	47·48n	22·53e
Săveni	47·57n	26·52e
Săvîrșin	46·01n	22·14e
Scărișoara	44·00n	24·35e
Sebeș	45·58n	23·34e
Sebiș	46·23n	22·08e
Șegarcea	44·06n	23·45e
Sfintu Gheorghe	45·52n	25·47e
Sibiu	45·48n	24·09e
Sighetul Marrnatiei	47·56n	23·54e
Sighișoara	46·13n	24·48e
Simeria	45·51n	23·01e
Șimleu Siivaniei	47·14n	22·48e
Sinaia	45.21n	25·33e
Sinnicolau Mare	46·05n	20·38e
Sintana	46·21n	21·30e
Siret	47·57n	26·04e
Slănic	45·15n	25·57e
Slănic Moldova	46·13n	26·26e
Slatina	44·26n	24·22e
Slobozia	44·34n	27·23e
Solacolu	44·23n	26·34e
Solca	47·42n	25·50e
Șorncuta Mare	47·31n	23·29e
Sovata	46·35n	25·04e
Ștefan Vodă	44·19n	27·19e
Ștrehaia	44·37n	23·12e
Suceava	47·39n	26·19e
Sulina	45·09n	29·40e
Țăndărei	44·38n	27·40e
Tarna Mare	48·05n	23·20e
Tășnad	47·29n	22·35e
Teaca	46·55n	24·31e
Techirghiol	44·03n	28·36e
Tecuci	45·50n	27·26e
Tirnișoara	45·45n	21·13e
Tîrgoviște	44·56n	25·27e
Tîrgu Buior		
Tîrgu *Cărbunești*	44·58n	23·31e
Tîrgu Frumos	47·13n	27·00e
Tîrgu Jiu	45·02n	23·17e
Tîrgu Lăpuș	47·27n	23·52e
Tîrgu Mureș	46·33n	24·33e
Tîrgu Neamț	47·12n	26·22e
Tîrgu Ocna	46·15n	26·37e
Tîrgu Secuiesc	46·00n	26·08e
Tîrgusor	44·28n	28·25e
Tîrnăveni	46·20n	24·17e
Titu	44·41n	25·16e
Toplița	46·55n	25·21e
Topolog	44·53n	28·22e
Topolovățu Mare	45·46n	21·37e
Trușești	47·46n	27·01e
Tulcea	45·11n	28·48e
Turda	46·34n	23·47e
Turnu Măgurele	43·45n	24·53e
Ulma	47·53n	25·18e
Ulmeni	45·04n	26·39e
Urlați	44·59n	26·14e
Urziceni	44·43n	26·38e
Videni	45·22n	27·56e
Valea lui Mihai	47·31n	22·09e
Vălenii de Munte	45·12n	26·03e
Vașcău	46·28n	22·28e
Vasile Roaită	44·03n	28·38e
Vaslui	46·38n	27·44e
Vața de Jos	46·10n	22·35e
Vatra Dornei	47.21n	25·21e
Vedea	44·47n	24·37e
Vetrisoaia	46·26n	28·13e
Victoria	45·45n	24·41e
Videle	44·16n	25·31e
Vidra	44·16n	26·11e
Vidra	45·55n	26·54e
Vinju Mare	44·26n	22·52e
Vintilă Vodă	45·28n	26·44e
Virfurile	46·19n	22·31e
Virtopu	44·12n	23·21e
Vișeu de Sus	47·44n	24·22e
Vlădeni	47·25n	27·20e
Voinești	47·05n	27·26e
Vulcan	45·23n	23·17e
Zăbalț	46.01n	21·55e
Zalău	47.11n	23·03e
Zărnești	45·34n	25·19e
Zerind	46·37n	21·31e
Zirnbor	47·00n	23·16e
Zimnicea	43·39n	25·21e

Physical features and points of interest

Alb, Crișul, *river*	46·36n	21·20e
Almăjului, Munții, mountains	44·43n	22·12e
Alma$. *river*	47·14n	23·19e
Argeș, *river*	44·04n	26·37e
Arieșul, *river*	46·26n	23·59e


ROMANIA
Size of symbol indicates relative size of town
Elevations in metres
© Rand McNally & Co.
0 25 50 100 150 km
0 25 50 100 mi
BLACK SEA
SOVIET UNION
BESSARABIA
HUNGARY
YUGOSLAVIA
BULGARIA
CARPATHIAN MOUNTAINS
TRANSYLVANIA
MOLDAVIA
WALLACHIA
DOBRUDJA
BANAT
Danube
BUCHAREST
Constanţa
Galaţi
Brăila
Iaşi
Cluj-Napoca
Timişoara
Braşov
Ploieşti
Craiova
Oradea
Arad
Sibiu
Tulcea
Suceava
Bacău
Piteşti
Tîrgu Mureş
Satu Mare
Baia Mare
Buzău
Drobeta-Turnu Severin
Giurgiu

MAP INDEX (continued)

Banat, physical regïon	45·30n	21·00e
Baraoltului, Muntii, mounfains	46·15n	25·45e
Başeu, river	47·44n	27·15e
Bega, river	45·13n	20·19e
Beretău, river	47·14n	21·52e
Bessarabia, historic region	47·00n	28·30e
Bicaz, Izvorul Munteiui reservoir	47·00n	26·00e
Bîstra, river	45·29n	22·11e
Bistriţa, river	46·30n	26·57e
Black Sea	44·10n	29·20e
Borcea, *river*	44·40n	27·53e
Brăilei, Balta, marsh	45·00n	28·00e
Bran, Pasul, pass	45·26n	25·17e
Brateş, Lacul, lake	45·30n	28·05e
Bucegi, Muntii, *mountains*	45·27n	25·26e
Bukovina, historic region	48·00n	25·30e
Buzău, river	45·26n	27·44e
Buzăului, Muntii, mountains	45·35n	26·10e
Călimani, Muntii, mountains	47·07n	25·03e
Călmătui, *river*	44·50n	27·50e
Căpăţînii, Munţii, mountains	45·20n	24·00e
Carpathian Mountains	47·00n	25·30e
Cerna, river	45·00n	22·35e
Cernei, Muntii, mountains	45·02n	22·31e
Chilia, Braţul, stream	45·18n	29·40e
Cimpia Romînă, see Walachia		
Ciuc, Depresiunea, depression	46·20n	25·45e
Ciucaş, peak	45·31n	25·55e
Codru-moma, Muntii, mouhtains	46·30n	22·20e
Cotmeana, *river*	44·24n	24·45e
Crasna, river	47·46n	22·27e
Crişul Aib, *river*	46·42n	21·17e
Crişul Negru, *river*	46·38n	21·26e
Crişul Repede, nver	47·03n	21·42e
Curcubăta, *peak*	46·28n	22·36e
Danube (Dunărea), *river*	45·20n	29·40e
Dîmboviţa, *river*	44·14n	26·27e
Dobruja, physical region	44·20n	2810e
Dranov, Ostrovul, island	44·52n	29·15e
Dunărea, see Danube, river		
Dunărea Veche, *river*	45·17n	28·02e
Dunării, Delta	45·10n	29·20e
Eastern Carpathians, mountains	47·00n	26·00e
Elan, river	46·07n	28·04e
Făgăras, Munfii, mountains	45·35n	25·00e
Farcaui, peak	47·55n	24·27e
Giurge, Depresiunea, depression	46·45n	25·30e
Gorul, *peak*	45·48n	26·25e
Gurghiului, Muntii, mountains	46·41n	25.12e
Harghita, Muntii, mountains	46·31n	25·33e
Haţeg, Depresiunea, depression	45·35n	23·00e
lalomiţa, river	44·42n	27·51e
lalomiţei, Balta, mars*h*	44·30n	28·00e
Iron Gate, gorge	44·41n	22·31e
Iron Gate, reservoir	44·30n	22·00e
Jijia, river	46·54n	28·05e
Jiu, river	43·47n	23·48e
Letea, Ostrovul, island	45·18n	29·15e
Lotrului, Munţii, mountaïns	45·30n	23·52e
Maramureşului, Muntii, mountains	47·50n	24·45e
Mare, Muntele, mountain	46·29n	23·14e
Moldavia, historic region	46·30n	27·00e
Moldova, river	46·54n	26·58e
Moldoveanu, peak	45·36n	24·44e
Motru, river	44·33n	23·27e
Muntelui, Virful, peak	46·29n	23·14e
Mures, river	46·10n	20·28e
Neajlov, river	44·11n	26·12e
Nemira, peak	46·15n	26·19e
Oaşului, Muntii, mountains	47·58n	23·15e
Oituz, Pasul, pass	46·03n	26·23e
Olt, river	43·43n	24·51e
Omu, oeak	45·26n	25·26e
Orientaliă, Poarta, pass	45·06n	22·18e
Pădurea Craiului, Munţii, *mountains*	46·55n	22·20e
Parîngului, Munţii, mountains	45·20n	23·40e
Peleaga. Vîrful, *peak*	45·22n	22·54e
Perşani, Muntii, mountains	45·40n	25·15e
Pietrosu, peak	47·36n	24·38e
Pogăniş, river	45·41n	21·22e
Poiana Ruscăi, Muntii, mountains	45·41n	22·30e
Prahova, river	44·43n	26·27e
Pricopan, Culmea, hills	45·14n	28·20e
Prut, river	45·28n	28·12e
Razelm, Lacul, lake	44·54n	28·57e
Retezat. Muntii, mountains	45·25n	23·00e
Rirnna, river	45·39n	27·19e
Rodnei, Muntii, mountains	47·35n	24·40e
Sebeşului, Munţii, mountains	45·38n	23·27e
Semenic, Muntii, mountains	45·05n	22·05e
Şes, Muntii, mountains	47·05n	22·30e
Sfintu Gheorghe, Ostrovul, island	45·07n	29·22e
Sfintu Gheorghe, Braţul, stream	44·53n	29·36e
Sinoe, Lacul, lake	44·38n	28·53e
Siret, river	45·24n	28·01e
Sitna, river	47·36n	27·08e
Someş, river	47·48n	22·43e
Someşul Cald, river	46·44n	23·22e
Someşul Mare, rïver	47·12n	24·12e
Someşul Mic, rïver	47·09n	23·55e
Someşul Rece, river	46·44n	23·22e
Southern Carpathians, mountains	45·30n	24·15e
Stinisoara, Muntii, mountains	47·10n	26·00e
Subcarpathians, mountains	45·00n	27·00e
Suceava, river	47·32n	26·34e
Sulina, Braţul, stream	45·09n	29·41e
Svinecea, peak	44·48n	22·04e
Tarcăului, Muntii, mountains	46·45n	26·20e
Teleorman, *river*	43·52n	25·26e
Tibleş, Munţii, mountaîns	47·38n	24·05e
Timiş, river	45·28n	20·52e
Tisa, river	47·59n	23·32e
Tisza, river	47·50n	21·00e
Toaca, *peak*	46·59n	25·57e
Transylvania, *historic* region	46·30n	24·00e
Trască, Munţii, mountains	46·23n	23·33e
Trotus, river	46·03n	27·14e
Turnu Roşu, Pasul, pass	45·33n	24·16e
Tutova, river	46·06n	27·32e
Vedea, river	43·53n	25·59e
Vlădeasa, oeak	46·45n	22·48e
Vrancei, Muntii, mountains	46·00n	26·30e
Walachia (Cirnpia Romînă), historic *region*	44·00n	25·00e
Western Carpathians, mountains	46·00n	23·00e
Zăbala, river	45·51n	26·46e
Zărand, Muntii, *mountains*	46·10n	22·15e
Zeletin, river	46·03n	27·23e

kovina to the Soviet Union. A large Romanian population remains in Bessarabia and northern Bukovina, and as long as they remain under Soviet control (Moldavian and Ukrainian Soviet Socialist republics) Romania's national unity cannot be considered complete.

Demographic trends. Within the present frontiers, a growth of population can be traced from 8,600,000 in 1859 to 14,250,000 in 1930, 17,500,000 in 1956, and an estimated 21,245,000 in 1975. By the mid-1970s, more than one-quarter of the population was under 15 years of age, a reduction in the proportion shown in the years of high birth rate immediately following World War II. The bulk of the population—just over 60 percent—was of adult working age (15 to 59 years old), and there was a slight preponderance of females over males. Substantial changes in the social composition of the population have taken place as a result of increasing industrialization, reflected in the rise of the working class population. Similarly, the collectivization of agriculture transformed the rural population, the proportion of peasants with individual households falling from more than half to something less than 5 percent during 1945–70.

Age distribution of population

In education and geographical distribution of the labour force there have been substantial changes since World War II. By the end of the 1960s, for example, official census figures showed a sharp rise in the proportion (some 16 percent) that had achieved some kind of higher education. Geographically, the picture is more complex. Differing rates of economic development in different parts of Romania have produced a movement toward towns and cities, largely for daily and seasonal work. Government planners seek to reduce migration across county boundaries by trying to ensure that each area has its share of development, and the benefits of modernization are consciously spread out over both favoured and unfavoured areas of the country.

This ideal equalization has not yet been attained, however, and the heavy reliance placed on the most advanced regions maintains a modest flow of migrants, and a more substantial flow of temporary workers, coming especially from those backward areas that have relatively high birth rates.

The population density of the country as a whole was by 1970 virtually double that of 1900 although, in contrast to other central European states, there was still considerable room for further growth. The overall density figures, however, conceal considerable regional variation. Population densities are naturally highest in the towns. But in the plains (up to altitudes of some 700 feet) even the rural population has a density of about 250 per square mile (100 per square kilometre), a figure that may rise to as high as 350 (135) in areas with intensive agriculture or a traditionally high birth rate (*e.g.*, northern Moldavia and the "contact" zone with the Subcarpathians). At altitudes of 700 to 2,000 feet the mineral resources, along with the orchards, vineyards, and pastures, support densities of some 75 per square mile (29 per square kilometre).

Variations in density

Settlement types. In 1975 some 57 percent of the population still lived in rural areas. Following administrative reforms of 1968, these areas were regrouped into some 2,700 communes, of which about 150 were classified as suburban. The average population per commune is about 4,500.

The average population for a village is just under 1,000, but an average figure has little relevance because of the sharp regional contrasts in settlement. A dispersed type of rural settlement is generally found in the foothill, ta-

Table 1: Romania, Area and Population

	area		population	
	sq mi	sq km	1966 census	1975 estimate
Districts (*județe*)				
Alba	2,406	6,231	383,000	404,000
Arad	2,955	7,654	481,000	497,000
Argeş	2,626	6,801	530,000	607,000
Bacău	2,549	6,603	598,000	686,000
Bihor	2,909	7,535	586,000	628,000
Bistrița-Năsăud	2,048	5,305	269,000	294,000
Botogani	1,917	4,965	452,000	492,000
Brăila	1,824	4,724	340,000	378,000
Braşov	2,066	5,351	443,000	505,000
Buzău	2,344	6,072	483,000	524,000
Caras-Severin	3,287	8,514	359,000	374,000
Cluj	2,568	6,650	631,000	695,000
Constanta	2,724	7,055	466,000	554,000
Covasna	1,431	3,705	177,000	195,000
Dîmbovița	1,443	3,738	422,000	469,000
Dolj	2,862	7,413	691,000	750,000
Galați	1,708	4,425	474,000	566,000
Gorj	2,178	5,641	299,000	334,000
Harghita	2,552	6,610	282,000	315,000
Hunedoara	2,709	7,016	475,000	519,000
Ialomița	2,398	6,211	363,000	396,000
Iagi	2,112	5,469	617,000	736,000
Ilfov	3,176	8,225	757,000	811,000
Maramureş	2,400	6,215	428,000	492,000
Mehedinți	1,892	4,900	309,000	330,000
Mureş	2,585	6,696	562,000	617,000
Neamt	2,274	5,890	472,000	543,000
Olt	2,126	5,507	477,000	521,000
Prahova	1,812	4,694	699,000	794,000
Sălaj	1,486	3,850	263,000	273,000
Satu Mare	1,701	4,405	359,000	392,000
Sibiu	2,093	5,422	415,000	460,000
Suceava	3,303	8,555	573,000	652,000
Teleorman	2,267	5,872	521,000	543,000
Timig	3,351	8,678	608,000	650,000
Tulcea	3,255	8,430	237,000	263,000
Vaslui	2,046	5,300	432,000	484,000
Vîlcea	2,203	5,705	369,000	408,000
Vrancea	1,878	4,863	351,000	388,000
Municipality (*municipiu*)				
Bucharest	234	605	1,452,000	1,700,000
Total Romania	91,699*	237,500	19,103,000†	21,245,000†

'Converted area figures do not add to total given because of rounding. †Figures do not add to total given because of rounding.
Source: Official government figures.

bleland, and upland regions. The scattered village proper is found at the highest levels and reflects the rugged terrain and the pastoral economic life. Small plots and dwellings are carved out of the forests and on the upland pastures wherever physical conditions permit. Where the relief is less difficult, the villages are slightly more concentrated, although individual dwellings still tend to be scattered among agricultural plots.

The Subcarpathian region, with hills and valleys covered by plowed fields, vineyards, orchards, and pastures and dotted with dwelling places, typically has this type of settlement. The more familiar concentrated villages, marked by uniform clustering of buildings, are to be found in the plains, particularly those given over to cereal cultivation.

Urban settlements

The first urban settlements were situated at points of commercial or strategic significance, and the great majority of present-day towns are either on or in the immediate neighbourhood of the ruins of ancient settlements, whether of fortress or market town. The oldest towns were founded on the Black Sea shores, and urban development only later spread to the plains and then to the mountains. The turbulent history of the country favoured some of these early settlements, which grew into modern towns and cities, while other, once important towns have regressed to become villages or have simply vanished. In the mid-1970s there were 236 towns, of which 47 had the special status of municipality. Many of the towns are small (208 have fewer than 50,000 inhabitants, and 152 have fewer than 20,000), but 15 towns have more than 100,000 inhabitants and apart from the capital (1,706,-818) five exceed 200,000: Cluj–Napoca, Iaşi, Brasov, Galafi, and Timişoara.

THE ECONOMY

Overall development. A program of economic development has transformed the nation since independence. A formerly backward and largely agricultural economy has been transformed into a modern economy, with a strong emphasis on industry.

A very radical land reform was carried out in 1921 (and completed in 1948), although the independence of the peasantry has since been compromised. The restructuring of the economy since the Communist takeover included compulsory collectivization of agriculture, carried out between 1949 and 1962. The means of production were nationalized, including the banks and the main branches of industry, and a system of medium-term central planning was introduced. The result has been an acceleration of economic growth but along lines that had been clearly established by World War II.

Economic growth

The industrial base of the economy has been developed by expanding both those industrial branches that have a raw material base within the country—notably the chemical, power, building materials, and food industries—and those that depend on imported raw materials; the metallurgical industry is probably most important in this respect. Parallel with this process, an attempt has been made to achieve a rational distribution of industrial centres. The growth of new plants has accelerated since the late 1960s (some 1,500 new industrial units were commissioned in the last half of that decade), and this process has been facilitated by the purchase of licenses and patents from firms all over the world.

Economic growth, and in particular the expansion of industry, has required a major program of capital investment. The proportion of national income so allocated rose to about 20 percent in 1960 and to 30 percent in 1970, at which level it is projected to remain for some years. This high rate is thought by many experts to be the only feasible method by which intensive development can take place and reduce the gap that still separates Romania from other industrialized nations. The intensification of industrial activity, however, has caused some problems of air and water pollution, particularly in the case of old industries sited in narrow valleys or low-lying areas. Strict environmental controls have been introduced, although the problem is major only in certain restricted areas.

The role of industry. By the mid-1970s, about 65 percent of the national income was being produced by industry, in marked contrast to the pre-World War II situation, when the percentage was only 30. The main emphasis has been on the development of heavy industry (see Table 2), and the metallurgical, machine-building, metal-processing, and chemical industries have shown the strongest rate of growth.

Coal. The largest reserves are those of bituminous coal; half of Romania's bulk coal production comes from the Petroşani Depression alone. Reserves of poorer quality lignite are being tapped more and more to meet energy requirements. Except for the Baraolt-Vîrghiş Basin, which lies within the Carpathians, most deposits are found along the fringe of the mountain areas. A large lignite field in the Motru Valley (Gorj) supplies some of

Table 2: Industrial Structure of Romania*
(percent)

item	1938	1950	1960	1975
Electric power	1.1	1.9	2.5	2.7
Fuel	16.8	11.3	9.1	3.6
Metallurgy (including ore production)	6.7	7.5	8.4	7.9
Machine building and metal processing	10.2	13.3	24.0	32.4
Chemicals	2.7	3.1	6.1	11.3
Building materials	1.2	2.4	3.2	3.1
Wood operation and processing	9.5	9.9	7.5	4.7
Textiles and ready-made clothes	12.8	18.6	13.5	11.9
Food	32.4	24.2	18.9	13.1
Other	6.6	7.8	6.8	9.3
Total industry	100.0	100.0	100.0	100.0

*For the years 1938, 1950, and 1960, calculated in prices comparable to 1955.

the largest power stations in the country, Rovinari and Turceni.

Oil. Oil deposits are found in the flysch formations that run in a band along the outer rim of the Carpathians and through the Subcarpathians. Deposits in the plains, notably Videle, have been tapped since World War II. Bacău and Prahova districts have long been famous for their oil-refining industry, and they have been joined in recent years by production from Argeg (Pitegti). Natural gases—mainly methane—are produced in the centre of the Transylvanian Plateau, and gases produced as by-products of the oil industry are becoming of increasing importance. Finally it is worth mentioning that oil shales are to be mined at Anina, in the Caraş-Severin district, to supply a new power station to be built at Oravifa.

Hydro-
electric
projects

Electric power. One of the greatest problems facing Romania after World War II, when the Soviet Union demanded the delivery of Romanian petroleum as war reparations, was the very limited development of power stations based on other fuels. Under a plan spanning the years 1951–60 and supplemented by later plans, a remarkable rise in power output took place. The latter totalled 1,100,000,000 kilowatt-hours in 1938, 2,100,000,000 in 1950, and almost 54,000,000,000 by 1975. The foundation for this increase was a series of large power projects, each having 200,000 to 1,000,000 kilowatts' capacity. The most important projects have been the hydroelectric projects of the Argeg, Bistriţa, Danube, and Olt and Lotru rivers and the thermal stations based on Motru lignite. Nuclear power is envisaged, but the time and cost involved in developing a largely independent program have brought delays and completion of a nuclear power station cannot be expected before the 1980s.

The metallurgical industry. Romanian iron industry has particularly strong connections with Galafi as well as with Hunedoara and Reşiţa, where iron was smelted in classical times. Small units exist at Brăila, Cîmpia Turzii (near Turda), Iaşi, Roman, and Tîrgovişte, and a complex is projected for Călăraşi. The nonferrous metallurgical industry, which also dates from the Dacian–Roman period, is largely concentrated in the southwest and west, with copper, gold, and silver production still very active. Aluminum production is a recent development; alumina factories at Oradea and Tulcea supply the aluminum reduction complex at Slatina in the Olt district. Small quantities of lead, mercury, and zinc are also produced.

The machine-building and metal-processing industry is the main branch of the industrial economy, accounting for (by the mid-1970s) nearly a third of bulk industrial production. It provides a good index of the changing priorities in the Romanian economy; before World War II it accounted for only 10 percent of the total, being exceeded in importance by food processing (32 percent) and even the textile and ready-made clothing industry (13 percent). Contemporary centres of production are Bucharest, Bragov, Ploieşti, Cluj, Craiova, Arad, Reşiţa, and many others, with a considerable degree of regional specialization. There has been a strong tendency to concentrate on such modern branches as the electronics industry, as well as to widen and diversify the range of production.

Timber
industry

Other industries. In contrast to metallurgy (which relies on imports of ore and coke to supplement the modest domestic resources), the timber industry can rely on domestic raw materials. In recent decades the emphasis, in what is a traditional industry, has switched from production of sawn timber to finished products. A chain of modern wood industrialization combines turns out a range of products, including furniture and chipboard, which have done well in foreign markets. The building materials industry also utilizes a wide range of resources across the country; cement manufacture represents the most important sub-branch. The main centres are at Turda, Medgidia, Bicaz, Fieni, and Tîrgu Jiu. The long-established textile industry has also undergone a steady development since its radical overhaul in the 1930s. The closely connected ready-made clothing industry has undergone considerable expansion, with a heavy investment in new plant. Finally, the food industry—formerly the foundation of the economy—has been all but eclipsed by the rapid development of other branches. It has, nevertheless, continued to grow in absolute terms, and centres are distributed throughout the country.

Developments in agriculture and fisheries. The natural conditions within Romania make possible a great diversity of agriculture. The resources of the plains, hills, and mountains tend to be complementary, and, despite a strong subsistence element in peasant agriculture, exchanges of staple products are traditional.

Field crops. The climate and relief of the extensive Romanian plains are most favourable to the development of cereal crops, although these are also found in the Subcarpathians and in the Transylvanian Plateau, where they occupy a high proportion of the total arable land. Wheat and corn (maize) are most important, followed by barley, rye, and oats. Two-row barley is cultivated in the Braşov, Cluj, and Mureş areas, where it is used for brewing. The tendency is for the acreage of cereals to fall as yields increase and industrial crops require more land. In 1975, for example, cereals, dominated by wheat and maize, were planted on 64.1 percent of the total area of arable land; in 1938, by contrast, they occupied almost 85 percent.

Vegetables. Vegetables—peas, beans, and lentils—are planted on a relatively small area. Peas are the predominant crop; being capable of early harvest, they allow a second crop, usually fodder plants, to be grown on the same ground. Vegetable cultivation is particularly marked around the city of Bucharest, with specialization in the production of early potatoes, tomatoes, onions, cabbages, and green peppers. Similar gardening areas are found around Timigoara, Arad, Craiova, Galafi, Brăila, and other cities. The most important potato-growing areas are Bragov, Sibiu, Harghita, and Mureş districts. Other related crops include sugar beets; sunflower seed, mostly on the Danube, Tisa, and Jijia plains; and hemp, flax, rape, soybeans, and tobacco.

Wines

Viticulture. With some 730,000 acres (295,000 hectares) under cultivation by the mid-1970s, Romania can be counted among the main wine-producing countries of Europe. It specializes in the production of high-quality wines, using modern methods; with the growth of the tourist trade, its wines are becoming known to, and appreciated by, a larger international public. Large quantities are exported annually. The major vineyards are at Odobeşti, Panciu, and Nicoregti, with a half-dozen or more other major centres. Both white and red wines have won various international awards.

Orchards. At an altitude of between 1,000 and 1,600 feet (300 and 500 metres), orchards are found on almost all the hillslopes on the fringe of the Carpathians. There is increasing specialization in fruits with a high economic yield, and there are plans to extend the orchards, particularly those of apples and pears, into higher altitudes. Orchards have solved problems of soil erosion on many unstable hillsides.

Livestock. Livestock raising has a very long history in Romania. Sheep can be raised wherever grass is available, whether in the alpine pastures or the Danube plain and valley. Though there are only around 6,000,000 cattle, about half of these are beef and important to exports. Bees are also raised throughout the country, and silkworm production retains a modest importance despite the introduction of man-made fibres. Silk, the weaving of which was long the occupation of peasant women in the south and southwest, has lent much to the beauty of local folk costumes, especially the richly embroidered blouses and headscarves.

Fishing. The rivers of Romania, its lakes—especially the group around Razelm—and its Black Sea coastal region support a well-developed fishing industry. The largest quantity of fish is obtained from the Danube and its delta, and about 80 percent of the annual catch is consumed fresh. The canneries that process the remainder, especially the marine species, are located at Tulcea, Constanţa, and Galafi. Ocean fishing in foreign waters is developing rapidly to supplement the production from home waters and allow more meat to be exported.

Transportation. Railways provide the main method of transportation for both freight and passengers in Romania. Since World War II, diesel and electric motors have been placed in service, and the major lines have been electrified. Romania also has a system of national roads, the majority of which have been brought up to modern standards. The main lines of communication tend to focus on Bucharest and include many scenic routes. The country has maritime connections with many countries, and the port of Constanfa, undergoing major expansion, plays a major role in the national economy. Finally, the Danube River continues to be a major transportation route, which eventually will be connected with Constanfa by the Danube–Black Sea Canal.

Bucharest is the main centre for air transportation. In addition to local routes, its international traffic—again aided by the growing tourist trade—has been growing in significance. The great majority of flights by the national airline (Tarom, derived from Transporturile Aeriene Române) are to Europe, North Africa, and the Middle East, but there are also services to the U.S. and China.

Role in Comecon

Foreign trade and tourism. The modernization of the Romanian economy has resulted in a considerable upsurge in its foreign trade and commercial contacts, which by the mid-1970s involved more than 100 countries. The nation has also taken part in international fairs and exhibitions. Romania is an active participant in Comecon, the Communist bloc international trade group, under the Romanian policy of "Socialist internationalism." Great attention has also been paid to broadening trade with the developing countries. Total foreign trade in fact increased by more than two and a half times during the 1960s and doubled in the first five years of the 1970s, and there have been radical changes in exports, notably toward emphasis on machinery, industrial equipment, and other durable goods.

Tourism has become of special significance to Romania, with more than 3,800,000 persons a year visiting the country in the mid-1970s. Tourist attractions range from winter sports in the mountains to summer seaside activities in the resort belt fringing the Black Sea, with health spas receiving special emphasis. The Danube Delta, too, has become increasingly popular because of the growing worldwide interest in ecology and conservation. Special features of interest to tourists include the mountain lakes and underground cave systems that are features of the Carpathians and the fine churches and monasteries, with frescoes dating from the 14th to the 16th century, that are found in northern Moldavia. More generally, the folk costumes and the ancient folklore of Romanians, notably in the Carpathian Mountains, provide a reminder of the country's long traditions. Foreign tourists have been encouraged by much-improved hotels and by favourable tourist rates of exchange. Compulsory currency exchange regulations and a prohibition on the use of private accommodation, however, prove frustrating to the individual tourist.

The financial system. The basic financial vehicle for Romanian economic policy is the state budget presented annually to the National Assembly. By the mid-1970s, 19 percent of the budget income was derived from the profits of state enterprises, the remainder coming from taxation and insurance. Some 66 percent of annual budget expenditure goes toward financing economic development, some 22 percent for state services and cultural activity, about 5 percent for national defense, and less than 2 percent for administration. The National Bank of Romania, founded in 1880, is the heart of the banking system, managing budgetary cash resources and issuing currency. It also establishes foreign exchange rates and engages in foreign exchange operations. It is supported by an investment bank, which finances the investment projects of all state and cooperative organizations; an agricultural bank; and a foreign trade and loan bank, which also handles the money incomes--deposited as current and savings accounts—of individual citizens.

Prices

Lnterest rates do not: reflect scarcity of money or the element of risk; they are used by the government as one of the economic levers intended to motivate enterprises toward greater efficiency; penalties are built into the system to allow discrimination against enterprises that are poorly managed. Prices, too, are set arbitrarily. They have tended to assure high profits for many enterprises and provide resources from which unprofitable enterprises can be subsidized. Prices of agricultural commodities and other raw materials have usually been set low. In this way the price system has served to transfer resources from agriculture to industry and keep consumption low for the benefit of investment. This strategy, however, gives industrial enterprises little incentive to cut costs, and the government's drive for economic efficiency is thereby compromised.

ADMINISTRATION AND SOCIAL CONDITIONS

Constitutional and political framework. The constitutional framework derives from the state constitution adopted in 1965, which characterized Romania as a Socialist republic. The constitution gives equal rights to all citizens, without regard to nationality, sex, or religion, and state power is said to rest on an alliance of peasants, workers, and intellectuals. The National Assembly is the supreme organ of state power, but in the intervals between its sessions the State Council, composed of a president, four vice presidents, and 22 members, exercises supreme power.

The administration of state affairs is in the hands of the Council of Ministers, which, on the authority of the National Assembly, acts as the national executive. Local people's councils are elected at the district level. Judicial functions are headed by the Supreme Court—elected by the National Assembly—which supervises the activities of all district courts, lawcourts, and military tribunals.

The Communist Party

The most fundamental fact of the political system, however, is the constitutional status of the Communist Party of Romania (Partidul Comunist Romhn) as the leading force of society and of every organized group within it. The CPR was founded in 1921 and pursued an underground existence from 1924 to 1944. In 1948 the Communist Party merged with the Social Democratic Party to form the Romanian Workers Party (Partidul Muncitoresc Romhn), which in 1965 reverted to its original name. Other political parties, including the Liberal and Peasant parties that previously dominated the political life of the country, have been suppressed, although exiles continue to propagate their philosophies. Left-wing parties survived but lost their independent existence through amalgamation into a large grouping dominated by the Communists.

Education. With the exception of certain university courses, education is free and universal in Romania, and its development has been a key to the economic transformation of the country in modern times and to the gradual elimination of illiteracy. One in five of the country's inhabitants is a pupil or student, whether at the obligatory general school; the middle schools (general or specialized, four or five years); or the wide range of professional and technical schools and institutes of higher education. Associated with this educational system is an extensive national library network.

The major institution of academic research is the Academia Republicii Socialiste Romhnia (Academy of the Socialist Republic of Romania), which traces its origins to the Societatea Literară Română founded in 1866. Its publishing house, Editura Academiei, produces research papers and more than 75 journals. The academy's library contains almost 7,500,000 volumes and is the national depository for all Romanian and United Nations publications. The Central State Library, founded in 1955, is the copyright depository, with more than 6,800,000 volumes and periodicals.

Of the institutions of higher learning, the University of Bucharest was founded in 1864. Other important schools are Babes–Bolyai University in Cluj and the Gheorghe Gheorghiu–Dej Technical Institute in Bucharest.

Health services. As in other Communist countries, medical care is provided free by the state. The quality of service has improved with the training of more doctors and the construction of new hospitals in the main towns and administrative centres and with the new drugs that have become available from the country's growing pharmaceutical industry. The antibiotics factory built at Iaşi in 1955 is a particularly important postwar achievement. Facilities have also been improved in rural areas; before World War II, doctors were available only in the main villages, but now every commune has its clinic and dispensary. Moreover, to reduce the formerly high rate of infant mortality, all babies are delivered in hospitals. Several new hospitals in the mountains serve tuberculosis patients. Overall there has been a considerable increase in life expectancy. The standard of living, however, is still relatively poor by European and North American standards, and many kinds of sophisticated medical treatment are not widely available.

CULTURAL LIFE AND INSTITUTIONS

The cultural milieu. The authorities have emphasized the need to bring the broad mass of the populace in contact with the nation's contemporary culture and with its heritage. This includes emphasis on open public expression of varying viewpoints on cultural life, as well as access to the works of culture. Sadly, this policy has been gravely compromised in the Communist period by party control of all cultural activity and the consequent necessity for all contributors to the national culture to support party prescriptions. This control is exercised through the Council on Socialist Culture and Education, a government ministry, and the professional unions to which all practicing artists must belong.

Emphasis on mass cultural activity

Cultural institutions and the folk heritage. There are more than 150 institutions of mass cultural education putting on performances, often in remote regions and sometimes in the languages of minority ethnic groups (Hungarian, German, Ukrainian, etc.). They include theatres and puppet shows, operas, music hall shows, song and dance ensemble productions, and musical performances ranging from folk music to symphony concerts. The institutions in which these performances take place are village clubs, houses of culture, and clubs run by trade unions and other mass organizations. The "people's universities" in both towns and villages also emphasize mass cultural life.

The number of museums has undergone a dramatic increase, a third of the 1979 total of more than 300 having been set up during the 1966–70 Five-Year Plan alone and nearly 70 more added during 1971–75. The film industry, present in Romania since 1912, is controlled by the state. Two motion-picture studios in the Bucharest area produce documentaries and some feature-length films. There are more than 6,000 cinemas in the country; a special feature is the village film festival of the winter months.

In spite of these modern developments, Romania still offers a variety of customs, traditions, and forms of folk art. Wood carvings, brightly ornamented costumes, skillfully woven carpets, pottery, and other elements of traditional Romanian culture remain popular and, with the onset of tourism, have become known internationally. Folk art is characterized by abstract or geometric designs and stylized representations of plants and animals. In embroidery and textiles, designs and colour schemes can be associated with particular regions of the country. Special folk arts of Romania are the decoration of highly ornamental Easter eggs and painting on glass, which, however, is becoming a lost skill. Folk music includes dance music, laments and ballads, and pastoral music. Major instruments are the violin, the *cobza* (a stringed instrument resembling a lute), the *ţambal* (a dulcimer played with small hammers), and the flute. Folk melodies are preserved in the music of modern Romanian composers such as Georges Enesco.

The Romanian language, although developing over the centuries in difficult historical conditions, is as Latin as any other Romance language and, like the culture as a whole, continues to exhibit a remarkable vitality. The fact is perhaps paralleled by some of the Modernist tendencies in the Romanian fine arts; the sculptor Constantin Brancusi, a promoter of absolute Modernism coupled with a firm sense of classical Mediterranean values, had great international influence early in the 20th century. Romanian poets and writers, too, have operated in a cultural tradition somewhat different from that in neighbouring countries; in architecture, the elegant new Bucharest television centre is but one example of another Modernist trend.

THEOUTLOOK

Romanian prospects for the 1980s, although obviously subject to international conditions, rest basically on the long-term economic plans worked out for the country. These call for continued expansion of foreign trade, introduction of newer and better techniques throughout the economy, and concentration on consumer products to supplement the continued emphasis on industrialization. The emphasis on technical efficiency will be reflected throughout the economy. Culturally the prospects are less promising, with the "détente" envisaged under the Helsinki agreement (1975) being frustrated by growing ideological activity and very strict controls on the issue of passports.

BIBLIOGRAPHY. General and geographical works include H.J. FLEURE and R.A. PELHAM (eds.), *Eastern Carpathian Studies: Roumania* (1936); ANDRE BLANC, PIERRE GEORGE, and HENRI SMOTKINE, *Les Républiques socialistes d'Europe centrale* (1967); TIBERIU MORARIU, VASILE CUCU, and ION VELCEA, *Géographie de la Roumanie* (1966; Eng. trans., 2nd ed., 1969); VICTOR TUFESCU, *România* (1974); ION SANDRU, *România: geografie economica* (1975); IAN M. MATLEY, *Rornania: A Profile* (1970). Other works are listed in STEPHEN A FISCHER-GALATI (comp.), *Rumania: A Bibliographic Guide* (1969). Information on the physical geography of the Romanian territory may also be found in RAUL CALINESCU et al., *Biogeografia Romciniei* (1969); PETRE GISTESCU, *Lacurile din R.P.R.: geneză şi regim hidrologic* (1963), on the lakes of Romania; VINTILA MIHAILESCU, *Geografia fizică a Romciniei*, vol. 1 (1970); and VICTOR TUFESCU, *Subcarpaţii şi depresiunile marginale ale Transilvaniei* (1966). Basic aspects of economic and human geography are discussed in NICOLAE A RADULESCU, ION VELCEA, and N. PETRESCU, *Geografia agriculturii României* (1968), with French summary; HENRY L. ROBERTS, *Rumania: Political Problems of an Agrarian State* (1951, reprinted 1969); J.M. MONTIAS, *Economic Development in Communist Rumania* (1967); ERVIN HUTIRA, *The Development of the National Economy in the Rumanian People's Republic* (1963); M. PEARTON, *Oil and the Romanian State* (1971); E. DOBRESCU and I. BLAGA, *Structural Patterns of Romanian Economy* (1973); M. CONSTANTINESCU et al., *Urban Growth Processes in Romania* (1974); VIOLETTE REY, *La Roumanie: Essai d'analyse régionale* (1975); D. TURNOCK, *An Economic Geography of Romania* (1974); and VASILE CUCU, *Oraşele României* (1970). Attention is also drawn to the *Atlasul Republicii Socialiste România*, which is being produced in installments, and the statistical annual *Anuarul statistic al R.S.R.*

(V.S.C./D.T.)

Roman Law

The term Roman law denotes first of all the law of the city of Rome and of the Roman Empire: in the West, the law in force at any period from the foundation of the city (traditional date 753 BC) until the fall of the Western Empire in the 5th century AD, and the fall'ofthe Eastern Empire in 1453.

The range of influence of Roman law

The term Roman law today, however, often refers not merely to the law of those political societies to which the name Roman may in some sense be applied; for the legal institutions evolved by the Romans have had influence on the law of other peoples in times long after the disappearance of the Roman Empire as a political entity, and even in countries that were never subject to Roman rule. To take the most striking example, in a large part of Germany, until the adoption of a common code for the whole empire in 1900, the Roman law was in force as "subsidiary law"; that is, it was applied unless excluded by contrary local provisions. This law, however, which was in force in parts of Europe long after the fall of the Roman Empire, was not the Roman law in its original form. Its basis was indeed always the *Corpus Juris Civilis*—the codifying legislation of the emperor Justinian I (see below The law of Justinian)—but from the 11th century this legislation was interpreted, developed, and adapted to later conditions by generations of jurists and received additions from non-Roman sources. All of the forms it assumed in different countries and at different epochs can be included under Roman law.

Roman law is important not only as the system once in force in many places but also as an influence on the development of law in general. Even today, the legal systems of Western civilization (with some exceptions, especially the Scandinavian) fall into two groups, in one of which the main elements are of Roman origin; the other is the English common law, which itself is not without Roman influences. To the English group belong England, nearly all of the United States of America, and most British territories; to the Roman group belong the rest. Nearly all of the nations of continental Europe have legal codes that are Roman in structure, fundamental categories, and general method of thought. Within the British territories, Scotland has a system largely derived from the Roman; Quebec and most countries of Latin America have systems of French law built largely with Roman materials; and South Africa has a system known as "Roman–Dutch," based on Roman law as developed by Dutch jurists.

This article deals with Roman law as it developed in Rome up to the age of the emperor Justinian (reigned 527–565). (Subsequent developments in the Eastern Empire were mainly modifications of Justinian's codes.)

SOURCES OF ROMAN LAW

Development of the jus civile and jus gentium. In the great span of time during which the Roman Republic and Empire existed, there were naturally many phases of development. During the period of the Republic (753–31 BC), there first developed the *jus civile* (civil law), which was based on custom or legislation and which applied exclusively to Roman citizens. By the mid-3rd century BC, however, there had developed another type of law, *jus gentium*, which the Romans applied both to themselves and to foreigners. *Jus gentium* was not the result of legislation but rather was developed by the magistrates and governors who were responsible for administering justice in cases in which a foreigner was involved. The *jus gentium* became, to a large extent, part of the massive body of magisterial or praetorian law that was developed by the magistrates for citizens as well as for foreigners as a flexible alternative to *jus civile.*

Roman law, like other ancient systems, adopted originally the principle of personality—that is, that the law of the state applied only to its citizens. The foreigner was strictly rightless and, unless protected by some treaty between his state and Rome, could be seized like an ownerless piece of property by any Roman. But from early times there were treaties with foreign states guaranteeing mutual protection. Even in cases in which there was no treaty, the increasing commercial interests of Rome forced it to protect, by some form of justice, the foreigners who came within its borders. A magistrate could not simply apply Roman law because that was the privilege of citizens; even had there not been this difficulty, foreigners, especially those coming from Greek cities and used to a more developed and freer system, would probably have objected to the cumbrous formalism that characterized the early *jus civile.*

The law that the magistrates applied probably consisted of three elements: (1) an already existing mercantile law used by the Mediterranean traders; (2) those institutions of the Roman law that, after being purged of their formalistic elements, could be applied universally to any litigant in a Roman court, whether Roman or foreigner; (3) in the last resort, a magistrate's own sense of what was fair and just. This system of *jus gentium* was also adopted when Rome began to have provinces and provincial governors administered justice to the *peregrini* (foreigners). This word came to mean not so much persons living under another government (of which, with the expansion of Roman power, there came to be fewer and fewer) as Roman subjects who were not citizens. In general, disputes between members of the same subject state were settled by that state's own courts according to its own law, whereas disputes between provincials of different states or between provincials and Romans were resolved by the governor's court applying *jus gentium*. By the 3rd century AD, when citizenship was extended throughout the empire, the practical differences between *jus civile* and *jus gentium* ceased to exist. Even before this, when a Roman lawyer said that the contract of sale was *juris gentium*, he meant that it was formed in the same way and had the same legal results whether the parties to it were citizens or not. This became the practical sense of *jus gentium*. Because of the universality of its application, the idea was also linked with a theoretical sense, that of a law common to all peoples and dictated by nature—an idea that the Romans took from Greek philosophy.

Written and unwritten law. The Romans themselves divided their law into *jus scriptum* (written law) and *jus non scriptum* (unwritten law). By "unwritten law" they meant custom; by "written law" they meant not only that derived from legislation but, literally, that which was based on any source in writing.

Types of written law

There were various types of written law, the first of which consisted of leges (singular lex), enactments of one of the assemblies of the whole Roman people. The wealthier classes, or patricians, dominated these assemblies; and the common people, or plebeians, therefore had their own council, wherein they enacted resolutions called *plebiscita*. Only after the passage of the Lex Hortensia in 287 BC did *plebiscita* become binding on all classes of citizens; thereafter, *plebiscita* were generally termed leges along with other enactments. In general, legislation was a source of law only during the republic. When Augustus established the empire in 31 BC, the assemblies did not at once cease to function, but their assent to any proposal was a mere formal ratification of the emperor's wishes. The last known lex was passed under the emperor Nerva (AD 96–98).

It should be mentioned that the earliest and most important legislation or body of leges was the Twelve Tables, enacted in 451–450 BC during the struggle of the plebeians for political equality. It represents the desire to obtain a written and public code that patrician magistrates could not alter at will against plebeian litigants. Little is known of the actual content of the Twelve Tables; the text of the code has not survived, and only a few fragments are extant, collected from allusions and quotations in the works of such authors as Cicero. From the fragments it is apparent that numerous matters were discussed, among them family law, delict (tort, or offense against the law), and legal procedure.

A second type of written law consisted of the *edicta* (edicts), proclamations issued by a superior magistrate (praetor) on judicial matters. The office of praetor was created in 367 BC to take over the expanding legal work involving citizens; later, a separate praetor was created to deal with foreigners. Upon entering office, a praetor issued an edict that was, in effect, his program of his year in office. In addition, the curule aediles, who were the magistrates responsible for the care and supervision of the markets, also issued edicts. By the later part of the republic, these praetorian and magisterial edicts had become an instrument of legal reform, and leges ceased to be a major source of private law.

The Roman system of procedure gave the magistrate great power over the provision or refusal of judicial remedies, as well as over the form that such remedies were to take. The result was a new body of rules that existed alongside, and often superseded, the civil law. The *edicta* remained a source of law until about AD 131, when the emperor Hadrian commissioned their reorganization and consolidation and declared the result to be unalterable except by the emperor himself.

A third type of written law was the *senatus consulta,* resolutions of the Roman senate that had no legislative force during the Republic but were instead suggestions to various magistrates that could be given force by the magistrates' edicts. In the early empire, as the assemblies declined and the position of the emperor increased, *senatus consulta* became resolutions endorsing the proposals of the emperor. The approval of the senate became increasingly automatic, for the emperor's proposals were the true instrument of power. Consequently, the emperors ceased referring proposals to the senate and, not long after the early imperial period, ended the practice of legislating through the senate.

A fourth type of law consisted of the constitutiones *principum*, which were, in effect, expressions of the legislative power of the emperor. By the middle of the 2nd century AD, the emperor was, essentially, the sole creator of the law. The chief forms of imperial legislation were edicts or proclamations; instructions to subordinates, especially provincial governors; written answers to officials or others who consulted the emperor; and decisions of the emperor sitting as a judge.

The last type of written law was the *responsa prudentium*, answers on legal questions given by learned lawyers to those who consulted them. Although law, written and unwritten, was originally a rather secretive monopoly of the pontiffs, a college of priests, there had developed by the early 3rd century BC a recognizable class of legal advisors, *juris consulti* or *prudentes*, who were not professionals as such but men of rank who sought popularity and advancement in a public career by giving free legal advice. They interpreted statutes and points of law, especially unwritten law, and advised the praetor on the content of his edict and assisted parties and judges in litigation. Augustus empowered certain jurists to give *responsa* with the emperor's authority, and this increased their prestige, although this practice lapsed as early as AD 200.

During the early empire, numerous commentaries were written by the great jurists on individual leges, on the civil law, on the edict, and on law as a whole. In the 5th century, a law was passed stipulating that only the works of certain jurists could be cited. Postclassical law produced little in comparison to the classical world.

The law of Justinian. When the emperor Justinian I assumed rule in AD 527, he found the law of the Roman Empire in a state of great confusion. It consisted of two masses that were usually distinguished as old law and new law.

Old law and new law under Justinian

The old law comprised (1) all of the statutes passed under the republic and early empire that had not become obsolete; (2) the decrees of the senate passed at the end of the republic and during the first two centuries of the empire; (3) the writings of the jurists and, more particularly, of those jurists to whom the emperor had given the right of declaring the law with authority; these jurists, in their commentaries, had practically incorporated all that was of importance. All of these records and writings, however, were very numerous; many had become exceedingly scarce or had been altogether lost, and some were of doubtful authenticity. They were so costly that even the public libraries did not contain complete collections. Moreover, these writings contained many inconsistencies.

The new law, which consisted of the ordinances of the emperors promulgated during the middle and later empires, was in a condition not much better. These ordinances or constitutions were extremely numerous and conflicting. No complete collection existed, for the earlier codices did not include all of them; others had to be obtained separately, and many of them were probably inaccessible to a private person. It was thus necessary to collect into a reasonable corpus so much of the law, both new and old, as was regarded as binding and to purge away its contradictions and inconsistencies.

Immediately after his accession, Justinian appointed a commission to deal with the imperial constitutions. The commissioners, ten in number, went through all of the constitutions of which copies existed, selected such as were of practical value, cut these down by retrenching all unnecessary matter, got rid of contradictions by omitting one or the other of the conflicting passages, and adapted all the provisions to the circumstances of Justinian's own time. The resulting Codex Constitutionum was formally promulgated in 529, all imperial ordinances not included in it being repealed. This Codex has been lost, but a revised edition of 534 exists as part of the so-called Corpus Juris Civilis.

The success of this first experiment encouraged the Emperor to attempt the more difficult enterprise of simplifying and digesting the writings of the jurists. Thus, beginning in 530, a new commission of 16 eminent lawyers set about this task of compiling, clarifying, simplifying, and ordering; and the results were published in 533 in 50 books that became known as the Digest (*Digesta*) or Pandects. In enacting the Digest as a lawbook, Justinian repealed all of the other law contained in the treatises of the jurists and directed that those treatises should never be cited in the future even by way of illustration; at the same time, he abrogated all of the older statutes that had formed a part of the old law.

About the same time, there was published an outline of the elements of Roman law called the Institutes of Justinian (or simply *Institutiones*). In addition, between 534 and his death in 565, Justinian himself issued a great number of ordinances, dealing with all sorts of subjects and seriously altering the law on many points. These ordinances are called, by way of distinction, new constitutions (Novellae constitutiones) and in English are called the Novels.

Corpus Juris Civilis: the civil law code of Justinian

All of these books—the revised Codex Constitutionum, the *Digesta*, the *Institutiones*, and the Novellae—are collectively known as the Corpus Juris Civilis. This Corpus Juris of Justinian, with a few additions from the ordinances of succeeding emperors, continued to be the chief lawbook of the Roman world. In the 9th century a new system that is known as the Basilica was prepared by the emperor Leo VI the Wise. It is written in Greek and consists of parts of the Codex and of the Digest, joined and often altered in expression, together with some matter from the Novels and imperial ordinances posterior to Justinian. In the western provinces, the law as settled by Justinian held its ground.

CATEGORIES OF ROMAN LAW

The law of persons. Slavery. "The main distinction in the law of persons," said the 2nd-century jurist Gaius, "is that all men are either free or slaves." The slave was, in principle, a human chattel who could be owned and dealt with like any other piece of property. As such, he was not only at the mercy of his owner but rightless and (apart from criminal law) dutiless. But if the slave was in law a thing, he was in fact a man, and this modified the principle. A slave could not be a party to a contract nor own property, but he could be given a de facto patrimony, which could be retained if he were freed; if he made a "commitment," it could ultimately be enforced against his master. A slave could be manumitted and became, in most instances, not only free but a citizen.

Citizenship. The definition of citizenship was important for the purposes of private law because certain parts of private law applied only to citizens (jus *civile*). Noncitizens could be either Latini, inhabitants of Roman settlements that had the rights of members of the original Latin League, or peregrini, who were members of foreign communities or of those territories governed but not absorbed by Rome. The great extension of the citizenship by the emperor Caracalla in AD 212 reduced the importance of this part of the law.

Patria potestas: the power of the paternal head of the Roman family

Family. The chief characteristic of the Roman family is the *patria* potestas (paternal power in the form of absolute authority), which the elder father exercised over his children and over his more remote descendants in the male line, whatever their age might be, as well as over those brought into the family by adoption—a common practice at Rome. This meant originally not only that he had control over the persons of his children, amounting even to a right to inflict capital punishment, but that he alone had any rights in private law. Thus, any acquisitions made by a child under potestas became the property of the father. The father might indeed allow a child (as he might a slave) certain property to treat as his own, but in the eye of the law it continued to belong to the father.

By the 1st century AD there were already modifications of the system: the father's power of life and death had shrunk to that of light chastisement, and the son could bind his father by contract with a third party within the same strict limits that applied to slaves and their masters.

Sons too could keep as their own what they earned as soldiers and even make wills of it. In Justinian's day, the position as regards property had changed considerably; what the father gave to the son still remained in law the father's property, but the rules concerning the son's own earnings had been extended to many sorts of professional earnings; and in other acquisitions (such as property inherited from the mother), the father's rights were reduced to a life interest (usufruct). At all times, *patria potestas* ceased normally only with the death of the father; but the father might voluntarily free the child by emancipation, and a daughter ceased to be under her father's *potestas* if she came under the *manus* of her husband.

Marriage with and without *manus,* or husbandly authority

There were two types of marriage known to the law, one with *manus* and one without, but the *manus* type was rare already in the late republic and had disappeared long before Justinian's day. *Manus* was the autocratic power of the husband over the wife, corresponding to *patria potestas* over the sons.

Marriage without *manus* was by far the more common in all properly attested periods. It was formed (provided the parties were above the age of puberty and, if under *potestas,* had their father's consent) simply by the beginning of conjugal life with the intention of being married, and this was normally evidenced by the bringing of the bride to the bridegroom's house. The wife remained under her father's *potestas* if he were still alive; if he were dead, she continued (so long as guardianship of women continued) to have the same guardian as before marriage. Both spouses had to be citizens, or if one was not, he or she must have *conubium* (the right, sometimes given to non-Romans, of contracting a Roman marriage). In marriage without *manus,* the property of the spouses remained distinct, and even gifts between husband and wife were invalid.

Divorce was always possible at the instance of the husband in cases of marriage with *manus*; in marriage without *manus,* either party was free to put an end to the relationship at will. A formal letter was usual, but any manifestation of intention to end the relationship—an intention made clear to the other party and accompanied by actual parting—was all that was legally necessary. The Christian emperors imposed penalties on those who divorced without good reason, but the power of the parties to end the marriage by their own act was not taken away.

Concubinage was recognized in the empire as a "marriage" without a dowry, with a lower status for the woman, and with provisions that the children were not legally the father's heirs. **A** man could not have both a wife and a concubine. The emperor Constantine in the 4th century first enacted that the children of such unions might be legitimated by the subsequent marriage of their parents, a rule that the medieval civil law extended to all illegitimate children.

Guardianship

Persons under the age of puberty (14 for males, 12 for females) needed *tutores* if they were not under *patria potestas.* Such tutors could be appointed under the will of the father or male head of the household; failing such appointment, the guardianship went to certain prescribed relatives; if there were no qualified relations, the magistrates made an appointment. Originally, children were considered adult at the age of puberty; but, by a long development, it became usual for those above puberty and under 25 to have guardians who were always magisterially appointed. Originally, all women not under *patria potestas* or *manus* similarly needed *tutores,* appointed in the same way as those for children. By the early empire, this provision was little more than a burdensome technicality, and it disappeared from Justinian's law.

Corporations. The Romans did not develop a generalized concept of juristic personality in the sense of an entity that had rights and duties. They had no terms for a corporation or a legal person. But they did endow certain aggregations of persons with particular powers and capacities, and the underlying legal notion hovers between corporate powers, as understood in modern law, and powers enjoyed collectively by a group of individuals. The source of such powers, however' was always an act of state,

Four types of corporation may be distinguished:

1. *Municipia* (the citizen body, originally of the conquered cities and later of other local communities) possessed a corporateness that was recognized in such matters as the power to acquire things and to contract. In imperial times, they were accorded the power to manumit slaves, take legacies, and finally—though this became general only in postclassical law—to be instituted heir.

2. The *populus Romanus,* or the "people of Rome," collectively could acquire property, make contracts, and be appointed heir. Such property included property of the treasury.

3. *Collegia—private* associations with specialized functions, such as craft or trade guilds, burial societies, and societies dedicated to special religious worship—seem to have carried on their affairs and to have held property corporately in republican times and were apparently numerous. The emperors, viewing the *collegia* with some suspicion, enacted from the beginning that no *collegium* might be founded without state authority, and such rights as the manumission of slaves and the taking of legacies were closely regulated.

4. Charitable funds became a concern of postclassical law. Property might be donated or willed—normally, but not necessarily, to a church—for some charitable use, and the church would then (or so it appears from the evidence) have the duty to supervise the fund. Imperial legislation controlled the disposition of such funds so that their alienation in excess of powers was void. Ownership is thought in such cases to have been vested in the administrators for the time being.

The Roman concept of absolute ownership

The law of property and possession. In Roman law (today as well as in Roman times), both land and movable property could be owned absolutely by individuals, This conception of absolute ownership (*dominium*) is characteristically Roman, as opposed to the relative idea of ownership as the better right to possession, which underlies the Germanic systems, including English law.

Methods of acquisition by the jus civile. *Mancipatio* (or formal transfer of property) involved a ceremonial conveyance needing for its accomplishment the presence of the transferor and transferee, five witnesses (Roman citizens of full age), a pair of scales, a man to hold them, and an ingot of copper. The transferee grasped the object being transferred and said, "I assert that this thing is mine by Quiritarian [Roman] law; and be it bought to me with this piece of copper and these copper scales." He then struck the scales with the ingot, which he handed to the transferor "by way of price." Clearly, this was a symbolical sale and the relic of a real sale.

In jure cessio was a conveyance in the form of a lawsuit. The transferee claimed before the magistrate that the thing was his, and the transferor, who was the defendant, admitted the claim. The magistrate then adjudged the thing to the transferee. (The sham-lawsuit theory, however, is not acceptable to all modern scholars, principally because the judgment of ownership was valid against any possible private claimant, not merely against the defendant, as in a true lawsuit.)

Usucapio referred to ownership acquired by length of possession. In early Roman law, two years' continuous possession gave title in the case of land, one year in the case of movables. In the developed law, possession must have begun justifiably in good faith, and the thing must not have been stolen (even though the possessor himself may have been quite innocent of the theft) or occupied by violence.

Methods of acquisition by the jus gentium. In terms of *occupatio,* ownerless things, provided that they were susceptible to private ownership (excluding such things as temples), became the property of the first person to take possession of them. This applied to things such as wild animals and islands arising in the sea. In some views, it also applied to abandoned articles.

Accessio worked in this manner: if an accessory thing belonging to A was joined to a principal one belonging to B, the ownership in the whole went to B. For example, if A's purple were used to dye B's cloth, the dyed cloth belonged wholly to B. By far the most important application of this rule asserted that whatever is built on land becomes part of the land and cannot be separately owned.

Specificatio was somewhat different. If A made a thing out of material belonging to B, one school of thought held that ownership went to A, and another held that it remained in B. Justinian adopted a "middle opinion," according to which B retained ownership if reconversion to the original condition was possible (a bronze vase can be melted down); A obtained ownership if it was not (wine cannot be reconverted into grapes).

According to *thesauri inventio* or treasure trove, the final rule was that if something was found by a man on his own land, it went to him; if it was found on that of another, half went to the finder, half to the landowner.

Traditio was simple delivery of possession with the intention of passing ownership and was the method of conveyance of the *jus gentium.* If A sold and merely delivered a slave to B, *A* under the *jus civile* remained owner of the slave until a specified length of time had elapsed. The praetors, however, devised procedural methods of protecting B's possession in such a way that A's title became valueless, and B was said to have the thing *in bonis.* This was a remarkable triumph for informality in the granting of title. From the phrase *in bonis,* later writers coined the expression "bonitary ownership." Justinian abolished the theoretical distinction between civil and bonitary ownership.

Forms of property in land other than ownership. The ordinary leaseholder had no protection beyond a contractual right against his landlord, and he could not assign his tenancy. But there were certain kinds of tenure that did provide the tenant protection and that were assignable: there were agricultural and building leases granted for a long term or in perpetuity, whereby the leaseholders often enjoyed rights hardly distinguishable from ownership.

Servitudes: rights to specific, limited use or enjoyment of another's property

There were also servitudes, in which one person enjoys certain rights in property owned by another. *Ususfructus* was the right to use and take the fruits (such as crops) of a thing and corresponded to the modem notion of life interest. *Usus* was a more restricted right, likewise not extending beyond the life of the holder but merely to the use of a thing; thus, a person could live in a house himself but could not let it, as that would be equivalent to taking the fruits.

Since ownership was absolute, it was sharply distinguished from possession, which the civil law did not protect as such. Any owner wishing to interfere with an existing possessor, however, must bring an action and prove his title. If he interfered on his own authority, the praetor would see that the original state of affairs was restored before adjudication upon the title.

Obligations. Obligations were classified by the classical jurists into two main categories, according to whether they arose from delict or contract. Justinian's law recognized two further classes of obligation termed quasi-delict and quasi-contract.

Delict. As early as the 6th and 5th centuries BC, Roman law was experiencing a transition from a system of private vengeance to one in which the state insisted that the person wronged accept compensation instead of vengeance. Thus, in the case of assault *(injuria),* if one man broke another's limb, *talio* was still permitted (that is, the person wronged could inflict the same injury as he had received); but in other cases, there were fixed money penalties. Theft involved a penalty of twice the value of the thing stolen, unless the thief was caught in the act, in which case he was flogged and "adjudged" to the person wronged.

By the early empire, reforms had substituted a fourfold penalty in the case of thieves caught in the act, and the court assessed all penalties for *injuria* (which now included defamation and insulting behaviour). The law of damage to property was regulated by statute, which in turn was much extended by interpretation. Additionally, there were situations in which a person might be held liable for damages even though he was not personally responsible.

Contract. In the early republic, a law of contract hardly existed. There was, however, an institution called *nexum,* of which little can be said with certainty except that it was a kind of loan so oppressive in character that it might result in the debtor's complete subjection to the creditor. It was obsolete long before imperial times. The contracts of classical law were divided into four classes: literal, verbal, real, and consensual. The literal contract was a type of fictitious loan formed by an entry in the creditor's account book; it was comparatively unimportant and was obsolete in Justinian's day. The verbal contract or *stipulatio* was of great importance, for it provided a form in which any agreement (provided it was lawful and possible) might be made binding by the simple method of reducing it to question and answer: "Do you promise to pay me 10,000 sesterces?" "I promise." Originally, it was absolutely necessary that the words be spoken, but it may be said that by Justinian's day a written memorandum of such a contract would be binding, even though in fact there had been no speaking at all.

If an agreement was not clothed in the form of a stipulation, it must, to be valid, fall under one of the types of real or consensual contracts. A real contract was one requiring that something should be transferred from one party to the other and that the obligation arising should be for the return of that thing. Real contracts included such items as loans of money or loans of a horse. Consensual contracts needed nothing except verbal or written agreement between the parties, and though there were only four such known to the law, these were the most important in ordinary life—those dealing with sale, hire, partnership, and mandate (acting upon instructions). In Justinian's day there was a further principle that in any case of reciprocal agreement—such as an agreement for exchange (but not sale)—if one party had performed, he could bring an action to enforce performance by the other. In addition to the foregoing contracts, a few other specific agreements were recognized as enforceable, but the general recognition of all serious agreements as binding was never achieved by the Romans.

Quasi-delict: This covered four types of harm, bracketed together on no clearly ascertainable principle. They included the action against an occupier for harm done by things thrown or poured from his house on to a public place; and the action against a shipowner, innkeeper, or stablekeeper for loss caused to customers on the premises through theft or damage by persons in his service.

Quasi-contract: This miscellaneous category embraced obligations that had no common feature save that they did not properly fall under contract, there being no agreement, nor under delict, there being no wrongful act. The most noticeable instances were, first, *negotiorum gestio,* which enabled one who intervened without authority in another's affairs for the latter's benefit to claim reimbursement and indemnity; second, the group of cases in which an action *(condictio)* is allowed for the recovery by A from B of what would otherwise be an unjustified enrichment of B at A's expense, as when A, mistaking the facts, paid B something that was not due *(condictio indebiti).* This notion of unjust enrichment as a source of legal obligation was one of the most pregnant contributions made by Roman law to legal thought.

The law of succession. The first requirement of any Roman will of historical times was the appointment of one or more heirs. An heir, in the Roman sense of the term, is a universal successor; that is, he takes over the rights and duties of the deceased (insofar as they are transmissible at all) as a whole. On acceptance, the heir becomes owner if the deceased was owner, creditor if he was creditor, and debtor if he was debtor, even though the assets be insufficient to pay the debts. It was thus possible for an inheritance to involve the heir in loss. Until Justinian's day this consequence could be avoided only by not accepting the inheritance, but Justinian made one of his most famous reforms by providing that an heir who made an inventory of the deceased's assets need not pay out more than he had received. Freedom of testation, furthermore, was not complete; a man was obliged to leave a certain proportion of his property to his children and in some cases to ascendants and brothers and sisters.

With regard to intestate succession, or succession without a will, those first entitled in early times were the deceased's own heirs—that is, those who were in his *potestas* or *manus* when he died and became free from that power at his death. Failing these, the nearest agnatic relations (relations in the male line of descent) succeeded, and, if there were no agnates, the members of the gens, or clan, of the deceased succeeded. Later reforms placed children emancipated from *potestas* on an equality with those under *potestas* and gradually gave the surviving spouse (in marriage without *manus*) greater and greater rights of succession. By Justinian's day the system had evolved as follows: descendants had the first claim, and failing these, a composite class consisting of ascendants, brothers and sisters of full blood, and children of deceased brothers and sisters. Next came brothers and sisters of the half blood and, finally, the nearest cognates (relations in the female line). Husband and wife were not mentioned, but their old rights were kept alive in the absence of any of the preceding categories. Justinian also gave to the poor widow a right to one-quarter of her husband's estate unless there were more than three children, in which case she shared equally with them. If, however, the heirs were her own children by the deceased, she only received the *ususfructus* (life interest) in what she took.

The law of procedure. In the earliest forms of procedure there were two stages: a preliminary one before the jurisdictional magistrate, in which the issue was developed; then the actual trial before the *judex,* or judge. The system required that set forms of words be spoken by the parties and, sometimes at least, by the magistrate. The parties making an assertion of ownership, for instance, would grasp the thing in dispute and lay a wand on it, after which the magistrate would intervene and say, "Let go, both of you." So formal was the procedure that a plaintiff who made the slightest mistake lost his case. Under the system it was also the responsibility of the plaintiff to produce the defendant physically in court and often to carry out the sentence of the court.

Instructions to the judge

Under new procedures developed in the 2nd and 1st centuries BC, the issue at the magisterial stage was formulated in written instructions to the *judex,* couched in the form of an alternative: "If it appears that the defendant owes the plaintiff 10,000 sesterces, the *judex* is to condemn the defendant to pay the plaintiff 10,000 sesterces; if it does not so appear, he is to absolve him." A draft of these written instructions was probably prepared for the plaintiff before he came into court, but there could be no trial until it was accepted by the defendant, for there was always a contractual element about a lawsuit under both the new and the older system. Pressure, however, could be exercised by the magistrate on a defendant who refused to accept instructions of which the magistrate approved, just as a plaintiff could be forced to alter instructions of which the magistrate disapproved, by the magistrate's refusal otherwise to give his order to the *judex* to decide the case.

In late republican times, still another system developed, first in the provinces, then in Rome. Under the new system, the magistrate used his administrative powers, always large, for the purpose of settling disputes. He could command, and thus if one man brought a complaint against another before him, he could investigate the matter and give the order he thought fit. As imperially appointed officers superseded republican magistrates, this administrative process became more common. The result was that the old contractual element in procedure disappeared as well as the old division into two stages. Justice was now imposed from above by the state—not, as originally, left to a kind of voluntary arbitration supervised by the state. (H.F.J./R.Po./M.A.M.)

BIBLIOGRAPHY

Sources: For modern collections of the Twelve Tables, see C.G. BRUNS, *Fontes Iuris Romani Antiqui,* 7th ed., 2 vol. (1919); P.F. GIRARD, *Textes du Droit Romain,* 6th ed. (1937); and S. RICCOBONO, *Fontes Iuris Romani Anteiustiniani*, 3 vol. (1940–43). The best edition of Justinian's Digest is that of THEODOR MOMMSEN (1868–70); of the Codex, that of P. KRUGER (1875–77). For an English translation of the first 15 books of the Digest, see C.H. MONRO, *The Digest of Justinian,* 2 vol. (1904–09). S.P. SCOTT, *The Civil Law* (1932), is a translation of the whole Corpus Juris, but very faulty. For the Institutes of Justinian, see J.B. MOYLE'S annotated text, 5th ed. (1913)—some of the notes are now out of date—with translation in a separate volume, 5th ed. (1913). For the Institutes of Gaius, see F. DE ZULUETA, vol. 1 (1946), text and translation; vol. 2 (1953), commentary. A masterly reconstruction of the praetorian edict was accomplished by O. LENEL, *Das Edictum perpetuum,* 3rd ed. (1956).

Textbooks: W.W. BUCKLAND, *A Textbook of Roman Law from Augustus to Justinian,* 3rd ed. (1964), is the standard reference in English; detailed but densely packed. More elementary are: W.W. BUCKLAND, *Manual of Roman Private Law,* 2nd ed. (1939); R.W. LEAGE, *Roman Private Law Founded on the Institutes of Gaius and Justinian,* 3rd ed. (1961); and M. KASER, *Römisches Privatrecht,* 4th ed. (1965; Eng. trans., *Roman Private Law,* 1965). Also P.F. GIRARD, *Manuel élémentaire de droit romain,* 8th ed. (1929). For a vigorous summary of Roman law in the classical era viewed in the light of the textual researches of German and Italian scholars, see F. SCHULZ, *Classical Roman Law* (1951). For civil procedure, see L. WENGER, *Institutionen des römischen Zivilprozessrechts* (1925; Eng. trans., *Institutes of the Roman Law of Civil Procedure,* rev. ed., 1940).

History: H.F. JOLOWICZ, *Historical lntrodrtction to the Study of Roman Law,* 2nd ed. (1952), is a most readable and scholarly account. For a briefer account, see H.J. WOLFF, *Roman Law: An Historical Introduction* (1951). L. WENGER, *Die Quellen des romischen Rechts* (1953), *is* a detailed historical study that relates Roman legal development to that of the surrounding legal systems. For the earlier development of the law, see the studies of the patriarchal joint family by C.W. WESTRUP, *Introduction to Early Roman Law,* 5 vol. (1934–54). See also WOLFGANG KUNKEL, *Römische Rechtsgeschichte,* 4th ed. (1964; Eng. trans., *An Introduction to Roman Legal and Constitutional History,* 1966), with its discriminating bibliographical appendix.

General: For the reception of Roman law in Europe, see P. KOSCHAKER, *Europa ztnd das römische Recht,* 4th ed. (1966); and briefly, P. VINOGRADOFF, *Roman Law in Mediaeval Europe,* 2nd ed. (1929, reprinted 1968). An interesting short comparative study may be found in W.W. BUCKLAND and ARNOLD D. MCNAIR, *Roman Law and Common Law,* 2nd ed. (1952). For classified references to modern literature, see A. BERGER, *Encyclopedic Dictionary of Roman Law* (1953). Periodical literature from 1800 is elaborately indexed by LUCIEN CAES and ROGER HENRION, *Collectio Bibliographica Operum ad Ius Romanum Pertinentium* (1949–). The volumes of *Iura* from 1950 (issued semiannually) contain ample bibliographies of current literature, including articles and reviews.

(M.A.M.)

Roman Religion

This article deals with the beliefs and practices of the inhabitants of the Italian peninsula from ancient times until the ascendancy of Christianity, 4th century AD.

NATURE AND SIGNIFICANCE

The Romans, according to the 1st-century-BC Roman orator and politician Cicero, excelled all other people in the unique wisdom that made them realize that everything is subordinate to the rule and direction of the gods. Yet Roman religion was not based on divine grace but instead on mutual trust (*fides*) between god and man. The object of Roman religion was to secure the cooperation, benevolence, and "peace" of the gods (*pax deorunz*). The Romans believed that this divine help would make it possible for them to master the unknown forces around them that inspired awe and anxiety *(religio),* and thus they would be able to live successfully. Consequently, there arose a body of rules, the *jus divinum* (divine law), ordaining what had to be done or avoided.

Object of Roman religion

These precepts for many centuries contained scarcely any moral element; they consisted of directions for the correct performance of ritual. Roman religion laid almost exclusive emphasis on cult acts, endowing them with all the sanctity of patriotic tradition. Roman ceremonial was so obsessively meticulous and conservative that, if the various partisan accretions that grew upon it throughout the years can be eliminated, remnants of very early thought can be detected near the surface.

This demonstrates one of the many differences between Roman religion and Greek religion, in which such remnants tend to be deeply concealed. The Greeks, when they first began to document themselves, had already gone quite a long way toward sophisticated, abstract, and sometimes daring conceptions of divinity and its relation to man. But the orderly, legalistic, and relatively inarticulate Romans never quite gave up their old practices. Moreover, until the vivid pictorial imagination of the Greeks began to influence them, they lacked the Greek taste for seeing their deities in personalized human form and endowing them with mythology. In a sense, there is no Roman mythology, or scarcely any. Although recent discoveries, notably at Caere (Cerveteri), confirm that Italians were not entirely unmythological, their mythology is sparse. What is found at Rome is chiefly only a pseudomythology (which, in due course, clothed their own nationalistic or family legends in mythical dress borrowed from the Greeks). Nor did Roman religion have a creed; provided that a Roman performed the right religious actions, he was free to think what he liked about the gods. And, having no creed, he usually deprecated emotion as out of place in acts of worship.

In spite, however, of the antique features not far from the surface, it is difficult to reconstruct the history and evolution of Roman religion. The principal literary sources, antiquarians such as the 1st-century-BC Roman scholars Varro and Verrius Flaccus, and the poets who were their contemporaries (under the late Republic and Augustus), wrote 700 and 800 years after the beginnings of Rome. They wrote at a time when the introduction of Greek methods and myths had made erroneous interpretations of the distant Roman past unavoidable. In order to supplement such conjectures or facts as they may provide, scholars rely on surviving copies of the religious calendar and on other inscriptions. There is also a rich, though frequently cryptic, treasure-house of material in coins and medallions and in works of art.

HISTORY

Early Roman religion. For the earliest times, there are the various finds and findings of archaeology. But they are not sufficient to enable scholars to reconstruct archaic Roman religion. They do, however, suggest that early in the 1st millennium BC, though not necessarily at the time of the traditional date for the founding of Rome (753 BC), Latin and Sabine shepherds and farmers with light plows came from the Alban Mount and the Sabine Hills, and that they proceeded to establish villages at Rome, the Laiins on the Palatine Hill and the Sabines (though this is uncertain) on the Quirinal and Esquiline hills. About 620 the communities merged, and *c.* 575 the Forum Romanum between them became the town's market.

Deification of functions. From such evidence it appears that the early Romans, like many other Italians, sometimes saw divine force, or divinity, operating in pure function and act, such as in human activities like opening doors or giving birth to children, and in nonhuman phenomena such as the movements of the sun and seasons of the soil. They directed this feeling of veneration both toward happenings that affected human beings regularly and, sometimes, toward single, unique manifestations, such as a mysterious voice that once spoke and saved them in a crisis (Aius Locutius). They multiplied functional deities of this kind to an extraordinary degree of "religious atomism," in which countless powers or forces were identified with one phase of life or another. Their functions were sharply defined; and in approaching them it was important to use their right names and titles. If one knew the name, one could secure a hearing. Failing that, it was often best to cover every contingency by admitting that the divinity was "unknown" or adding the precautionary phrase "or whatever name you want to be called" or "if it be a god or goddess."

Veneration of objects. The same sort of anxious awe was extended not only to functions and acts but also to certain objects that inspired a similar belief that they were in some way more than natural. This feeling was aroused, for example, by springs and woods, objects of gratitude in the torrid summer, or by stones that were often believed to be meteorites—*i.e.*, had apparently reached the earth in an uncanny fashion. To these were added products of human action, such as burial places and boundary stones, and inexplicable things, such as Neolithic implements (probably the mysterious meteorites were often these) or bronze shields (artifacts that had strayed in from more advanced cultures).

High gods

To describe the powers in these objects and functions that inspired the horror, or sacred thrill, the Romans eventually employed the word *numen*, suggestive of a god's nod, *nutus*; though so far there is no evidence that this usage was earlier than the 2nd century BC. The application of the word spirit to *numen* is anachronistic in regard to early epochs because it presupposes a society capable of greater abstraction. Nor must the term mana, used by Melanesians to describe their own concept of superhuman forces, be introduced too readily. The two societies are not necessarily analogous and, besides, the deduction from such comparisons that the Romans experienced an impersonal, pre-deistic, primordial stage of religion that neatly preceded the personal stage cannot be regarded as correct. On the contrary, from the very earliest times, the supernatural forces that they envisaged included a number of deities in analogous human forms; among them were certain "high gods." Foremost among these was a divinity of the sky, Jupiter, akin to the sky gods of other early Indo-European-speaking peoples, the Sanskrit Dyaus and Greek Zeus. Not yet, probably, a Supreme Being, though superior in some sense to other divine powers, this god of the heavens was easily linked with the forces of function and object, with lightning and weather, or with the uncanny stone that came from on high and was called Jupiter Lapis.

Purpose of sacrifice and magic. These gods and sacred functions and objects seemed charged with power because they were mysterious and alarming. In order to secure his food supply, physical protection, and increase, the early Roman believed that such forces had to be propitiated and made allies. Sacrifice was necessary. The product sacrificed would revitalize the divinity, which was seen as a power of action and therefore likely to run down unless so revitalized. By this nourishment he or it would become able and ready to fulfill requests. And so the sacrifice was accompanied by the phrase *macte* esto! ("be you increased!").

Prayer was a normal accompaniment of sacrifice, and as a conception of the divine powers gradually developed, it contained varying ingredients of flattery, cajolery, and attempted justification; but it also was compounded by magic—the attempt not to persuade nature, but to coerce it. Though the authorities (*e.g.*, c. 451–450 BC, Twelve Tables) sought to limit its noxious aspects, magic continued to abound throughout the ancient world. Even official rites remained full of its survivals, notably the annual festival of the Lupercalia and the ritual dances of the Salii in honour of Mars. Romans in historical times regarded magic as an oriental intrusion, but Italian tribes, such as the Marsi and Paeligni, were famous for such practices. Among them curses figured prominently, and curse inscriptions from c. 500 BC onward have been found in large numbers. There were also numerous survivals of taboo, a negative branch of magic: have no dealings with strangers, corpses, newborn children, spots struck by lightning, etc., lest harm befall you.

Religion in the Etruscan period. The amalgamation of the Latin and Sabine (?) villages of Rome coincided with, or more probably was soon followed by, a period in which Rome was under the control of at least one dynasty (the Tarquins) from Etruria, north of the Tiber (c. 575–510 BC, though some scholars would extend this domination to c. 450).

Importance of ritual. The Etruscans felt profound religious anxieties and were more devoted to ritual than any other people of the ancient Western world. Though sources are, again, late and unsatisfactory, it appears that they possessed a comprehensive collection of rules regulating these rites. Etruscan culture was heavily based on influences from Greece in its orientalizing period, supplemented by direct contacts with Phoenicia and other eastern lands. But the religion of Etruria proclaims a very un-Greek view of the abasement and nonentity of man before the gods and their will.

Divination and views of the afterlife

To the Etruscans the whole fanatical effort of life was directed toward forcing their deities, led by Tinia or Tin (Jupiter), to yield up their secrets by divination. They saw an intimate link existing between heaven and earth, which seemed to echo one another within a unitary system, and they were more ambitious than either Greeks or Romans in their claims to foretell the future. They also formed an exceptionally complex, rich, and imaginative picture of the afterlife. The living were perpetually obsessed by their care for the dead, expressed in elaborate, magnificently equipped and decorated tombs and lavish sacrifices. For, in spite of beliefs in an underworld, or Hades, there was also a conviction that the individuality of the dead somehow continued in their mortal remains; and it was therefore imperative that they should take pleasure in their graves or tombs and not return to haunt the living. From the 4th century BC onward, after the Etruscans had lost their political power to Rome, their art depicts horrors indicating an increasing fear of what death might bring.

Influence *on Roman* religion. The Roman religion continued to display certain obvious debts to the period when the city had been under Etruscan control: It is true that the Roman shades (Di Manes) were much less substantial than the fantastic Etruscan conceptions and. although Etruscan divination by the liver and entrails survived and later became increasingly fashionable in Rome, Roman diviners in general, products of a more realistic and prosaic society, never aspired to such precise information about the future as the Etruscans had hoped to gain. Yet, it was the Etruscans who first gave a vigorous definition to Italian religious forms. Indeed, many of the religious features that patriotic historians preferred to ascribe to the mythical King Numa Pompilius (who was supposed to have been Romulus' Sabine successor in the 8th century BC—the man of peace following the man of war) date, in fact, from the period of Etruscan domination two centuries later. Nevertheless, Romans acknowledged a debt to Etruria that included much ceremony and ritual, and the plan, appearance, and decoration of a number of temples, notably the great shrine of the Capitoline Triad, Jupiter, Juno, and Minerva. The Romans also were indebted to the Etruscans for their first statues of gods, including the cult image of Jupiter commissioned from an Etruscan for the Capitoline temple. Such statuary, showing the gods in human shape, encouraged the Romans to think of their gods in this way, with the consequent possibility of investing them with myths, which thereafter gradually accumulated around them in the form of Hellenic stories often infused with a native patriotic eiement.

The calendar

Above all, Rome owed to its Etruscan kings its religious calendar. In addition to poetical works discussing the calendar in antiquarian fashion, such as the Fasti of Ovid, there are extant fragments of about 40 copies of the calendar itself, in a revised shape established by Julius Caesar. Besides the Julian revision, there is an incomplete pre-Caesarian, Republican calendar, the Fasti Antiatos, discovered at Antium (Anzio); it dates from after 100 BC. It is possible to detect in these calendars much that is very ancient, including a pre-Etruscan ten-month solar year. However, the basis of the calendars, in their surviving form, is later, since it consists of an attempt to reconcile the solar and lunar year, in accordance with Babylonian calculations. This endeavour belongs to the period of Etruscan domination of Rome—for example, the names of the months April and June (in their Roman form) come from Etruria. Moreover, the presence or absence of certain festivals permits a dating approximating to the time of Etruscan domination in the later 6th century BC. Additional modifications were introduced in the following century and again when the calendar was subsequently published (30 BC).

The festivals it records, of which the earliest are indicated in large letters, reflect a period of transition between country and town life. Though local cult continued to remain active, many forms of worship hitherto maintained by families and farms had now been taken over by the comparatively mature Roman state. The state management blocked any tendency toward spiritualization and removed the need for any vigorous individual participation; however, by ensuring that the gods were conciliated by a schedule corresponding to the regular process of nature, it made the individual citizens feel for centuries that relations with the supernatural were being maintained safely.

Influence of Greek religion

Religion in the early Republic. Even if, as tradition records, a coup d'etat dislodged the Etruscan kings before 500 BC, the first half of the 5th century saw no weakening of trade relations with Etruria. Its southern cities, such as Caere (Cerveteri) and Veii close to Rome, had long used the Greek city of Cumae in Campania as a commercial outlet, converting it into an important grain supplier. And now Rome, faced with a shortage of grain, arranged for it to be imported from Cumae. The same city also influenced the foundation of Roman temples in the Greek style. Rome, which had already become accustomed to Greek religious customs in the Etruscan epoch, now showed a willingness to absorb them. This forms a strange contrast to its deeply ingrained religious conservatism. Moreover, at some quite early stage (though there is no positive evidence of the practice until the 3rd century), Romans borrowed from elsewhere in Italy a special ritual (evocatio) for inviting the patron deities of captured towns to abandon their homes and migrate to Rome.

In an emergency in 399 BC, during a difficult siege of Veii, Rome carried Hellenization further by importing a Greek rite in which, as an appeal to emotional feeling, images of pairs of gods were exhibited on couches before tables spread with food and drink; this rite (*lectisterni*urn) was designed to make them Rome's welcome guests. From the same century onward, if not earlier, pestilences were averted by another ritual (supplicatio), in which the whole people went around the temples and prostrated themselves in Greek fashion. Later the custom was extended to the celebration of victories.

Importation of mystery religions and astrology

Religion in the later Republic: crises and new trends. The *lectisternium* was repeated, with increased elaboration and pomp, in 217 BC during a period in which emotional religion was running rampant because of Hannibal's invasion of Italy in the Second Punic War. Faced with a flood of fears and anxieties and reports of many alarming and extraordinary events, Rome took precautions to secure the favour of all manner of gods. Among them, as a desperate attempt at novelty when appeals to the usual deities seemed stale, was the introduction of the Great Mother of Asia Minor, Cybele (204). Eighteen years later, the equally orgiastic worship of Dionysus (Bacchus) was coming in so rapidly and violently, by way of southern Italy, that the Senate, scenting subversion, repressed its practitioners. But these and other mystery religions, promising initiation, afterlife, and an excitement that Roman national cults could not provide, had come to stay and, although there were long periods of official disapproval before acclimatization was completed, they gradually played an immense part upon the religious scene. Eastern astrology, too, became extremely popular. It was based on the conviction that, since there is cosmic sympathy between the earth and other heavenly bodies, and since, therefore, the emanations of these bodies influence the earth, men must learn how to foresee their dictates—and outwit them.

Influence of Stoicism

Astrological practices received encouragement from Stoic philosophy, which was introduced to Rome in the 2nd and early 1st centuries BC, notably by Panaetius and Poseidonius. The Stoics saw this pseudoscience as proof of the Platonic unity of the universe. Stoicism affected Roman religious thinking in at least three other ways. First, it had a deterministic effect, encouraging a widespread belief in Fate and also, somewhat illogically, in Fortune, both of which were revered in other parts of the Mediterranean and Near Eastern world. Second, Stoicism infused a new spirituality into religious thinking by its insistence that the human soul is part of the universal spirit and shares its divinity. Third, the moral implication of this, as the Stoics pointed out, was that all men are brothers and must treat each other accordingly. This demonstration struck a chord in the psychology of the Romans, who possessed strongly ethical inclinations and now, at last, saw this trend supported and justified by a philosophical sanction that their formalistic religion had not provided. In changing times of imperialism, materialism, and widespread heart-searching, the state religion had failed to fill the vacuum, and philosophy stepped in instead. At the same time the negative approach of Roman religion to the afterlife was counteracted by an influx of speculations that blended theology, mysticism, and magic and claimed the legendary Orpheus and Pythagoras as prophets.

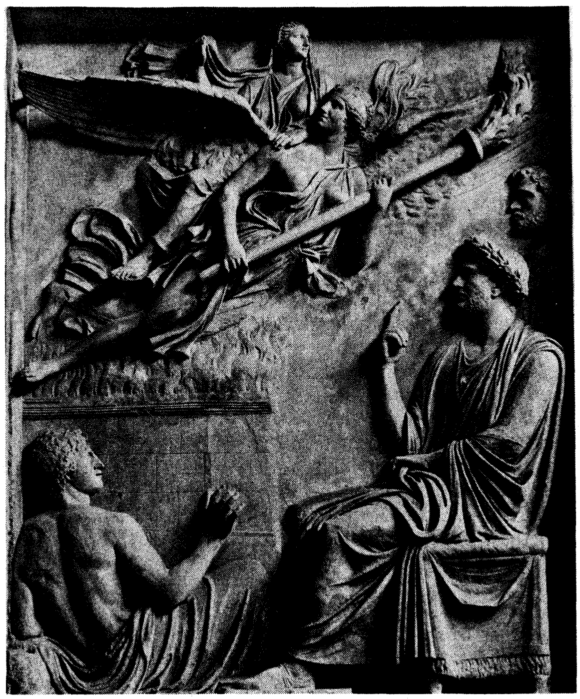
While their national poet Ennius *(239–169* BC) helped to diffuse such beliefs, he and the comic dramatist Plautus ridiculed the traditional Roman gods on the stage. The upper class attitude of the times was expressed by the historian Polybius, the priestly lawyer Scaevola, the scholarly Varro, and the orator and philosopher Cicero, who maintained that the importance of religion was political, residing in its power to keep the multitude under control, to prevent social chaos, and to promote patriotic feeling.

The imperial epoch: the final forms of Roman paganism. After the prolonged horrors of civil war had ended *(30* BC), the victorious Octavian, the adoptive son of the dictator Caesar and founder of the imperial regime or principate, decided, correctly, that the ancient religion was far from dead and that the restoration of all its forms would respond to a strong popular, instinctive belief that the disasters of the past generations had been due to the neglect of religious duties.

Deification of Caesar and Augustus

The imperial cult. Octavian himself took the name Augustus, a term indicating a contrast to *humanus.* This did not make him a god in his lifetime, but, combined with the insertion of his *numen* and his *genius* (originally the procreative power that enables a family to be carried on) into certain cults, it prepared the way for his posthumous deification, just as Caesar had been deified before him. Both were deified by the state because they seemed to have given Rome gifts worthy of a god. From earliest times in Greece there had been an idea that if someone saved you, you should pay him the honours you would offer to a god. Alexander the Great and his successors had demanded reverence as divine saviours, and Ptolemy II Philadelphus of Egypt *(285–246* BC) introduced a cult of his own living person. The Stoic belief that the human soul was part of the world soul was a corollary of the view that great men possessed a larger share of this divine element. Moreover, the philosopher Euhemerus *(c.300* BC) had elaborated a theory that the gods themselves had once been human; this idea was readily adapted to the supposed careers of Hercules and of Castor and Pollux; and the Romans applied it to their own gods Saturn and Quirinus, the latter identified with the national founder, Romulus, risen to heaven. And so it became customary, if emperors (and empresses) were approved of in their lives, to raise them to divinity after their deaths. They were called *divi,* not *dei* like the Olympian gods; the latter were prayed to, but the former were regarded with veneration and gratitude.

As the empire proceeded and the old religion seemed more and more irrelevant to people's personal preoccupations and successive national emergencies, the cult of the *divi,* subsequently grouped together in a single Hall of Fame, remained foremost among the patriotic cults that were increasingly encouraged as unifying forces. Concentrating on the protectors of the emperor and the nation, they included the worship of Rome herself, and of the *genius* of the Roman people; for the army a number of special military celebrations are recorded on the Calendar of Doura-Europus in Mesopotamia (Feriale Duranum, *c.* AD *225–27).* As for the ruling emperors, they were more and more frequently treated as divine, with varying degrees of formality, and officially they often were compared with gods. As monotheistic tendencies grew, however, this custom led not so much to their identification with the gods as to the doctrine that they were the elect of the divine powers, who were defined as their companions *(comites).* In pursuance of this way of thinking, as official paganism approached its last days, the emperors Diocletian (AD *284–305)* and Maximian took the names Jovius and Herculius respectively, after their Companions and Patrons Jupiter and Hercules.


Apotheosis of Faustina, wife of Marcus Aurelius, ancient bas-relief. In the Capitoline Museum, Rome.
Anderson— Alinari

Introduction of *Christianity and Mithraism.* By now, however, the humanistic idea that men could *become* gods had ceased to have any plausibility. Plotinus and his Neoplatonism, the dominant philosophy of the pagan world from the mid-3rd century AD, had given powerful, mystical shape to the Platonic and Stoic conception that the universe is governed by a single force. On the other hand, the greatest religious figure of the century, the Iranian Mani, who had started to preach in Mesopotamia *c. 240,* dramatically preached the opposing dualistic idea that the world is the creation not only of a good power but of an evil one as well. Mani's church, which alarmed Diocletian and for a time attracted the great Christian theologian St. Augustine, absorbed many of the innumerable cults of Gnostics who claimed special knowledge (*gnōsis*) by illumination and revelation and taught how man can purge the nonspiritual from within him and escape his earthly prison. More impressively, the cult of the Persian Mithra blended the dualism of Mani with the emotional initiations of the mystery religions (corrected by a much sterner tone of moral endeavour) and became a strong link between the cult of the sun (which appealed to contemporary monotheists) and the fashionable revulsion from the senses that was shortly to lead to Christian monasticism. Like Christianity, Mithraism had its sacraments; but the life of Mithra exercised a less far-reaching appeal than the life of Christ, and Mithra's cult excluded women.

Speculative religious thought

Christianity, unique in its universal charity and unique also in its demand for a noble effort of faith in Jesus' blend of divinity and humanity, was the religion that prevailed in the Roman world. It satisfied the emperor Constantine's impulsive need for divine support, and from AD 312 onward, by a complex and gradual process, it became the official religion of the empire.

The survival of Roman religion. For a time, coins and other monuments continued to link Christian doctrines with the worship of the sun, to which Constantine had been addicted previously. But even when this phase came to an end, Roman paganism continued to exert other, permanent influences, great and small. The emperors passed on to the popes the title of chief priest, *pontifex maximus.* The saints, with their distribution of functions, often seemed to perpetuate the many *numina* of ancient tradition. The ecclesiastical calendar retains numerous remnants of pre-Christian festivals—notably Christmas, which blends elements including both the feast of the Saturnalia and the birthday of Mithra. But, most of all, the mainstream of Western Christianity owed ancient Rome the firm discipline that gave it stability and shape, combining insistence on established forms with the possibility of recognizing that novelties need not be excluded, since thev were implicit from the start;

ROMAN GODS AND GODDESSES

Jupiter, Mars, and Quirinus

The earliest divinities. The early Romans, like other Italians, worshipped not only purely functional and local forces but also certain high gods. Chief among them was the sky god Jupiter, whose cult, at first limited to the communities around the Alban Mount, later gained Rome as an adherent. The Romans gave Jupiter his own priest (*flamen*), and the fact that there were two other senior *flamines,* devoted to Mars and Quirinus, confirms other indications that the cults of these three deities, envisaged perhaps in some sort of association, belonged to a very early stratum (though the theory of their correspondence to the three-class social division of the early Indo-European-speaking peoples is generally unacceptable). Mars, whose name may or may not be Indo-European, was a high god of many Italian peoples, as liturgical bronze tablets found at Gubbio, the Tabulae Iguvinae (c. 200–c. 80 BC), confirm, protecting them in war and defending their agriculture and animals against disease. Later, he was identified with the Greek god of war, Ares, and also was regarded as the father of Romulus. Mars Gradivus presided over the beginning of a war and Mars Quirinus over its end, but earlier Quirinus had apparently, as a separate deity, been the patron of the Quirinal village before its amalgamation with the Palatine; subsequently he was believed to have been the god that Romulus became when he ascended into heaven.

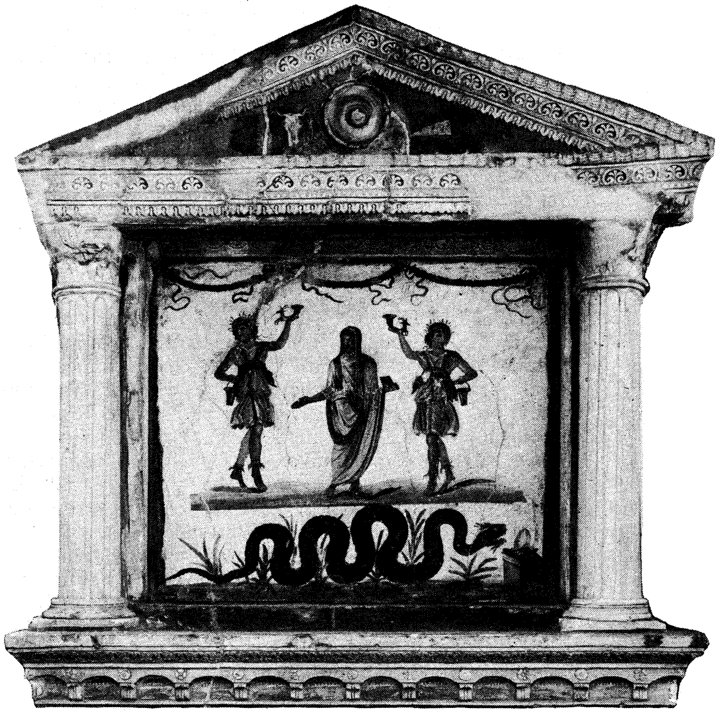
Janus and Vesta

Two other forces that belong to an early phase were Janus and Vesta, the powers of the door and hearth, respectively. Janus, who had no Greek equivalent, was worshipped beside the Forum in a small shrine with double doors at either end, and originated either from a divine power that regulated the passage over running water or rather, perhaps, from sacred doorways like those found on the art of Bronze Age Mycenae. Janus originally stood for the magic of the door of a private house or hut and later became a part of the state religion. The gates of his temple were formally closed when the state was at peace, a custom going back to the primitive war magic that required armies to march out to battle by this properly sanctified route, Vesta, too, passed from the home to the state, always retaining a circular temple reminiscent of the primitive huts whose form can be reconstructed from traces left in the earth and from surviving funerary urns. Vesta's shrine contained the eternal fire, but the absence of a statue indicates that it preceded the anthropomorphic period; its correspondence with the Indian *garhapatya,* "house-father's fire," suggest an origin prior to the time of the differentiation of the Indo-European-speaking peoples. The cultic site just outside the area of the primitive Palatine settlement indicates that there had been a form of fire worship even earlier than Vesta's (dedicated to the deity Caca) on the Palatine itself. The cult of Vesta, tended by her Virgins, continued to flourish until the end of antiquity, endowed with an important role in the sacred protectorship of Rome.

Manes, Indigetes, Lares, and Penates

The Di Manes, collective powers (later "spirits") of the dead, may mean "the good people," an anxious euphemism like the Greek name of "the kindly ones" for the Furies; an alternative connection with the Phrygian god Men seems less probable. As a member of the family or clan, however, the dead man or woman would, more specifically, be one of the Di Parentes; reverence for ancestors was the core of Roman religious and social life. Di Indigetes was a name given collectively to these forebears, as well as to other deified powers or spirits who likewise controlled the destiny of Rome. For example, the name Indiges is applied to Aeneas, whose mythical immigration from Troy led to the eventual foundation of the city. According to an inscription of the 4th century BC (found at Tor Tignosa, 15 miles south of Rome), Aeneas is also called Lar, which indicates that the Lares, too, were originally regarded as divine ancestors and not as deities who presided over the farmland. The Lares were worshipped wherever properties adjoined, and inside every home their statuettes were placed in the domestic shrine (*lararium*). Under state control they moved from boundaries of properties to crossroads (where Augustus eventually associated his own *genius* with the cult) and were worshipped as the guardian spirits of the whole community (Lares Praestites). The cult of the Di Penates likewise moved from house to state. From very early times the Penates, the powers that ensured that there was enough to eat, were worshipped in every home. They also came to be regarded as national protectors, the Penates Publici. Originally they were synonymous with the Dioscuri (Castor and Pollux). The legend that they had been brought to Italy by Aeneas was imported from Lavinium (Pratica di Mare) when the early Romans incorporated that town into their own state.

Brogi—Alinari



Altar of the Lares, depicting two Lares on either side of the Genius, AD 69–72. In the House of the Vettii, Pompei.

Diana and Fors Fortuna

The divinities of the later Regal period. Two other deities whose Roman cults tradition attributed to the period of the kings were Diana and Fors Fortuna. Diana, an Italian wood goddess worshipped at Aricia in Latium and prayed to by women who wanted children, was in due course identified with the Greek Artemis. Her temple on the Aventine Hill (*c.* 540 BC) with its statue, an imitation of a Greek model from Massilia (Marseille), was based on the Temple of Artemis of Ephesus. By establishing such a sanctuary, the Roman monarch Servius Tullius hoped to emulate the Pan-Ionian League among the Latin peoples. Fors Fortuna, whose temple across the Tiber was one of the few that slaves could attend, was similar to the oracular shrines of Fortuna at Antium (Anzio) and Praeneste (Palestrina). Originally a farming deity, 'she eventually represented luck. She came to be identified with Tyche, the patroness of cities and goddess of Fortune among the Hellenistic Greeks.

In Roman tradition, Servius Tullius reigned between two

Etruscan kings, Tarquinius Priscus and Tarquinius Superbus. The Etruscan kings began and perhaps finished the most important Roman temple, devoted to the cult of the Capitoline Triad, Jupiter, Juno, and Minerva (the dedication was believed to have taken place in 509 or 507 BC after the expulsion of the Etruscans). Such triads, housed in temples with three chambers (cellae), were an Etruscan institution. But the grouping of these three Roman deities seems to be owed to Greek anthropomorphic ideas, since Hera and Athena, with whom Juno and Minerva were identified, were respectively the wife and daughter of Zeus (Jupiter). In Italy, Juno (Uni in Etruscan) was sometimes the warlike high goddess of a town (*e.g.*, Lanuvium in Latium), but her chief function was to supervise the life of women, and particularly their sexual life. The functions of Minerva concerned craftsmen and reflected the growing industrial life of Rome. Two gods with Etruscan names, both worshipped at open altars before they had temples in Rome, were Vulcan and Saturn, the former a fire god identified with the Greek blacksmiths' deity Hephaestus, and the latter an agricultural god identified with Cronus, the father of Zeus. Saturn was worshipped in Greek fashion, with head uncovered.

Juno, Minerva, Vulcan, and Saturn

The focal point of the cult of Hercules was the Great Altar (Ara Maxima) in the cattle market, just inside the boundaries of the primitive Palatine settlement. The altar may be traced to a shrine of Melkart established by traders from Phoenicia in the 7th ccntury BC. The name of the god, however, was derived from the Greek Heracles, whose worship spread upward from southern Italy, brought by traders who venerated his journeys, his labours, and his power to avert evil. In a market frequented by strangers, a widely recognized divinity of this type was needed to keep the peace. The Greek cult, at first private, perhaps dates from the 5th century BC.

The divinities of the Republic. An important series of temples was founded early in the 5th century BC. The completion of the temple of the Etruscan Saturn was attributed to this time (497). A shrine honouring the twin horsemen, the Dioscuri (Castor and Pollux), was also built in this period. An inscription from Lavinium describing them by the Greek term ***kouroi*** indicates a Greek origin (from southern Italy) without Etruscan mediation. In legend, the Dioscuri had helped Rome to victory in a battle against the Latins at Lake Regillus, and in historic times, on anniversaries of that engagement, they continued to preside over the annual parade of knights (equites). From southern Italy, too, came the cult of Ceres, whose temple traditionally was vowed in 496 and dedicated in 493. Ceres was an old Italian deity who presided over the generative powers of nature and came to be identified with Demeter, the Greek goddess of grain. She owed her installation in Rome to the influence of the Greek colony of Cyme (Cumae), from which the Romans imported grain during a threatened famine. The association of Ceres at this temple with two other deities, Liber (a fertility god identified with Dionysus) and Libera (his female counterpart), was based on the triad at Eleusis in Greece. The Roman temple, built in the Etruscan style but with Greek ornamentation, stood beside a Greek trading centre on the Aventine Hill and became a rallying ground for the plebeians, the humbler section of the community who were hard hit by the grain shortage and who, at this time, were pressing for their rights against the patricians.

Ceres, Apollo, and Venus

Cumae also played a part in the introduction of Apollo. The Sibylline oracles housed in Apollo's shrine at Cumae allegedly were brought to Rome by the last Etruscan kings. The importation of the cult (431 BC) was prescribed by the Sibylline Books at a time when Rome, as on earlier occasions, had requested Cumae for help with grain. The Cumaean Apollo, however, was primarily prophetic, whereas the Roman cult, introduced at a time of epidemic, was concerned principally with his gifts as a healer. This role may possibly have been derived from the Etruscans, whose Apollo is known from a superb statue of *c.* 500 BC from Veii, Etruria's nearest city to Rome. In 82 BC the Sibylline Books were destroyed and replaced by a collection assembled from various sources. Later, Augustus elevated Apollo as the patron of himself and his regime, intending thereby to convert the brilliant Hellenic god of peace and civilization to the glory of Rome.

Unlike Apollo, Aphrodite did not keep her name when she became identified with an Italian deity. Instead, she took on the name Venus, derived, without complete certainty, from the idea of *venus*, "blooming nature" (the derivation from ***venia***, "grace," seems less likely). She gained greatly in significance because of the legend that she was the mother of Aeneas, the ancestor of Rome, whom statuettes of the 5th century BC from Veii show escaping from Troy with his father and son. From the time of the Punic Wars 200 years later the Trojan legend grew, for long before the 1st-century-BC dictators Sulla and Caesar claimed Venus as their ancestors, the story was interpreted as the preface to the Carthaginian struggle.

A number of gods were spoken of as possessing accompaniments, often in the feminine gender; *e.g.*, Lua Saturni, Moles Martis. These attachments, sometimes spoken of as cult partners, were not the wives of the male divinities but rather expressed a special aspect of their power or will. A similar origin could be ascribed to the worship of divine powers representing "qualities." Fides (Faith, Loyalty), for example, may at first have been an attribute or aspect of a Latin-Sabine god of oaths, Semo Sanctus Dius Fidius; and in the same way Victoria may come from Jupiter Victor. Some of these concepts were worshipped very early, such as Ops (Plenty, later associated with Saturn, and equated with Hebe), and Juventas (who watched over the men of military age). The first of these qualities to receive a temple, as far as is known, is Concordia (367), in celebration of the end of civil strife. Salus (health or well-being) followed in c. 302, Victoria in c. 300, Pietas (dutifulness to family and gods, later exalted by Virgil as the whole basis of Roman religion) in 191. The Greeks, too, from the earliest days, had clothed such qualities in words; *e.g.*, Shame, Peace, Justice, Fortune. In the Hellenic world they had a wide variety of signification, ranging from full-fledged divinity to nothing more than abstractions. But in early Rome and Italy they were in no sense abstractions or allegories and were likewise not thought of as possessing-the anthropomorphic shape that the term personification might imply. They were things, objects of worship, like many other functions that were venerated. They were external divine forces working upon men and affecting them with the qualities that their names described. Later on, under philosophical (particularly Stoic) influences that flooded into ethically minded Rome, they duly took their place as moral concepts, the Virtues and Blessings which abounded for centuries and were depicted in human form on Roman coinage as part of the imperial propaganda.

Divine qualities

The sun and stars. Little or no contribution to cosmology was made in the Roman world, and the demonstration of Aristarchus of Samos (*c.* 270 BC) that the earth revolves around the sun received virtually no support. The complicated geocentric interpretation that held sway in Rome was summed up in Cicero's ***Dream*** of *Scipio*. It formed the basis for the concept of the solar system on which the popular pseudoscience of astrology was founded, the sun being regarded as the centre of the concentric planetary spheres encircling the earth—not the centre of the cosmos in the sense of Aristarchus but its heart. From the 5th century BC onward this solar god was identified with Apollo in his role as the supreme dispenser of agricultural wealth. Possessor of a sacred grove at Lavinium, Sol Indiges was regarded as one of the divine ancestors of Rome. During the last centuries before the Christian era, worship of the sun spread throughout the Mediterranean world and formed the principal rallying point of paganism's last years. Closely associated with the sun cult was that of Mithra, the sun's ally and agent who was elevated to partake of communion and the love feast as the god's companion. Sun worship was popular in the army, and particularly on the Danube. Aurelian, one of the great military emperors produced by that area in the 3rd century, built a magnificent temple of

Sol

Sol Invictus (the Unconquered Sun) at Rome (274). Constantine the Great declared the sun his Comrade on empire-wide coinages and devoted himself to the cult until he adopted Christianity in its stead.

WORSHIP, PRACTICES, AND INSTITUTIONS

The ***rex sacrorum, flamines,*** and colleges of priests

Priests. Precedence among Roman priests belonged to the ***rex sacrorum*** ("king of the sacred rites"), who, after the expulsion of the kings, took over the residue of their religious powers and duties that had not been assumed by the Republican officers of state. Nevertheless, the hold exercised by the *rex sacrorum* and his colleagues was weakened by the Code of the Twelve Tables (*c.* 451–450 BC), which displayed the secular arm exercising some control over sacral law. As late as *c.* 275 BC the religious calendar was still dated by the ***rex sacrorum,*** but by this time he was already fading into the background.

Very early origins can also be attributed to some of the ***flamines,*** the priests of certain specific cults, and particularly to the three major ***flamines*** of Jupiter, Mars, and Quirinus. Jupiter's priest, the *flamen dialis*, was encompassed by an extraordinary series of taboos, some dating back to the Bronze Age, which made it difficult to fill the office in historic times.

Except for the ***rex sacrorum*** and *flamen dialis*, whose duties were unusually professional and technical, almost all Roman priesthoods were held by men prominent in public life. The social distinction and political prestige carried by these part-time posts caused them to be keenly fought for.

There were four chief colleges, or boards, of priests: the ***pontifices, augures, quindecimviri sacris faciundis,*** and ***epulones.*** Originally three, and finally 16 in number, the ***pontifices*** (whose name may recall antique tasks and magic rites in connection with bridges) had assumed control of the religious system by the 3rd century BC. The chief priest, the *pontifex maximus* (the head of the state clergy), was an elected official and not chosen from the existing *pontifices.* The ***augures,*** whose name may have been derived from the practice of magic in fertility rites and perhaps meant "increasers," had the task of discovering if the gods did or did not disapprove of an action. This they performed mainly by interpreting divine signs in the movements of birds ***(auspicia).*** Such divination was elevated, perhaps under Etruscan influence, into an indispensable preliminary to state acts, though the responsibility for the decision rested not with the priests but with the presiding state officials, who were said to "possess the auspices." In private life too, even as late as Cicero and Horace in the 1st century BC, important courses of action were often preceded by consultation of the heavens. The Etruscan method of divining from the liver and entrails of animals ***(haruspicina)*** became popular in the Second Punic War, though its practitioners (who numbered 60 under the empire) never attained an official priesthood.

Of the other two major colleges, the ***quindecimviri*** (Board of Fifteen, earlier Ten) ***sacris faciundis*** looked after foreign rites, and the members of the other body, the *epulones*, supervised religious feasts. There were also ***fetiales,*** priestly officials who were concerned with various aspects of international relationships, such as treaties and declarations of war. Also six Vestal Virgins, chosen as young girls from the old patrician families, tended the shrine and fire of Vesta and lived in the House of Vestals nearby, amid a formidable array of prehistoric taboos.

Annual festivals

Shrines and temples. The Roman calendar, as introduced or modified in the period of the Etruscan kings, contained 58 regular festivals. These included 45 Feriae Publicae, celebrated on the same fixed day every year, as well as the Ides of each month, which were sacred to Jupiter, and the Kalends of March, which belonged to Mars. Famous examples of Feriae Publicae were the Lupercalia (February 15) and Saturnalia (December 17, later extended). There were also the Feriae Conceptivae, the dates of which were fixed each year by the proper authority, and which included the Feriae Latinae (Latin Festival) celebrated on the Alban Mount, usually at the end of April.

Templum is a term derived from Etruscan divination. First of all, it meant an area of the sky defined by the priest for his collection and interpretation of the omens. Later, by a projection of this area onto the earth, it came to signify a piece of ground set aside and consecrated to the gods. At first such areas did not contain sacred buildings, but there often were altars on such sites, and later shrines. In Rome, temples have been identified from *c.* 575 BC onward, including not only the round shrine of Vesta but also a group in a sacred area (S. Omobono), close to the river Tiber beside the cattle market (Forum Boarium). The great Etruscan temples, made of wood with terra-cotta ornaments, were constructed later and culminated in the temple of the Capitoline Triad. Subsequently, more solid materials, such as tufa, travertine, marble, cement, and brick, gradually came into use. Temple archives, now vanished, play a large part in the historical tradition, and the anniversaries of the vows to build the temples and their dedication were scrupulously remembered and celebrated on numerous coins.

Sacrifice and burial rites. The characteristic offering of the Romans was a sacrifice accompanied by a prayer or vow. (The Triumph, associated with Jupiter, was regarded as a thanksgiving in discharge of a vow.) Sacrifices did not have to be animals, but they were regarded as more effective than anything else, the pig being the commonest victim, with sheep and ox added on important occasions. Considered best of all were the basic elements of life: heart, liver, and kidneys. Human sacrifice, on the whole, was extraneous to Roman custom, though its practice among the Etruscans may have contributed to the institution of gladiatorial funeral games in both Etruria and Rome, and it was resorted to in major crises, notably during the Second Punic War (216). Earlier in the century, and perhaps once before, a member of the family of the Decii had given up his life by human self-sacrifice ***(devotio)*** in a critical battle.

Cremation and inhumation

Although ancestors were meticulously revered, there was nothing resembling the comprehensive Etruscan attention to the dead. In spite of elaborate philosophizing by Cicero and Virgil about the possibility of some sort of survival of the soul (especially for the deserving), most Romans' ideas of the afterlife, unless they believed in the promises of the mystery religions, were vague. Such ideas often amounted to a cautious hope or fear that the spirit in some sense lived on, and this was sometimes combined with an anxiety that the ghosts of the dead, especially the young dead who bore the living a grudge, might return and cause harm. Graves and tombs were inviolable, protected by supernatural powers and by taboos. In the earliest days of Rome both cremation and inhumation were practiced simultaneously, but by the 2nd century BC the former had prevailed. Some 300 years later, however, there was a massive reversion to inhumation, probably due to an inarticulate revival of the feeling that the future welfare of the soul depended on comfortable repose of the body—a feeling that, as sarcophagi show, was fully shared by the adherents of the mystery cults, though, on the rational level, it contradicted their assurance of an afterlife in some spiritual sphere. The designs on these tombs reflect the soul's survival as a personal entity that has won its right to paradise.

Religious art. A vast gallery of architecture, sculpture, numismatics, painting, and mosaics illustrates Roman religion and helps to fill the gaps left by the fragmentary, though extensive, literary and epigraphic record. Starting with primitive statuettes and terra-cotta temple decorations, this array eventually included masterpieces such as the Apollo of Veii. Other works of art, over 400 years later, include paintings illustrating Dionysiac mysteries at Boscoreale near Pompeii, and the reliefs of Augustus' Ara Pacis at Rome; and with the Christian emblems of Constantinian sarcophagi and coinage a thousand years of ancient Roman religious art comes to an end.

CONCLUSION

Though Roman religion never produced a comprehensive code of conduct, its early rituals of house and farm engendered a feeling of duty and unity. Its idea of reciprocal understanding between man and god not only imparted the sense of security that Romans needed in order to achieve their successes but stimulated, by analogy, the concept of mutual obligations and binding agreements between one man and another. Except for rare aberrations, such as human sacrifice, Roman religion was unspoiled by orgiastic rites and savage practices. Moreover—unlike ancient philosophy—it was neither sectarian nor exclusive. It was a tolerant religion, and it would be difficult to think of any other whose adherents committed fewer crimes and atrocities in its name.

BIBLIOGRAPHY.

General works: R.M. OGILVIE, *The Romans and Their Gods in the Age of Augustus* (1970), a short up-to-date account; H.J. ROSE, *Ancient Roman Religion* (1948), a standard work; "Roman Religion, 1910–1960," in *Journal of Roman Studies*, 50:161–172 (1960); C. BAILEY, *Phases in the Religion of Ancient Rome* (1932), partly outdated; W. WARDE FOWLER, *The Religious Experience of the Roman People from the Earliest Times to the Age of Augustus*, 2nd ed. (1922); K. LATTE, *Römische Religionsgeschichte* (1960); M.P. NILSSON, *Geschichte der griechischen Religion,* vol. 2, *Die Hellenistische und Romische Zeit* (1950); G. WISSOWA, *Religion und Kultus der Romer*, 2nd ed. (1912), a basic collection of material.

Special periods and subjects: R. BLOCH, *The Origins of Rome* (1960); H. WAGENVOORT, *Roman Dynamism: Studies in Ancient Roman Thought, Language and Customs* (1947), a good but somewhat one-sided account; M. PALLOTTINO, *Etruscologia*, 5th ed. (1963); A. GRENIER, *Les Re'ligions étrusque et romaine* (1948); A.K. MICHELS, *The Calendar of the Roman Republic* (1967); W. WARDE FOWLEr, *Roman Festivals of the Period of the Republic* (1899), rather out-of-date hut not superseded; I.S. RYBERG, *Rites of the State Religion in Roman Art* (1955); F. CUMONT, *Les Re'ligions orientales dans le paganisme romain* (1906); L.R. TAYLOR, *The Divinity of the Roman Emperor* (1931); J. FERGUSON, *The Religions of the Roman Empire* (1970); A.D. NOCK, *Conversion: The Old and New in Religion from Alexander the Great to Augustine of Hippo* (1933); M. GRANT, *The Climax of Rome: The Final Achievements of the Ancient World A.D. 161–337* (1968); E.R. DODDS, *Pagan and Christian in an Age of Anxiety* (1965); A. MOMIGLIANO (ed.), *The Conflict Between Paganism and Christianity in the Fourth Century* (1963).

(M.Gr.)

Rome

A capital of kingdoms and of republics and of an empire the armies and polity of which defined the Western world in antiquity and left seemingly indelible imprints thereafter, a city called eternal, as the spiritual and physical centre of the Roman Catholic Church, and a name that evokes major pinnacles of man's artistic and intellectual achievement, Rome, in the late decades of the 20th century, retains all of these attributes: the capital of Italy, a font of religious authority, a memorial to the creative imagination of the past. Probably more than any other city in the West, possibly more than any other in the world, it is a city whose history continues to shape nearly every aspect of its being but, at the same time, whose contemporary consciousness of that history projects it into the very core of modem life.

Location and historical character of the city

Rome is located in central Italy on the Tiber (Tevere) River, 15 miles (24 kilometres) inland from the Tyrrhenian Sea. The Roman countryside, the Campagna, was one of the last areas of central Italy to be settled in antiquity. The city was built on a defensible hill dominating the last downstream, high-banked river crossing where traverse was facilitated by a midstream island.

For well over a millennium, Rome controlled the destiny of all civilization known to European man, then fell into dissolution and disrepair. Physically mutilated, economically paralyzed, politically senile, and militarily impotent by the late Middle Ages, Rome nevertheless remained a world power—as an idea. The force of Rome the lawgiver, teacher, and builder continued to radiate throughout Europe. Although the situation of the popes from the 6th century to the 15th was often precarious—at times tragic, ridiculous, or shameful—Rome knew glory as the fountainhead of Christianity and eventually won back its power and wealth and re-established itself as a place of beauty, a source of learning, and a capital of the arts.

This article focusses entirely on the city of Rome, covering features of the Roman Republic and Roman Empire only as they affected the life or structure of the city. It is divided into the following sections:

For information on related topics, see the articles ITALY; ITALY AND SICILY, HISTORY OF; and ROME, ANCIENT. Aspects of Rome's contributions to Western culture will be found in such articles as RENAISSANCE; VISUAL ARTS, WESTERN; and PAPACY. See also ROMAN RELIGION.

(B.E.)

I. The history of the city

ROME OF ANTIQUITY

Founding and the kingdom. Although the site of Rome was occupied as early as the Bronze Age (c. 1500 BC) and perhaps earlier, continuous settlement did not take place until the beginning of the 1st millennium BC. By the 8th century, separate villages of various iron-using Indo-European peoples appeared, first on the Palatine and the Aventine hills and soon thereafter on the Esquiline and Quirinal ridges. The artifacts and especially the funerary customs of these communities indicate that, from the beginning, diverse culture groups—including Latins, Sabines, and perhaps others—played important roles in the formation of the future city.

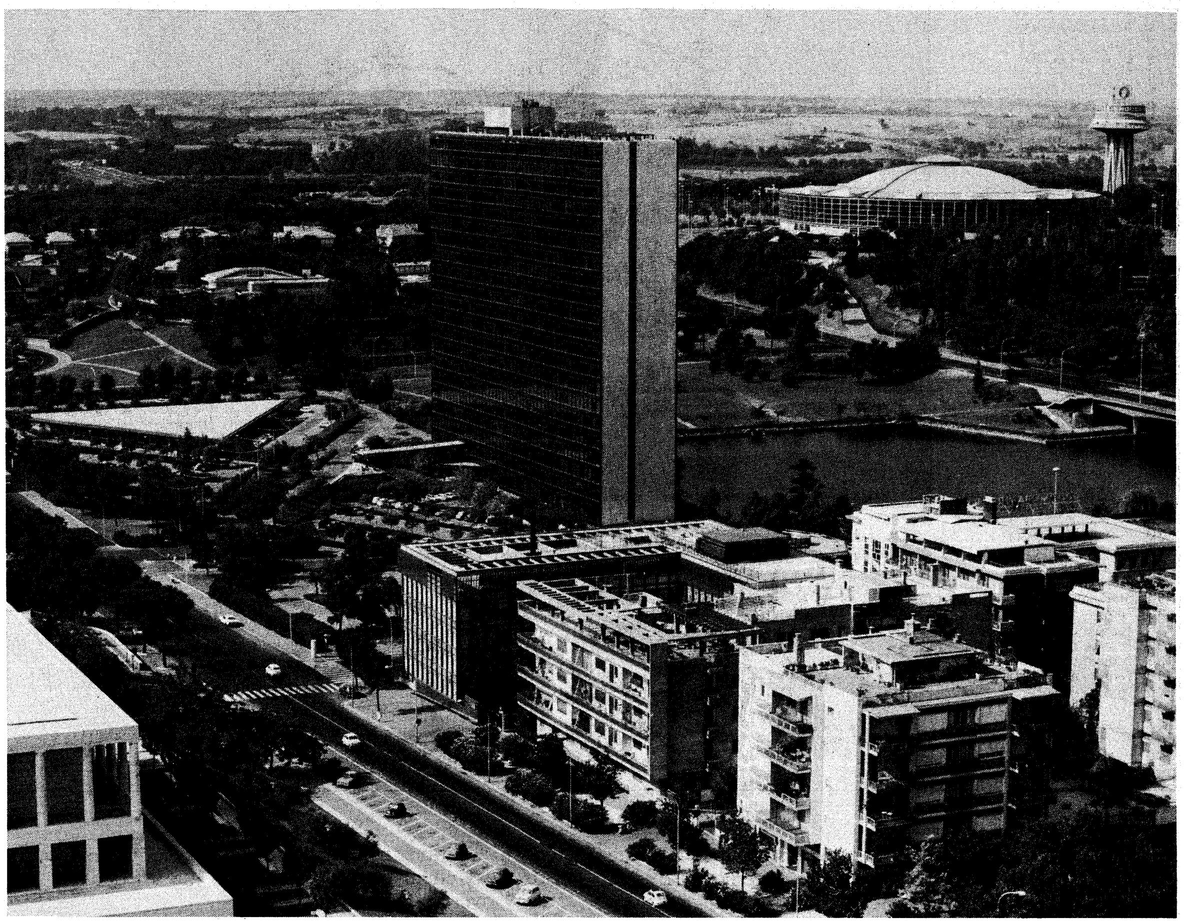
Early peoples and cultures

With the settlement of the valleys between the Palatine, Esquiline, and Caelian hills in the 7th century, the independent villages began to merge. Before the end of this century, the Forum valley, originally used as a cemetery, was partially drained and occupied by wattle-and-daub huts. The mixed agricultural and pastoral economies of the earliest settlements were slowly exposed to commercial contacts with both Etruscan and Greek traders. The formation of a politically unified city probably occurred in the early 6th century BC under the influence of the Etruscan city-states to the north. Under the rule of its kings, traditionally seven in number (the last three probably Etruscans), Rome became a powerful force in central Italy.

During the regal period, social and economic differences began to shape the two classes, patrician and plebeian, whose struggles for political power dominated the early republic. The tribal organization of the populace was replaced by one based on military units, whose composition in the late regal period depended on property qualifications.

The early Roman Republic. The overthrow of the last Roman king and the establishment of the republic, either in 509 BC or a generation or two later, coincided with the decline of Etruscan power in central Italy. The new government under the leadership of two patrician consuls was at first a mixed blessing. Although Etruscan techniques and symbols survived in republican Rome, commercial ties with the Etruscans and with the Greek colonies in southern Italy gradually withered. During the ensuing economic crisis, grain shortages occurred, a problem that was to plague the city intermittently for a millennium and more; the government was forced to make purchases from as far away as Sicily.

Political upheaval followed economic depression. The first major confrontation between the patricians and plebeians in the mid-5th century led to the writing down of the customary laws in the Law of the Twelve Tables (451–450) and to the formation of a plebeian political



The Esposizione Universale di Roma on the outskirts of Rome. The tall building (centre) serves as a corporate headquarters. The dome (right) is that of the covered stadium built for the 1960 Olympics.
Paolo Koch—Rapho Guillumette

organization whose leaders, the tribunes, acted to protect the plebeians from arbitrary patrician actions. In the last half of the 5th century, Rome began again to expand its control over neighbouring territories and peoples, a process that culminated in the conquest of the Etruscan city of Veii in 396.

In 390 Rome suffered a disastrous check when a Gallic army laid siege to the city. After seven months, during which only the Capitoline remained in Roman hands, the Gauls were bought off but left Rome in ruins. The Romans set about reconstructing their city almost immediately, surrounding it with a continuous wall of huge tufa blocks. Later writers attributed Rome's haphazard appearance to the rapid rebuilding during this period; Livy described Rome as looking more like a squatters' community than a planned community. For eight centuries, however, no foreign invader was to breach Rome's walls.

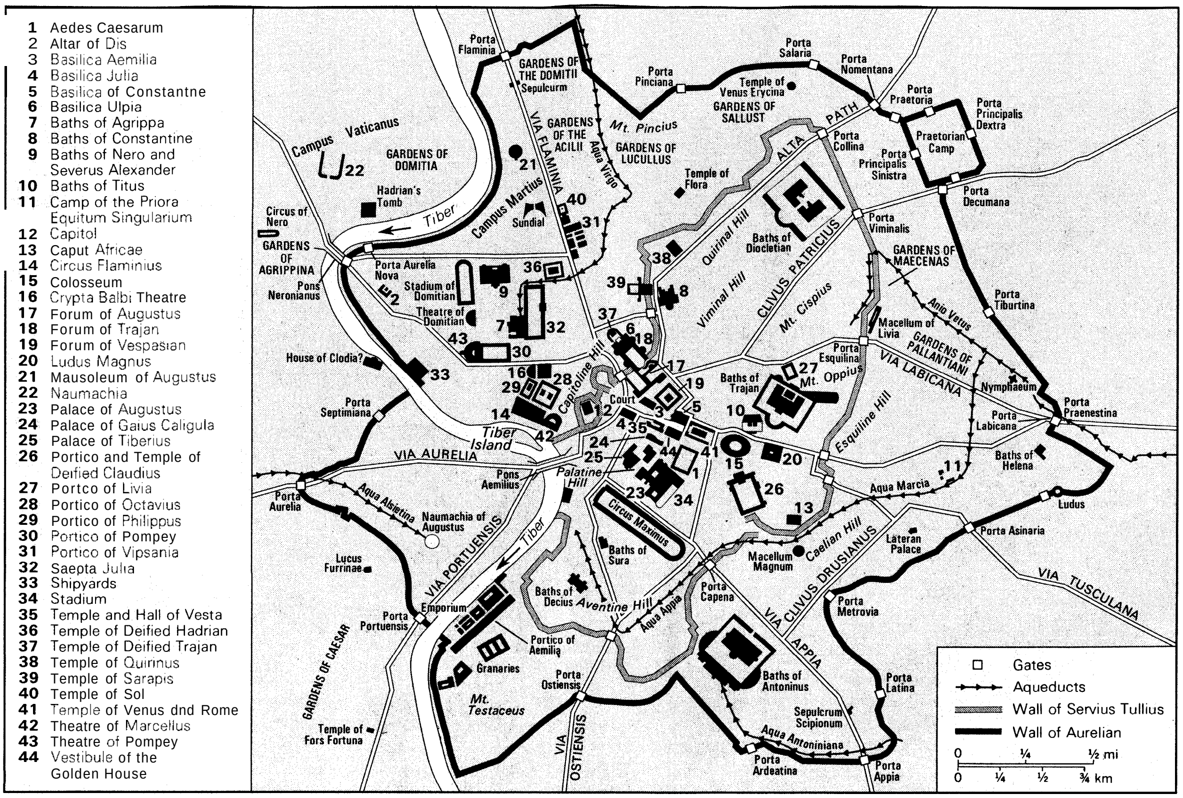
Origins of the haphazard municipal plan

The economic dislocation caused by the Gallic attack helped renew the conflict between the patricians and the plebeians; but, before the end of the 4th century, through a series of judicious compromises, the plebeians had won access to all of the offices of the state, and the actions of the plebeian assembly (plebicites) had been made legally binding on all Romans. Economic legislation dealing with debt and land distribution was directed toward relieving the distress of the lower classes.

The city of world power. The remarkable though largely unplanned territorial expansion of Rome between 375 and 275 brought lasting economic gains. With control of all of peninsular Italy, Rome established colonies on some of the conquered territories and elsewhere assigned lands to individual Roman citizens. The nearly 60,000 holdings distributed before the middle of the 3rd century helped to solve the pressure of Rome's land-hungry population; nevertheless, by c. 250 the city's population had grown to almost 100,000. The booty from conquests also helped to defray the costs of such public works as the building of temples and roads and the improvement of the city's water supply. By the early 3rd century, two aqueducts carried fresh water into the city.

In 264 Rome was drawn into a war with Carthage, the great Phoenician emporium in North Africa. After more than a century of conflict, Rome emerged as the strongest power in the Mediterranean; but the acquisition of an empire, which, for the most part, had not been the conscious desire of the Roman people, brought new social and economic problems to the city itself. The Second Punic War (218–201) saw the devastation of large areas of the peninsula by invading troops from Carthage, led by the famous general Hannibal; much land was abandoned and many peasants sought refuge in Rome. The growing requirements of a standing army depopulated the countryside and concentrated veterans in the city. The Roman nobility, prohibited by law and by custom from investing in commerce or industry, profited from the economic distress of the peasantry by buying up large tracts of land in central and southern Italy. Slaves, whom Rome's wars in the Mediterranean made available in large numbers, were introduced into Italy as farm labourers and herdsmen, causing further dislocation among the free peasantry. In general, the Roman economy lagged well behind the political development of both city and empire.

The late republic. During the 2nd century, the rapid growth of the urban population and the extension of Roman citizenship led to the effective disenfranchisement of the urban vote. The Senate, now the chief policy-making body of the Roman state, was preoccupied with the problems of the empire and too often ignored the needs of the city. With no separate municipal government, public works and the management of food and water supplies were left to private initiative or to amateur pub-



Rome during the imperial period.

lic officials. Nevertheless, some progress did occur. Some of the main streets were paved; drains were covered; and several large basilicas and a new row of shops were built in the Forum. The first stone bridge across the Tiber was completed in 142, and the first high-level aqueduct was erected in 144, allowing settlement on the higher ground of the city's eastern ridges. From the early 2nd century, the river port at the base of the Aventine acquired new warehouses and docking facilities.

These and other projects, however, were inadequate to deal with the growing urban proletariat increasingly swollen with slaves and freedmen. Crowded into jerry-built apartment houses (*insulae*) and with only minimum employment opportunities in what was an essentially non-industrial city, the lower classes were surviving on the sporadic public-works projects of the state and the largess of the rich before the end of the 2nd century. Rome had, moreover, neither police nor fire protection.

Impact of the Gracchi on municipal squalor

The Gracchi—Tiberius and later Gaius—attempted to deal with the problems of urban unemployment and rising food prices, first by advocating the re-establishment of a small farmer class in Italy, then through the subsidization of the grain supply for the poor. Gaius Gracchus also encouraged public expenditure on roads and buildings. Coupled with currency reforms and heavy government spending, these measures partially restored prosperity to Rome in the late 2nd century, but the basic structural faults in the city's economy and political life remained.

During the civil strife that occupied most of the first half of the 1st century BC, both population and problems multiplied in Rome. The creation of private armies attached to the Roman nobility offered employment to some of the urban lower classes but contributed greatly to the political violence that eventually spelled the end of the republic. Securing an adequate supply of cheap grain offered possibilities for the political manipulation of the urban masses. By the middle of the century, perhaps as many as 500,000 persons were receiving free grain. The upper classes became more interested in luxurious living and their tastes were matched in the public sphere by the building programs of Sulla and Pompey. Public buildings and theatres paid for with tribute and booty enhanced Rome's beauty but did not make a more livable city. In addition, heavy migration to Rome, especially from the Hellenistic east, added to the burdens of the already overcrowded city.

Municipal reforms of Augustus. Julius Caesar, the first to try to deal with the problems of Rome in a systematic way, did not live long enough to carry out his plans, which included canalizing the Tiber and building up the Campus Martius. His adopted son and successor, Augustus, attempted to transform Rome into a worthy capital for the new empire. Although his claim that he found the city brick and left it marble is exaggerated, Augustus and his colleagues did provide it with many fine public buildings, baths, theatres, temples, and warehouses. Such construction projects, together with the restoration of old buildings, provided employment for the urban masses; but the lack of any overall city planning left them to live in the unsafe and unsanitary tenements amid the narrow, winding streets and alleys of old Rome. Agrippa, a friend and supporter of Augustus, used his own immense wealth to further enhance the city's beauty and improve its water supply.

Advent of a professional city management

Augustus' reorganization of the administration of the city and his institution of certain public services were a significant break with the republican past. In 7 BC he divided Rome into 14 *regiones* ("wards") and these into *vici* ("precincts"), each with officials who performed both administrative and religious functions. The office of urban prefect, which Augustus revived in *c.* 26 BC, did not become permanent until later, but in the late empire the post became the most important in Rome.

In response to an obvious need, Augustus organized a fire brigade in 21 BC, placing a number of public slaves under the command of aediles, officials in charge of streets and markets; after a bad fire in AD 6, he established a corps of professional firemen *(vigiles),* comprising seven squads, or cohorts, of 1,000 freedmen apiece. The *vigiles* also had minor police duties, especially at night. He sought to impose order in the often violent streets by creating three cohorts under the command of the urban prefect; their main duty was to keep order in the city, and they could call on the Praetorian Guard for help if necessary. Altogether, Augustus saw to it that the amateur system of Roman municipal administration was replaced by a more professional and permanent set of institutions

—a work that probably contributed more to making Rome a great city than all of his marble monuments.

Contributions of later emperors. For the most part, the successors to Augustus continued his administrative policies and building program, though with less innovation and more ostentation. Claudius began a great port near Ostia, at the mouth of the Tiber, to facilitate grain shipments directly to Rome. Commerce remained largely in private hands, with public officials acting to ensure a regular supply and to prevent speculation.

Nero can be credited with introducing the most up-to-date ideas on town planning, though at a terrible price. The great fire of AD 64 destroyed large sections of the city. In the devastated areas Nero built new streets and colonnades as well as his fabulous Golden House, and he encouraged private citizens to build more spacious and more fireproof houses and apartment buildings with better access to the public water supply. Although Nero made Rome a more pleasant city in which to live, his measures did not prevent other devastating fires such as the one in 191 that gave Septimius Severus the opportunity to rebuild the city.

Other emperors in the late 1st and early 2nd centuries added to the glory of the imperial house and the amenities of Roman life—grandiose imperial forums, temples, arches, baths, and stadiums. Trajan's Forum, with its complex of buildings and courtyards, and his market, with its tiers of shops and its great market hall, represent, in the judgment of many historians, the supreme achievement of city planning in Rome. Trajan's Column, which narrates his victories beyond the Danube, was recognized as without peer even in the Christian Middle Ages. Hadrian left two enduring structures in Rome: the great domed Pantheon and his mausoleum, which in AD 590 was renamed Castel Sant'Angelo.

Populace and economy at the height of the empire

In the late 1st and early 2nd centuries Rome was at the peak of its grandeur and population, which has been estimated at over 1,000,000 persons but was probably less. It was kept at a high level by a steady stream of immigrants, both slave and free, from the provinces and beyond—although life expectancy in the city was probably lower than elsewhere in the empire. Rome's famous paved streets, water supply, and sewage system, however, should not be overestimated; even after the reforms of Nero, large numbers of the urban inhabitants continued to live in expensive, poorly built, overcrowded, and unheated slums without water or cooking facilities. The arena and the public bath relieved some pressures of high density and physical squalor, but Rome's refined technology was applied haphazardly to the problems of urban social organization. Garbage was usually dumped into the Tiber or pits on the city's outskirts.

Rome was a city of consumers, both rich and poor, and never a great industrial or commercial centre. The small shop was the basic unit of production and distribution through the imperial period, and the numerous trade associations served social and religious functions until they were enveloped in the economic regimentation of the late empire. Although Rome far surpassed any other ancient city in size and monumental splendour, its minimal economic and social achievement augured ill for the future.

Slow decline of the late empire. Rome's population probably began to decline in the late 2nd century. At the height of an outbreak of the plague in the reign of Marcus Aurelius, 2,000 persons a day are thought to have died. The economic and political disasters of the 3rd century did little good for Rome. In the 270s the walls built by Aurelian were more a symbol of the danger of barbarian attack than a restoration of Rome's grandeur.

By the time Diocletian (ruled 284–305) reformed the imperial government and ushered in the period of relative prosperity symbolized in his great baths, Rome was no longer the administrative capital of the empire. The founding of Constantinople merely confirmed Rome's loss of political primacy. Constantine, however, did much to restore the buildings and monuments of imperial Rome. In addition, his patronage of Rome's small Christian community laid the foundations of Christian and papal Rome of the medieval and modern periods.

Rome in the 4th century remained, nonetheless, a distinctly conservative and pagan city dominated by proud senatorial families. When the Visigothic army of Alaric first threatened the city in 408, the Senate and the prefect proposed pagan sacrifices to ward off the enemy, and even the pope would have allowed them to be performed in secret. In 410 Alaric seized Rome and allowed his troops to pillage the city for three days; much booty was taken, and many Romans fled.

It is unlikely, however, that the monuments of Rome suffered extensive damage. Its churches, for the most part, were spared. Even the longer, 14-day sack of Rome by the Vandals in 455 did less damage than the Romans themselves. In the 4th and 5th centuries, the emperors repeatedly legislated against those who were stripping buildings and monuments for their materials, especially the marble. By the middle of the 5th century, the population had dropped well below 250,000.

THE CITY OF THE POPES

Decay of imperial authority. Within a decade of Theodoric's death in 526, Justinian began his attempt to restore Roman imperial rule in the West. His ultimate success was disastrous for Italy and for Rome. Three times Rome was under siege; its aqueducts were cut, and once it was abandoned by its inhabitants. By the end of the century, with the urban population under 50,000, civil authority and the responsibility for protecting the city were in the hands of the church. Pope Gregory I tried to provide an adequate urban administration, and for nearly two centuries his successors played a similar role.

Bases of the papacy's temporal power

In the middle of the 8th century, when the Byzantines were no longer able or willing to supply Rome with adequate military aid, the papacy turned to the Franks. The "Donation" of Pepin—who owed his new title as king of the Franks in part to the Pope—and that of his son Charlemagne were the theoretical foundations of the temporal power of the papacy. In 774 Charlemagne conquered the Lombard kingdom, and in 800 he was crowned emperor by Pope Leo III and acclaimed by the people of Rome. The period of the late 8th and early 9th centuries was one of vigorous building and restoration of churches in Rome.

Factional struggles: papacy ad nobility. The decline of Carolingian authority in Italy led to the renewal of family and factional struggles. After the Muslims plundered St. Peter's and the outlying areas of Rome in 846, Pope Leo IV built a wall around the area of the Vatican, thus enclosing the suburb that came to be known as the Leonine City. From the late 9th through the middle of the 11th century, Rome and the papacy were controlled by various families from Rome's landed nobility, with brief interludes of intervention from the German emperors.

After decades of dispute between the Roman nobility and the papacy, the latter was able to establish an uneasy peace in Rome by the end of the 11th century. Much rebuilding was necessary after the Norman sack of 1084. Generally, the reformed papacy, begun under Leo IX (1049–54), was supported and financed by new Roman families such as the Frangipane and the Pierleone, whose wealth came from commerce and banking rather than landholdings. By the late 11th century the seat of the church had begun to draw many pilgrims and prelates to Rome, and their gifts and expenditures on food and housing stimulated a considerable flow of money. Although Rome had a population of fewer than 30,000 (occupying less than one-quarter of the lands within the old walls), it was becoming once again a city of consumers dependent on the presence of a governmental bureaucracy.

Emergence of the Roman commune. The Roman revolution in 1143 had fundamentally the same goals as other contemporaneous communal movements in northern Italy: freedom from episcopal (in Rome's case, papal) authority and control of the surrounding countryside. The revival of the Roman Senate and other echoes of the Classical past perhaps owed something to the preaching of Arnold of Brescia, a priest and monk, who said strong things against ecclesiastical property and church interference in temporal affairs. Rome's new republican constitution survived both papal and imperial attack alike, and in 1188 Pope Clement III recognized the communal government. In theory, the senators were to become papal vassals, but, in fact, the Pope had to make large cash payments to the senators and other communal officials. In the 1190s a single senator was able to exercise wide authority in the territories surrounding Rome.

Papal and communal authority in conflict

Pope Innocent III made it his first order of business to secure a firm papal position in Rome and in the Vatican. Only moderately successful, he found it expedient to support the Roman commune's expansionist policies. Territorial rivalry between Innocent's family and the Orsini led to rioting and finally open warfare in the streets of Rome in 1204, during which siege machines destroyed many ancient buildings. After a settlement, Innocent's many charitable projects won him Roman support. Gregory IX and the Roman commune clashed over Rome's expansionist policies and its claims to the right to tax the clergy and church property. Bitter struggles with the Hohenstaufen emperor Frederick II as well as the varied interests of Rome's leading families — the Orsini, the Savelli, the Annibaldi, and, above all, the Colonna—complicated the situation. After Frederick's death, an anti-papal regime promoted a rising middle class and a resurgence of the commune.

Period of the Avignon papacy. Few popes in the second half of the 13th century were able to reside in Rome. In the 1280s and 1290s, Rome was torn by the bitter rivalries among the Colonna, the Orsini, and the Annibaldi families, a discord encouraged by Pope Boniface VIII. In 1309 Clement V moved the papal residence to Avignon in France; Rome was left to its factional strife and its economic impoverishment.

In spite of sharp rivalries, Roman and papal interests had often coincided throughout the 13th century. Since Rome was never an important industrial or commercial city, its citizens, from the small shopkeepers and innkeepers to the great banking families, had depended economically on the presence of the papal Curia and the large numbers of pilgrims, prelates, and litigants it brought to Rome. The many brick campaniles of its Romanesque churches and the fortress towers on the palaces of its leading families symbolized Rome's ecclesiastical character; but, with a population of never more than 30,000 in the 13th century, it retained a village air for all its urbanity and classical aspirations. Most of the populace was concentrated around St. Peter's and in the low-lying areas of the Campus Martius and Trastevere; large sections of the city within the old Aurelian walls were pastures, gardens, and vineyards and wastelands.

Ambience of the city in the late Middle Ages

The popes in Avignon were able to maintain a tenuous rule over the city, especially under Benedict XII (1334–42). The brief popular revolution (1347) of Cola di Rienzo — who, styling himself tribune of Rome, combined apocalyptic visions with ideas of a renewal of Rome's ancient glories — had more dramatic than political impact. The terrible mortality of the Black Death reduced Rome's population to less than 20,000, and the city staggered through the last half of the 14th century still racked by factional strife. The return of the papacy from Avignon in 1377 did not help. Around 1400, Rome was described as a city filled with huts, thieves, and vermin, and in the neighbourhood of St. Peter's wolves could be seen at night.

The city of the Renaissance. The entry of Pope Martin V (a member of the Colonna family) into Rome in 1420 marked the beginning of the Renaissance city and of the absolute papal rule that lasted until 1870. Although Martin was neither a builder nor a patron of the arts, he laid the foundations of government that made Rome the capital of a Renaissance state. From this period, the apostolic vice chamberlain, as governor of Rome, controlled municipal offices, communal finances, and the statutes of the city. The Roman commune was transformed into a unit of authoritarian papal rule, and the papal states increasingly came under the firm control of papal officials.

From the pontificates of Nicholas V (1447–55) and, especially, Sixtus IV (1471–84), the squalid narrow streets of medieval Rome were widened and paved, and new Renaissance buildings replaced crumbling structures. At the same time, the monuments of ancient Rome suffered further damage as they were torn apart for their building materials, and their marble went too often into the lime kilns rather than into new structures. The popes attracted scholars and artists from across Italy, and, by the end of the 15th century, Rome was the principal centre of Renaissance culture. The high point was reached under Leo X (1513–21), with his plans for a new St. Peter's and his patronage of such artists as Michelangelo and Raphael. Rome flourished economically under the Renaissances popes. Banking and the exploitation of alum deposits near Civitavecchia by the popes (with the help of the Medici family of Florence) stimulated a flow of capital into the city. Although Rome once again had become a great consumer of imported luxuries, it still had little large-scale industry or commerce.

Destruction of ancient buildings

EVOLUTION OF THE MODERN CITY

Rebuilding and repopulation. The sack of Rome in 1527 by the armies of Emperor Charles V ended the city's pre-eminence as a Renaissance centre. In eight days, thousands of churches, palaces, and houses were pillaged and destroyed. But, even under the repressive rule of the Counter-Reformation papacy, Rome recovered; a new era of construction was begun, culminating in a vast program of city planning by Sixtus V (1585–90) and his architect Domenico Fontana. New streets and squares were laid out, obelisks raised, the Lateran and Vatican palaces rebuilt, and aqueducts repaired. Fortunately, his project to convert the Colosseum into a wool factory to provide employment for Rome's prostitutes came to nothing.

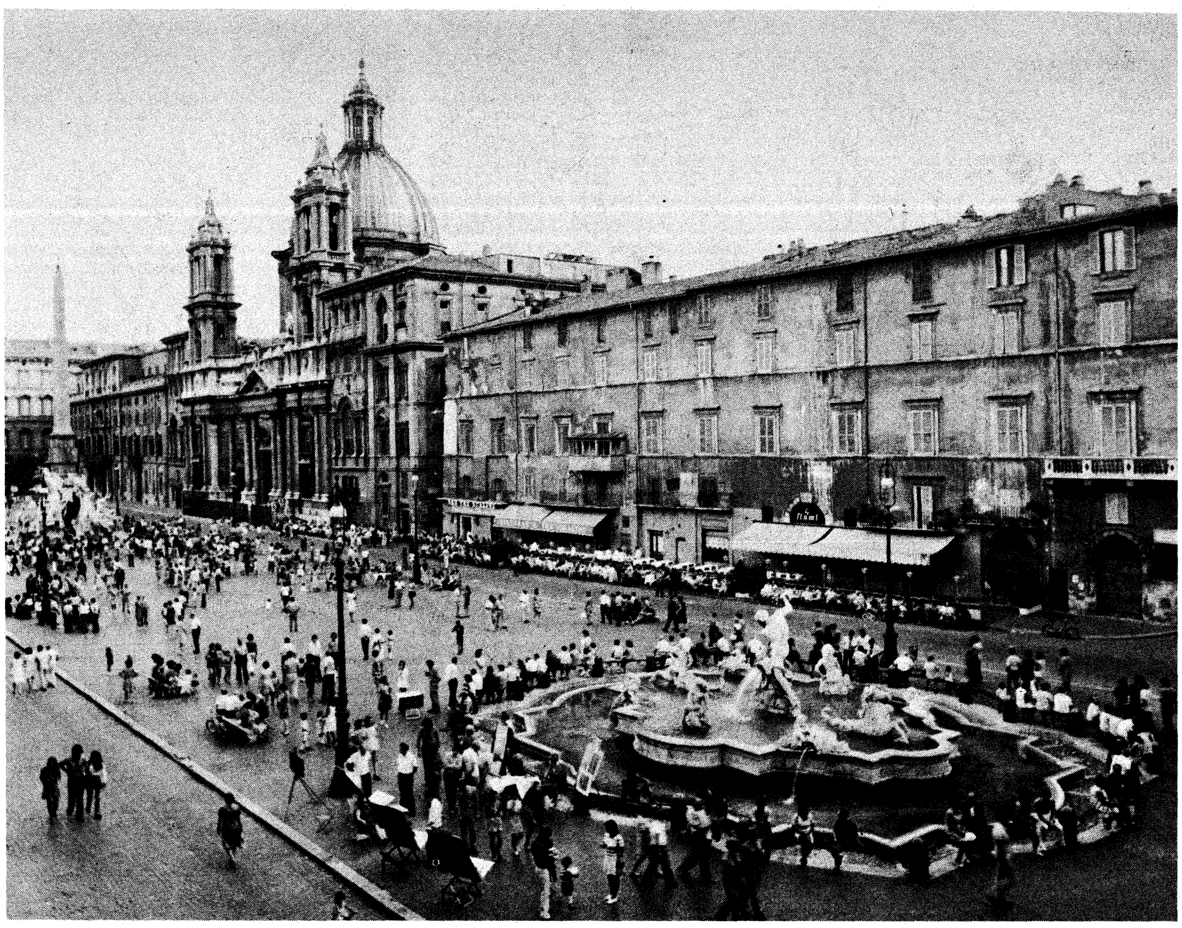
By 1600, Rome was again a prosperous cosmopolitan city. A great influx of new inhabitants attracted by employment opportunities in the papal bureaucracy and related service industries increased Rome's population to over 100,000. Much of the big business of the city remained in the hands of foreigners, however, for the wealth and power of the Roman nobility was based on land and ecclesiastical officeholding.

Decline and fall of the papal empire. In the 17th and 18th centuries Rome's noble families built fine palaces and patronized the arts while manoeuvring to win high positions in the church hierarchy. The highest prize of all, the papal crown, brought wealth and status to the wearer's family. But as corruption and bribery within these circles became a way of life, the influence of the papacy and of Rome declined throughout Europe and even throughout the Papal States. Although Sixtus V had created one of the best-planned cities in Europe, by the 18th century Rome was still a backward town, with poorly paved streets on which there were no road signs nor public lighting and little sanitation. To foreign observers, the Romans, from the most aristocratic families to the poorest classes, seemed to lead lives of provincial vacuity unconcerned with anything outside Rome. The population reached 165,000 by 1790, but as many as one-quarter of the inhabitants were employed in the petty bureaucracy that overran the city.

Continued backwardness and provinciality

The armies of Napoleon occupied Rome for the first time in 1798, and a republic was declared; but in 1809 Rome and the Papal States were annexed into the French Empire. The return of the pope to Rome in 1814 led to a long period of repression and reaction, though popes Leo XII and Gregory XVI promoted educational improvements and new public baths and hospitals. With the liberal attitude that characterized the early part of his reign, Pope Pius IX (1846–78) granted Rome a constitution in 1848; but, after the Revolution of 1848–49, he became an archconservative, attempting with French support to save the temporal power of the papacy and to stave off the modem world.

Capital of a united Italy. Most of the Papal States were included in the Kingdom of Italy, proclaimed in 1861, but Rome was excluded. Attempts by Garibaldi to



The Piazza Navona with the Church of S. Agnese designed by Borromini and, facing it, Bernini's "Fountain of the Four Rivers"; at right is Bernini's "Fountain of the Moor."
Garofalo—Grimoldi

capture the city in 1862 and 1867 were unsuccessful, but the withdrawal of the French garrison supporting Pius allowed Italian troops to enter Rome on September 20, 1870. After a plebiscite in October, Rome became the capital of a united Italy. Pius refused to accept the government's offer of settlement, choosing to style himself a prisoner in the Vatican. The situation was not resolved until 1929, when the Lateran Treaty between Pius XI and Mussolini recognized the pope's sovereignty within Vatican City.

Rome's population grew rapidly after 1870, passing the 500,000 mark before World War I and reaching more than 1,000,000 by 1930. Its area of settlement also expanded for the first time well beyond the old walls of the ancient city. During the Fascist regime of Benito Mussolini in the 1920s and 1930s, Rome was transformed into a modern capital, with grandiose new avenues and pompous buildings. Mussolini's encouragement of archaeological excavation contributed to the revelation and preservation of many of the monuments of classical Rome. Throughout the 20th century, Rome has been a great administrative and tourist centre, though it still lacks the large-scale commerce or industry characteristic of most modern urban development. (R.R.R.)

Mussolini's transformations

II. The contemporary city

The city of the seven hills, the city of treasures and tourists, the honey-coloured city of fountains and cupolas lies mostly within the old city walls. The so-called Servian Wall, built almost certainly 12 years after the Gauls' destruction of Rome in 390 BC, enclosed most of the Esquiline and Caelian hills and all the other five. It was built into ramparts that dated from the early republic or even the late kingdom. Although Rome grew beyond the Servian defenses, no new wall was constructed until Aurelian began building in brick-faced concrete in AD 270. Almost 12 miles long and girdling about four square miles, this is the wall that Italian troops had to assault and breach to claim their capital in 1870, and it is still largely intact.

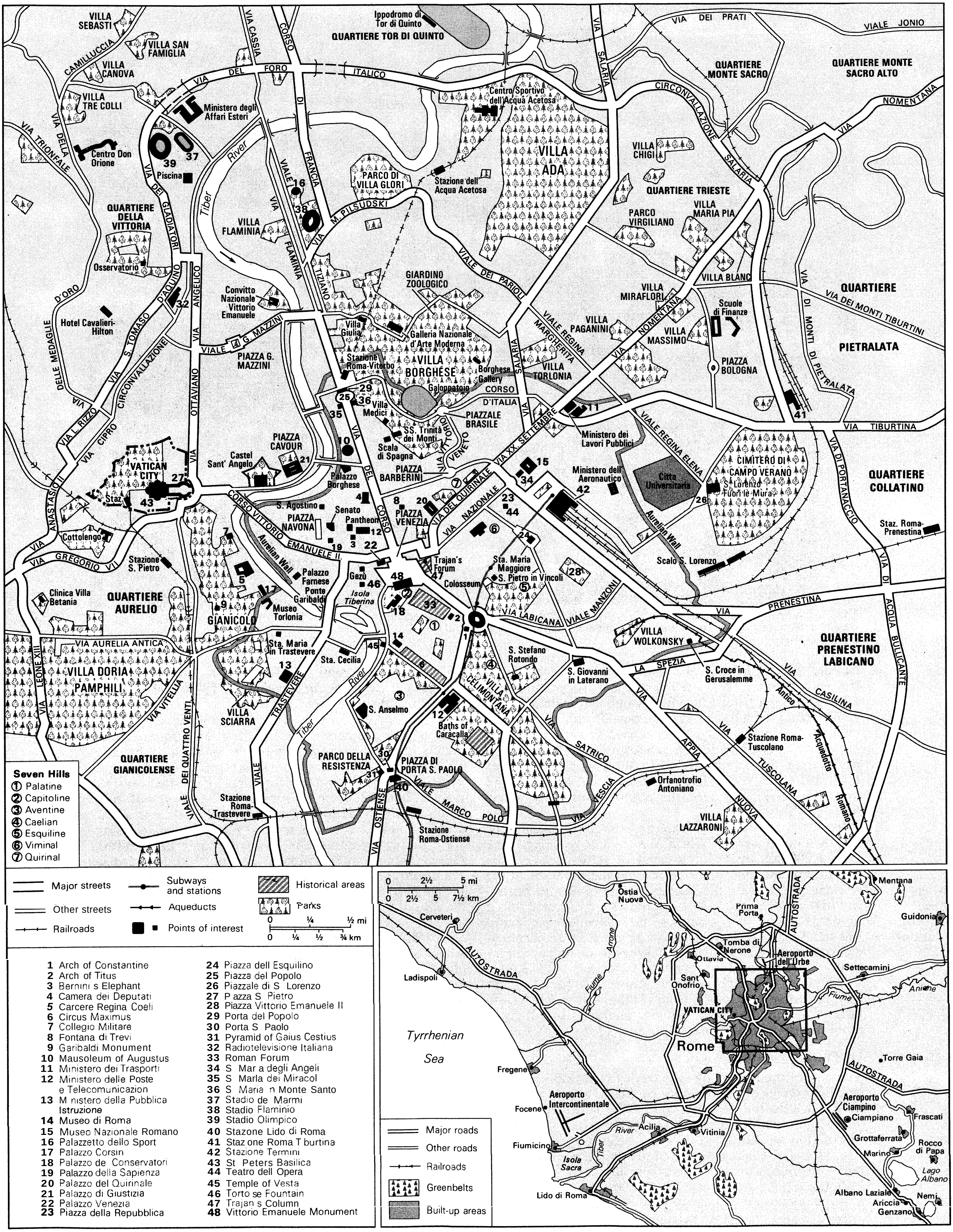
Extent of modern Rome

The ancient walled city of Rome embraces only 4 percent of the modern municipality's 582 square miles (1,508 square kilometres) and is the smallest of the city's 12 administrative zones, containing less than 10 percent of the more than 2,780,000 Romans. The walled centre is divided into 22 *rioni* ("districts"), the names of most dating from Classical times, while surrounding it are 32 *quartieri urbani* ("urban sectors") that began to be absorbed officially into the municipality after 1911. Within the city limits on the western and northwestern fringes are six large *suburbi* ("suburbs"), while beyond the municipal boundaries the commune of Rome about doubles the area of the city itself.

THE AMBIENCE OF ROME

About six miles out from the centre of Rome, a belt highway describes a huge circle around the capital, tying together the antique roads that led from everywhere to Rome, the Via Flaminia, Via Aurelia, Via Appia. Melancholy masses of modern apartment buildings rise in the districts outside the centre, in which the small amount of contemporary construction is inconspicuous. Street frontages and show windows are often rebuilt to keep pace with the times, and the Romans succeed in harmonizing the new, the simply old, and the antique with a talent that they have demonstrated since the first extensions of the republican Forum were made under the emperors.

Small as it is, the old city contains 300 hotels, 300 *pensioni,* more than 200 palaces, 20 churches, eight of the city's major parks, the residence of the Italian president, the houses of Parliament, offices of city and national government, and the great historical monuments, in addition to thousands of offices, workshops, restaurants, and bars,



Central Rome and (inset) its metropolitan area.

It is there that the millions of tourists seem to descend annually, to supplement and, in large measure, support the local populace.

Administrative mazes. Rome is one of the most beautiful and exciting capitals in the Western world. According to local authorities, it is also the filthiest, noisiest, and most heavily indebted city in Italy.

The city that introduced the public sewer lacks an adequate sewage system. Four treatment plants planned to end the flow of raw outfall into the Tiber remained unbuilt in the early 1970s because the $160,000,000 was not available to pay for them. The municipal debt in 1971 was $2,400,000,000; and interest on this debt amounted to $144,000,000 of the $200,000,000 annual municipal revenue in 1971.

Housing shortage

The city that invented both concrete and the apartment house (insulaj suffers a perennial housing shortage. At Christmas 1970 Pope Paul VI appealed to the municipality to house the 70,000–100,000 inhabitants of shantytowns on Rome's outskirts. The housing shortage persists because of the incessant arrival of job-seeking migrants from all over Italy but mostly from the impoverished south. Rome is not industrial, but it does have government and tourism to provide employment, and its growth continues unabated. By 1968 all the plans, powers, agencies, and even state building funds were in hand, but three things impeded construction: first, land cannot be built upon until the municipality supplies public services and schools (the city is so short of school space that schools operate classes in three successive shifts for 12 straight hours a day); second, Roman politics are more Byzantine — more labyrinthine and convoluted — than a 5th-century mosaic; third, the notorious glacier-slow Roman bureaucracy can, by paper shuffling alone, delay an approved project up to five years. Life in Rome remains an endless paper chase through the obscure corridors of petty authority.

Traffic becomes a typical Roman dilemma because much of the municipal revenue is derived from the 900,000 automobiles and 100,000 motor scooters that help render city life difficult. The average noise during waking hours is at or above the level that gradually induces deafness, whereas speed of motor traffic, in spite of the audacity and acuity of the drivers, is four miles per hour. A traffic analysis in 1971 projected that in six years Rome would have enough motor vehicles to cover every foot of its road surface.

In 45 BC Julius Caesar forbade any wagon to be led or driven during the daytime within the continuous built-up area of Rome. Unfortunately, the police force required for enforcement was seriously under strength so that generally during the daytime almost no traffic police were on duty. "Where can you find lodgings that give you a chance of sleep?" a celebrated writer demanded. "The roar of the wheeled traffic in the City's narrow, winding streets and the shouts of abuse . . . ," thus wrote Juvenal, who lived in Rome in the late 1st and early 2nd centuries AD. In 1970 and 1971 municipal authorities introduced stringent rules to keep some of the traffic out of the centre, and in 1972 the city experimented inconclusively with free buses to determine whether vehicular traffic could be reduced.

Deterioration of the city's monuments was being accelerated by traffic fumes and vibration, yet the monuments themselves impeded the one undertaking that could reduce road traffic: construction of a subway. Mussolini decreed the building of a subway from the central railway station, and by 1955 it was in operation. In 1959 a comprehensive metropolitan subway system was approved. After five years of tunnelling through the bureaucracy, the first line began tunnelling a route some 14 miles long under the streets. It was diverted to protect monuments, halted when it unearthed archaeological remains, and, at long last, resumed again, with completion expected in the mid-1970s. Two future lines will go three times deeper, to 60 feet, to avoid disturbing any more buried heritage.

The people and their pace of life. The knowledge that Rome is eternal, that nothing lasts but nothing changes, gives rise to the local watchword, pazienza ("patience"). In this overcrowded, understaffed city, pazienza is demonstrated everywhere, every day. Except for brief, lowering, summer-lightning flashes of an underlying volatility, the Roman is apt to be cheery and courteous, a little less operatic in his reactions than many other Italians.

Family life

In Rome, as in the rest of Italy, all children are godsends and are demonstrably, publicly loved, patted, petted, cuddled, and kissed. Unmonied families make sacrifices to provide the biggest possible dolls and the flashiest possible tricycles. This continues far into life, with the man playing the role of adored but respectful princeling to his queen mother and imperious but indulgent king to his wife and children. In society outside the family the important thing is bella *figura*, or keeping face. Thus the *dottore* (the only degree the university of Rome gives is the doctorate) salutes the street sweeper as capo ("chief"), a gesture of respect called for by the uniform.

For 1,000 years, to be a citizen of Rome was to hold the keys to the world, to live in safety, pride, and relative comfort. Today there is still considerable pride in being a Romano di *Roma*, a Roman Roman. Among such are the "black nobility," families with papal titles who form a society within high society, shunning publicity and not given to great intimacy with the "white nobility," whose titles were conferred by mere temporal rulers. They avoid the "jet set," the loose group of affluent international itinerants for whom Rome is a major and frequent stopover in their worldwide wanderings.

In the 1970s the cinematic wing of the jet set continued to congregate at the café tables ranged on the plane-tree-shaded sidewalks of the Via Vittorio Veneto, a street of grand hotels, airline offices, and government buildings. Laid out in 1887 from the Villa Borghese gardens to the Piazza Barberini, it runs downhill in a dogleg, Italians preferring the right-hand side, foreigners coagulating on the other. During the 15 or so years of peak prosperity in Italian film making, about 1950–65, international film celebrities abounded and clouds of beautiful career hopefuls drifted among the tables, making the Via Veneto one of the most intriguing — in both senses of the word — streets in the world. In the 1970s cinema production was depressed; but hope was thicker than the exhaust fumes and dreams as bright as the neon signs, and the street remained gaily and expensively animated until long after midnight.

At the same hour, less glittering Romans can be found in the Piazza Navona, on the flat plain in the bend of the Tiber that was the Campus Martius of classical times. The piazza retains the shape and some of the remains of Domitian's circus (AD 81–89), which remained intact until at least 1450. This is far more typical of central Rome than the Via Veneto, a mere centenarian and therefore a new street. Mussolini's regime cut some new routes through the city, mainly to render historic sites more accessible, but 20th-century streets are rare in the historic centre.

The most nobly Roman of the Romans

The inhabitants who consider themselves the most nobly Roman of them all are the people of Trastevere (Across the Tiber). They have been in their neighbourhood for a very long time, although they are of neither pure nor primordial stock. Trastevere was the quarter for sailors and foreigners, whereas the founding fathers eastward across the river were soldiers and farmers. In the Middle Ages a number of palaces were the homes of powerful families, and palaces continued to be built during the Renaissance (Palazzo Farnesina) and even in the 18th century (the Palazzo Corsini). Some authorities — not all from Trastevere — claim Sta. Maria in Trastevere as the oldest church in Rome, pointing out that under the empire the district was the home of Orientals with alien religions, among them a goodly number of Jews proselytized by SS. Peter and Paul. It is said that Alexander Severus (reigned 222–235) permitted Christians to foregather at this site under the leadership of Pope St. Calixtus I, and it is recorded that Pope St. Julius I either raised or rebuilt a church there in 341–352. Today's church is largely 12th-century Romanesque, with a beguiling mosaic facade.

Sidewalk cafe along the Via Vittorio Veneto.

Paolo Koch—Rapho Guillumette

Over the millennia the area has lost little of its vigour. The people have maintained the earthiest of Roman accents, their taverns have remained generally faithful to simple fare, robust wine, and the unison bawling of irreverent songs. One of Rome's few secular statues—a top-hatted marble effigy of Giuseppe Gioacchino Belli, a 19th-century satirical dialect poet—stands near the Ponte ("bridge") Garibaldi.

Most of the streets are still narrow and without sidewalks, appearing only on the most detailed maps and baffling taxi drivers who do not live there. Every 100 paces or so the haphazard cobbled lanes open upon some surprising, small plaza with a church, a palace, a cloister, or a group of cafés. In recent years rents have been rising, and the occasional commercially folkloric restaurant has appeared because artists, actors, and their friends have been moving to Trastevere in response to the unaffected village ambience and the warm humanity of the local folk —who are taking their profits and moving away to modern flats.

MAIN STREETS AND THEIR MONUMENTS

The Via del Corso

The main street in central Rome is the Via del Corso, an important thoroughfare since classical times, when it was the Via Flaminia, the road to the Adriatic. Its present name comes from the horseraces (corse) that were part of the Rome carnival celebrations. From the foot of the Capitoline Hill, the Corso runs to the Piazza del Popolo and through a gate in the city wall, the Porta del Popolo, there to resume its ancient name. It begins spectacularly with the Vittoriano, the monument to Victor Emmanuel II, first king of united Italy. The nation's unknown soldier was interred there after World War I. A Neo-Baroque marble mountain, it is the whitest, biggest, tallest, newest (1911), and possibly the most pompous of Rome's major monuments. Useful as well as ornamental, it contains a museum of the 19th-century cultural revival, a police station, and rare Roman public restrooms.

Along the street among the smart shops are five churches, eight palaces (and one palazzetto), and the column of Marcus Aurelius. The first church is S. Marco, the first of Rome's parish churches to be built (c. AD 336) on the plan of a classical basilica. The present church, third on the site, dates from the 9th century and was restored in the 15th by the Venetian Pope Paul II, who built the Palazzo and the Palazzetto Venezia around the church in 1445, when he was cardinal, enlarging the residence when he became pope. Thereafter, the basilica's priest was always a Venetian cardinal, sharing the palace with the Venetian embassy. Mussolini had his headquarters there and harangued the crowds from the balcony from which Paul II had cheered the carnival races and given his papal benediction. The palace is now a Renaissance art museum and the Biblioteca dell'Istituto Nazionale d'Archeologia e Storia dell'Arte (Library of the National Institute of Archaeology and Art History).

While her son Napoleon languished on St. Helena, Madame Laetitia languished in the Palazzo Bonaparte, now Palazzo Misciatelli. Across the way is the Palazzo Salviati, built by the Duc de Nevers in the 17th century, owned in the 19th by Louis Bonaparte. The Palazzo Doria is a late-15th-century building behind a 1734 facade. Four mornings a week the family admits the public—through a side door—to its state rooms and its art gallery, in which are many Titians, Bruegels, Caravaggios, a Bronzino, a Memling, and a Velázquez portrait and Bernini bust of the family pope, Innocent X. Behind S. Marcello, the Baroque reworking of a church founded in the 4th century, is the mid-17th-century Palazzo Ballestra, in which Bonnie Prince Charlie of Scotland was born in 1720 and to which he returned in 1788 to die.

The column of Marcus Aurelius, with reliefs showing his victory over Danubian tribes, was preserved from the assorted Christian looters of Rome because it was the property of a religious order. In the square around the column are the Palazzo Chigi (1562), for many years the Ministry of Foreign Affairs and now the offices of the Italian Cabinet, and the Palazzo Wedekind. Although built in the 19th century, the Wedekind, which now houses a daily newspaper, is not without its plundered antique columns.

The Piazza del Popolo

The Corso emerges onto the splendid oval Piazza del Popolo, which is monumental without being intimidating, a sort of toy theatre stage set magically magnified. Over a period of 300 years, it was constructed as the ceremonial entryway to Rome, and, although its elements are diverse in style and in age (13th century BC–19th century AD), a remarkable harmony prevails. In 1561 the Porta del Popolo, the medieval gate in the city wall, was rebuilt. Ninety-four years later its inner face was redone by Bernini for the grand entrance of Queen Christina, who had abandoned the Protestant throne of Sweden for the Catholic hospitality of Rome. In 1589 Pope Sixtus V punctuated the plaza centre with an obelisk (13th century BC) brought by Augustus from Heliopolis to the Circus Maximus.

The church next to the gate, Sta. Maria del Popolo, which stood for centuries before the piazza existed and gives its name to the area, was founded in 1227 to replace a 1099 chapel built over what was presumed to be Nero's tomb. It was replaced in 1472–77 by today's church, further disguised on the piazza frontage by a Neoclassical facade. The interioı is fraught with the works of great Renaissance and Baroque artists. The main chapel has tombs by Andrea Sansovino and frescoes by Pinturicchio. In the Cerasi Chapel are Caravaggio's "Conversion of St. Paul" and his "Crucifixion of St. Peter." The Chigi Chapel, unique for the early 16th century in being a miniature church, was designed by Raphael. Bernini sculpted two of the four prophets in the corners.

At the opposite end of the piazza stand "twin" churches (1662) framing the entrance to three streets. The streets were there first, so the churches were ingeniously squeezed into awkward, different-sized plots between them. Sta. Maria di Montesanto, on the narrower plot toward the Tiber on the west, has an oval plan and dome, while Sta. Maria dei Miracoli, on the east, has a round dome. Carlo Rainaldi, the architect, turned both facades slightly inward to frame the welcoming parades that would proceed up the Corso between the two churches. One of the streets, the Via del Babuino, was one of many built by Sixtus V (1585–90) to try to repopulate parts of Rome deserted after the Gothic wars.

Since lack of water had driven residents off the high ground, he restored the aqueduct of Alexander Severus, the Aqua Alexandrina, and gave it his own first name, Aqua Felice. He laid out new roads, the basis for the modern street plan of Rome. He also built the Vatican Library, saw to the completion of St. Peter's dome, rebuilt the papal palaces of the Vatican, the Quirinal, and S. Giovanni in Laterano (St. John Lateran), refurbishing the squares in front of the last two, and built a new square at Sta. Maria Maggiore. He re-erected four obelisks found among the ruins and restored a great number of fountains, dearly beloved of the Romans.

The Piazza di Spagna

An obelisk in the Piazza di Spagna is not his work but was discovered in the Piazza in Campo Marzio in 1778 and erected in 1857 to commemorate the 1854 promulgation of the dogma of the Immaculate Conception. The fountain is fed by the Aqua Vergine, Agrippa's aqueduct of 19 BC, which escaped Gothic destruction because it was mainly underground and which was repaired in 1447. When the fountain was planned in the early 1600s by Bernini (whether by father or son has not been established), there was insufficient water pressure for spouting jets, so the Barcaccia (Scow) was conceived, an ancient marble boat foundering endearingly in its marble bath.

The most striking architectural element in the piazza—indeed, one of the most striking in all Rome—is the renowned Scala di Spagna (Spanish Steps, or Stairs) which transmutes an irksome climb into a sensuous pleasure. The staircase is a rare case of the failure of French cultural propaganda, for while they are called the Spanish Steps—the Spanish Embassy moved onto the square in the 17th century—they are unequivocally French. First suggested by the French about the time the Spanish Embassy was being installed, the idea was approved by papal authorities 100 years later and paid for with a legacy from a French diplomat. The stairs ascend to the French-built church and convent of Trinità dei Monti, begun in 1495 with a gift from the visiting French king Charles VIII and restored by Louis XVIII.

Charles Dickens described the steps as thronged with unengaged "artist's models" in regional costume. They are still crowded with loiterers in distinctive dress, students from all over the world. Artists were among the first to move into the area, and some few who have not been shouldered out by galleries and ultra-modish shops retain their studios among the walled gardens of the Via Margutta. Since the end of the 16th century, the Piazza di Spagna, with its innkeepers who followed the artists, has been a stopping place for tourists. Young lords on the Grand Tour of Europe left their heavy touring coaches for refitting in a side street still called Via delle Carozze (Carriage Street). The room on the piazza in which John Keats died in 1821 has been made into a museum. The surrounding streets at both the top and the bottom of the steps are among the smartest shopping streets in Rome.

III. Monuments of the Roman past

The historic scourges of Rome

Many of the treasures of Rome no longer can be seen where they were placed originally, many can be seen only in other cities of the world, while many others still in Rome represent the spoils of conquest brought to the city from around the ancient world or the cannibalizing of one age or of one faith upon the creations of an earlier one. Rome was sacked first by the Gauls in 390 BC and subsequently by the Visigoths in AD 410, the Vandals in 445, the Normans in 1084, and Spanish troops in 1527. Muslims laid it under siege in 846. The Great Fire of Rome—Nero's fire—occurred in AD 64, and fires and earthquakes ravaged individual buildings or whole areas fairly often over the millennia. But, of all these scourges, it was the stripping of the structures of antiquity for building materials, especially from the 9th century through the 16th, that destroyed more of Classical Rome than any other force. The heritage of the past that survives in Rome is nevertheless unsurpassed in any city of the West, and it is so ubiquitous that its highlights must be comprehended in terms both of geography and of type.

1HE SEVEN HILLS

The Palatine. The origins of Rome, as of all ancient cities, are wrapped in fable. The Roman fable is of Romulus and Remus, twin sons of Mars, abandoned on the flooding Tiber and deposited by the receding waters at the foot of the Palatine. Suckled by a she-wolf, they were reared by a shepherd and grew up to found Rome, Romulus being obliged to execute Remus for disobeying one of the city's first laws. The Etruscan bronze statue of the maternally ferocious wolf (late 6th or early 5th century BC; Capitoline Museum) is one of the greatest works among the thousands of masterpieces in Rome. The nursing infants were sculpted and placed under the Etruscan statue in 1509.

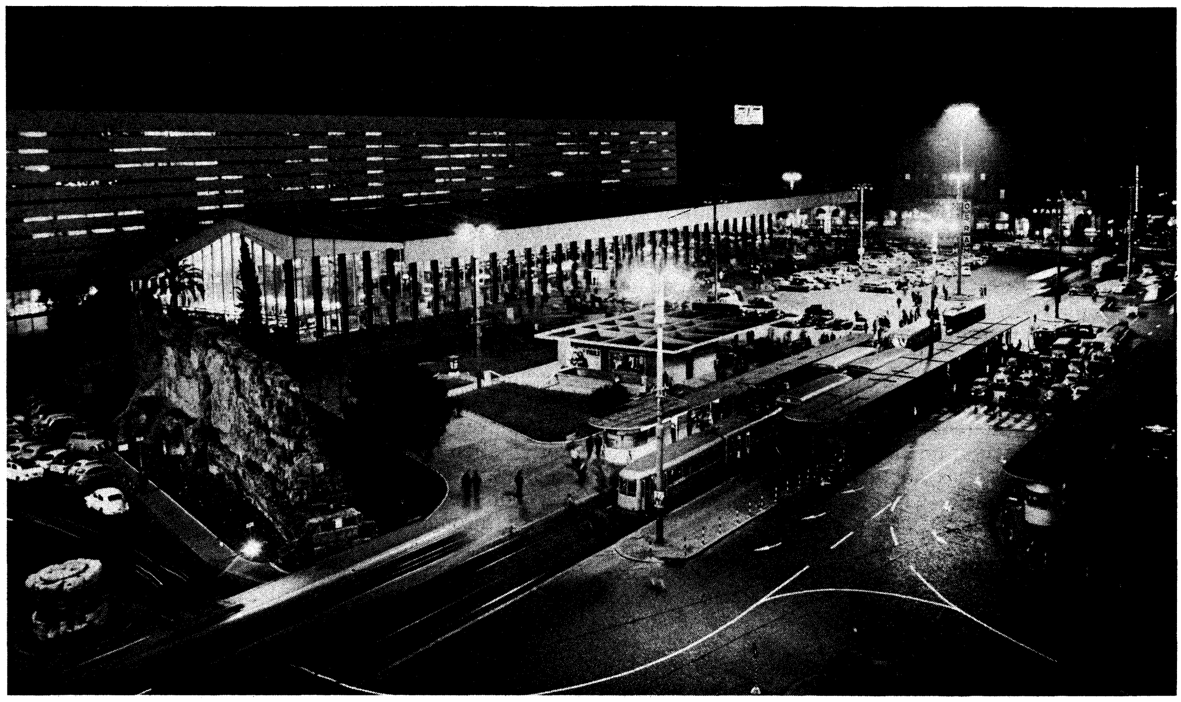
The wolf cave, the Lupercal, was maintained as a shrine at least until the fall of empire but is now lost. On the same side of the Palatine, "Romulus' House," a timber-framed circular hut covered in clay-plastered wickerwork, was kept in constant repair. Modern excavations have revealed the emplacement of just such Iron Age huts from the period (8th–7th century BC) given in the fable for the founding of Rome.

On this hill the columns of lost palaces rise in uncompromised beauty from fields of wildflowers and the dust of history. Ilex and pine and bay frame views of Rome. This is the landscape—classical, with figures—that has stirred romantics since it was first limned by 17th-century etchers and sketchers. Before the emperors departed, virtually the entire hill was one vast palace.

Residences of the 1st century

The Palatine was a superior residential district by the 3rd century BC. Augustus was born there in 63 BC and continued to live there after he became emperor. His private dwelling, built about 50 BC and never seriously modified, still stands. Known as the House of Livia, for his widow, it has small, graceful rooms decorated with paintings of fruit and flowers and mythological scenes. Other private houses, now excavated and visible, were incorporated into the foundations of the spreading imperial structures, which eventually projected down into the Forum on one side and onto the Circus Maximus on the other. The three crests of the hill were flattened in the course of building. The palace was begun by Tiberius, to whose work Nero, Caligula, Trajan, Hadrian, and Septimius Severus made their own additions.

The biggest and richest structure of all was created for Domitian (reigned AD 81–96), whose architect achieved feats of construction engineering not seen before in Rome. Parts of the lavish structure—the richly marbled, centrally heated dining hall of which is among the chambers visible today—were occupied by popes after there were no more emperors, and then the hill was abandoned.



The Stazione Termini in the Piazza dei Cinquecento and the remains of the 4th-century BC Servian Wall (left) that abuts the building.

Garofalo—Grirnoldi

After some six centuries the great Roman families returned to the Palatine, planting 16th-century pleasure gardens and pavilions over past glories. A whole set of rooms from the private wing of Domitian's palace was preserved by incorporation into the Villa Mattei. Atop Tiberius' palace the Farnese family built two aviaries and a garden house and laid out one of Europe's first botanical gardens—some parts of which have escaped archaeological excavation.

The Capitoline. The seat of Roman government, the Capitoline is little changed from Michelangelo's design and represents one of the earliest examples of modern town planning. The centrepiece of this piazza of three palaces is a bronze equestrian statue of Marcus Aurelius, which stood unmolested for ages by the barracks of the imperial guard (later the Lateran Palace) because it was believed to be a statue of Constantine, the first Christian emperor.

The Palazzo Senatorio incorporates remains of the facade of the Tabularium, a state-records office constructed in 78 BC and one of the first buildings to use concrete vaulting and employ the arch with the Classical architectural orders. After a popular uprising in 1143, a palace was built on the site for the revived 56-member Senate, supposedly elected by the people but by 1358 a body of one appointed by the pope; when it was rebuilt to Michelangelo's design, it was called the Palazzo Senatorio (Senate Palace).

The palace of the municipal councillors, the *conservatori*, is on the south side of the square opposite the Palazzo del Museo Capitolino (Capitoline Palace), which, as a papal collection of Classical works offered back to the citizens of Rome by Sixtus IV in 1471, became the first public museum of sculpture in the Western world. Now occupying both the Capitoline Palace and the Palazzo dei Conservatori, as well as a later private palace, the museum contains only objects found in Rome, including the famed Romulus and Remus wolf, the "Capitoline Venus," the "Dying Gaul" and the "Boy with Thorn," as well as the host of portrait busts that can, in imagination, repeople the Forum just below.

The hill was the fortress and asylum of Romulus' Rome. The northern peak was the site of the Temple of Juno Moneta (the word money derives from the temple's function as the early mint) and the citadel emplacements now occupied by the Victor Emmanuel monument and the Church of Sta. Maria d'Aracoeli. The southern crest, sacred to Jupiter, became, in 509 BC, the site of the Temple of Jupiter Optimus Maximus, the largest temple in central Italy. The tufa platform on which it was built, now exposed behind and beneath the Palazzo dei Conservatori, measured 203 by 174 feet (62 by 53 metres), probably with three rows of six columns across each facade and six columns and a pilaster on either flank. The first temple, of stuccoed volcanic stone quarried at the foot of the hill, had a timber roof faced with brightly painted terra-cottas. Three times it burned and was rebuilt, always of richer materials. The temple that Domitian built was marble with gilded roof tiles and gold-plated doors. It was filled with loot by victorious generals who came robed in purple to lay their laurel crowns before Jupiter after riding in triumph through the Forum. The Clivus Capitolinus, the antique pavings of which can be walked today, was lined with 40 elephants bearing torches to light the way for Caesar coming in triumph from Gaul. In this centre of divine guidance, the Roman Senate held its first meeting every year. When Petrarch was crowned with laurel among the ruins of the capitol in 1341, it was a harbinger of the Renaissance.

The church of Sta. Maria d'Aracoeli, built before the 6th century, remade in its present form in the 13th, is lined with columns rifled from Classical buildings. It is the home of "I1 Bambino," a much loved miracle-performing wooden Christ child who is called to save desperately ill children. At Christmas, adorned in jewels given by the grateful, he can be seen in the church's celebrated manger scene, where he is serenaded by shepherd pipers.

The Aventine. Though considerably built over with modem houses and travelled by modern bus lines, the Aventine still bespeaks a Rome of the past, if not the Classical past. The repeated fires that swept the city destroyed all the republican buildings, and the Temple of Diana remains only as a street name. Under the 4th-century Church of Sta. Prisca is one of the best preserved and maintained Mithraic basilicas in the city. The Basilica of Sta. Sabina, little altered since the 5th century, is lined with 24 magnificent matching Corinthian columns rescued out of Christian charity from an abandoned pagan temple or palace. The Parco Savello, a small public park luminous with orange trees, was the walled area of the Savello family fortress, one of 12 that ringed the city in medieval times.

The Knights of Malta

A romantic gem is the Piazza dei Cavalieri di Malta, designed in the late 1700s by Giambattista Piranesi, an engraver with the heart of a poet and the eye of an engineer. To the right of this obelisked and trophied square, set about with cypresses, is the Knight's Priory, residence of the grand master of the Knights of Malta. It is this order, rather than Vatican City, which is the smallest sovereign state in the world, whose extraterritorial rights, recognized by the Italian government, cover this building as well as the seat of the organization. In 1113 the newly founded order, the Knights Hospitaller of St. John of Jerusalem, were in the Holy Land, whence they were driven to Rhodes, which they held until 1522, thence to Malta until 1789, when they repaired to their stronghold in a Roman side street. The sovereign military order has its own flag, issues its own license plates and (rarely) passports, and continues its long history of international medical work.

The Caelian. Almost half parkland, the Caelian includes the public park of Villa Celimontana, once the garden of the Mattei family, who had another on the Palatine, a clutch of palaces in the Campus Martius, and another in Trastevere. The six churches on the hill date from the 4th to the 9th century.

In the medieval confines of the only fortified abbey left in Rome stands SS. Quattro Coronati, today sheltering nuns and their charges, deaf-mute children. The Basilica of SS. Giovanni e Paolo, from the 5th century, stands in a piazza that has few buildings later than the Middle Ages. Alongside the church are the remains of the platform of the Temple of Claudius, partly dismantled by Nero, completely by Vespasian. The round Church of S. Stefano Rotondo (460–483) may have been modelled on the Church of the Holy Sepulchre in Jerusalem.

The Hospital of St. John was founded in the Middle Ages as a dependence of S. Giovanni in Laterano (St. John Lateran), just off the hill, and maintains its Romanesque gateway. The Hospital of St. Thomas, established at the same period, has disappeared save for its mosaic gateway, signed by the original Cosmate of the Cosmati school of carvers and decorators and by his father Jacobus. Nearby stands the Arch of Dolabella (AD 10), and not far away are the ruins of Nero's extension of the Claudian aqueduct. Also on the hill is the extensive Military Hospital of Celia.

Nero's palace

The **Esquiline.** Between the Esquiline and the Caelian, the end of the Forum valley is filled by the Colosseum and the Arch of Constantine, with the Palatine edging down from the north. After the fire of AD 64 had destroyed so much of the city, Nero undertook to rebuild the end of it —200 acres (81 hectares)—as a palace for himself: seawater and sulfur water were piped into its baths; flowers were sprinkled down through its fretted ivory ceilings; and the facade was covered in gold, from which the name Domus Aurea, the Golden House. The expropriation so enraged the citizens that his successors hastened to efface all trace of Nero's incredible palace: the ornamental artificial lake was drained and on its bed the Colosseum was erected for free entertainment; Trajan built magnificent baths—also with free admission—atop the domestic wing of the Golden House; and Domitian converted the portico on the edge of the Forum into Rome's smartest shopping street. The obliterators were aided by the fire of AD 104. In 131 Hadrian erected his Temple of Venus and Rome where the vestibule had stood at this end of the Forum; the church and former convent buildings of Sta. Maria Nova were built on the western comer of the temple platform in the 10th century. Less than 70 years after the Golden House had been started, nothing was left of it but a 150-foot (50-metre) gilded statue of Nero. Popular tradition has it that the face was changed with each succeeding emperor, but it was destroyed by one of the early popes.

The removal was so complete that later Romans could not remember where the Golden House had stood. When the domestic wing was discovered under Trajan's Baths in the 15th century, the rooms painted in the Pompeiian style were thought to be decorated grottoes. Some years later, when Raphael and his friends were let down on ropes to look, the style they imitated in decorating the Vatican loggias was called ***grottesche.***

The Colosseum

The Colosseum that replaced Nero's lake is more correctly called the Flavian Amphitheatre. It was begun by Vespasian and inaugurated by Titus in AD 80. The oval stadium measures one-third of a mile around, with external dimensions of 615 by 415 feet. The 160-foot (49-metre) facade has three superimposed series of 80 arches and an attic story. The attached columns follow the order applied on the Theatre of Marcellus (13 BC): sturdy, unadorned Doric on the ground floor, more elegant Ionic next, and luxuriant Corinthian on top. The attic story bore corbels supporting masts from which royal sailors manipulated awnings to protect the 50,000 seats from the sun during the gladitorial contests, combats with wild animals, sham battles, and, when the arena was flooded, naval displays. The main structural framework and facade are travertine, the secondary walls of volcanic tufa, the inner bowl and the arcade vaults of concrete. Until Pius VIII (reigned 1829–30) began conserving what was left, it had been a convenient quarry for 1,000 years.

The nearby Arch of Constantine was erected hastily in 315 to celebrate a victory two years earlier. Almost all the sculpture on this splendid arch was snatched from earlier monuments: a battle frieze from the Forum of Trajan, a series of Hadrianic roundels, and eight panels from a Marcus Aurelius monument.

Not all the rooms of the Golden House on the Oppio have been excavated. Above them spread the remains of Trajan's Baths, theatrical decorations for the public garden, Parco di Traiano. They served as models for the baths of Caracalla (c. 212–217) and Diocletian (298–305/306), which, in turn, served as a pattern for the Basilica of Maxentius. The bath building that housed the hot, warm, cold, and exercise rooms and the swimming pool was a huge, rectangular concrete structure lined with marble. It was surrounded by a garden enclosed in an outer rectangle of libraries, lecture halls, art galleries, and other facilities of a big community centre.

Caracalla's baths on the river flats behind the Caelian Hill covered more than six acres, part of which area is occupied today by the modern glass-fronted buildings of the United Nations Food and Agriculture Organization and the Ministrv of Posts and Telecommunications. Among the towering remains set in a large park, the caldarium (steamroom) is used today for summer opera performances. Much of the famed Farnese collection of marbles was stripped from these baths.

The Baths of Diocletian are over the brow of the Viminal, and some idea of their size (130,000 square yards, or 110,000 square metres, for the main bath block) can be gained from the fact that the Church of S. Bernardo was built into one of the chambers some 500 feet west of the central hall of the 92-foot- (28-metre-) high *frigidarium* ("cold room"), into which Michelangelo built the cloister church of Sta. Maria degli Angeli in 1561.

The first seven halls of the Museo Nazionale Romano, also called the Museo delle Terme, are rooms of the *frigidarium* block. This matchless collection of antiquities includes wall paintings from villas, mosaics, sarcophagi, and sculptures, including the famous Ludovisi throne (Greek, 5th century BC), the Niobid from the Gardens of Sallust where the Via Veneto wends today, and the bronze "Pugilist" (2nd century BC), discovered in 1884 in a building site on the Quirinal.

The Basilica of Maxentius (also named after Constantine, who completed it after dispatching Maxentius) was started about 311. This massive hall of justice and commerce was an oblong 265 feet (81 metres) long and 120 feet (37 metres) high, covered by three groin vaults with three deeply coffered tunnel-vaulted bays on either side. Probably ruined by the earthquake of 847, it was also mined for its materials. One of the great Corinthian columns stands obelisk-like before Sta. Maria Maggiore on the Esquiline. The head of Constantine's 40-foot- (12-metre-) high statue reposes in the courtyard of the Palazzo dei Conservatori.

The Viminal and Quirinal. Like much of the Esquiline, the Virninal and Quirinal lie in the heart of modem

Rome. Heavily built upon and sclerotic with traffic, the former seems almost flattened under the Ministry of the Interior, the weighty department that directs the state's police forces. The Quirinal, pierced by a modern traffic tunnel, has been a distinguished address since Pomponius Atticus, recipient of Cicero's letters, was a resident. Starting with the Crescentii, who planted the family fortress there in the Middle Ages, powerful Roman families built their homes in this location. The Colonnas still live in theirs at the foot of the hill near the Corso, an art gallery open to the public, and their gardens, climbing the slope to the Piazza Quirinale, contain remnants of Caracalla's Temple of Sarapis. The piazza has been graced since antiquity with two large statues of men with rearing horses, "The Horsetamers" or "Castor and Pollux." Closed on three sides by palaces, the piazza opens on the fourth to a splendid view over the Tiber. The Quirinal Palace, built by Pope Gregory VIII in 1574 as a summer palace away from the heat and malaria of the Vatican, was enlarged and embellished over the next 200 years by a succession of noted architects. The palace, with many extensions and wings, is huge, and its garden is five times as big as the building. From 1550 to 1870, the Quirinal rather than the Vatican was the official papal residence. In 1870 it became the royal palace of the new Kingdom of Italy; and, in 1948, the presidential palace. Both monarchs and presidents, however, have preferred to inhabit the homier palazetto at the far end.

The hill of family strongholds

The handsome buildings opposite are the stables (1730–40), built on the site of the Crescenti 10th-century stronghold. The Palazzo della Consulta (1734) was erected for part of the papal administration. The Palazzo Pallavicini-Rospigliosi, built by a Borghese cardinal in 1603, is still a private house. The Palazzo Barberini farther up the hill, constructed 1629–33 on the site of the old Palazzo Sforza, was occupied by the family until 1949. Part of the collection of the Galleria Nazionale d'Arte Antica is housed here, the rest across the river in the Palazzo Corsini in Trastevere. The 1,700 pictures, most of them works by celebrated masters, were contributed by distinguished families, including the Barberinis. Architecturally, the palace is important, because it marks a departure from the heavy-set four-square town houses of the early and High Renaissance. In the Rome region, only country villas had previously been built on so open a plan, with two wings coming forward from an open, arcaded facade. Further, it pioneered the Baroque style in domestic architecture.

Carlo Maderno, who put the facade on St. Peter's, made the plans, which were carried out after his death by Bernini, assisted by Borromini. Each of these two rivals has a church just around the corner. After 20 years of apprenticeship, Borromini was given his first chance to do his own building. It was an impossibly tiny site at the crossroads of the Quattro Fontane (Four Fountains, one of which is built into a niche in the church wall), but S. Carlo alle Quattro Fontane was a triumph. To his revolutionary solutions of site problems, for which he employed a brilliant variation on the oval, Borromini added a facade in 1667, the year he died, which responded to the waves of motion generated by the spatially complex interior. Walls that flow and sway created a sensation, and the idea was seized upon by Baroque artists, especially from other nations.

Bernini's S. Andrea al Quirinale is also small, but it took 12 years to build (1658–70), late in his career. An oval building with the naves sculpted into the outer wall, it enlarges on concepts advanced by Michelangelo. Bernini's use of coloured marbles and shrewd lighting effects gives the small structure extra dimension.

Other hills. Behind the river plain of Trastevere is the Gianicolo (Janiculum), and behind the Piazza del Popolo across the river is the Pincio. Both are now parkland, with villas, gardens, and churches discreetly disposed. The Janiculum crest was made into a park in 1870 to honour Garibaldi for his heroic but unsuccessful defense of the Roman Republic in 1849. During the Roman Empire the Pincio was covered with villas and gardens, but it was made into a public park only in the 19th century. By day, nannies wheel their charges through the greenery, and toward sunset Romans arrive to carry on the turn-of-the-century tradition of the before-dining Pincio promenade. Down the road toward Trinità dei Monti is the 1554 Villa Medici, bought by Napoleon in 1803 to house the French Academy of Rome, which is still in occupation. This academy, founded in 1666, is the oldest of many national academies established from the 17th to the 19th century to give architects, artists, writers, and musicians the opportunity to study the vast textbook that is the city itself and to use its museums and libraries.

The French Academy of Rome

The Villa Giulia and the Villa Borghese are also on the hill, both housing art collections of world importance. The Villa Giulia was a typical mid-16th-century Roman suburban villa, conceived not as a dwelling but as a place for repose and entertainment during the afternoon and early evening. It has a collection of Etruscan art and artifacts of singular beauty and historical value. The Borghese collection is small but choice, with a roomful of Caravaggios and, in addition, Titian's "Sacred and Profane Love." Canova's Neoclassical nude statue of Pauline Bonaparte, for a time a Borghese princess, as Venus retains its capacity to scandalize. The Italian government bought the grounds, house, and contents in 1902.

THE FORUM

The Forum was the religious, civic, and commercial centre of pastoral, royal, and republican Rome. After Julius Caesar, though it became more imposing, it was only one (albeit the most distinguished) of several complexes serving the same functions. Essentially, it was a small, closed valley ringed by the Seven Hills. There were two meeting places, formal open spaces in the northwest corner, the political Comitium and the social Forum—the name later applied to the entire valley—with shops down both sides. At the other end of the valley was the precinct of the high priest next to the Vestals, the keepers of the sacred flame. Between these two were the temples of the gods. Various emperors opened up the ends of the valley, and there was more building; but the poles of activity did not alter.

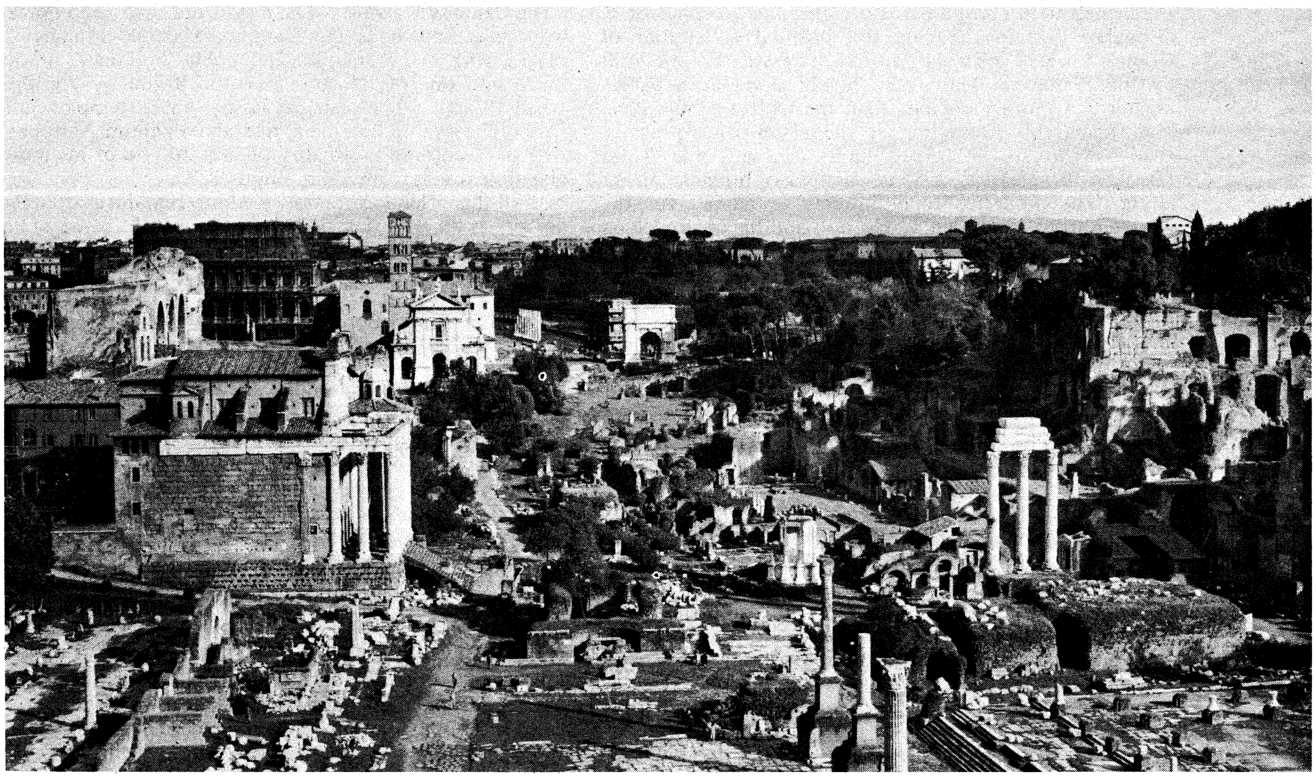
Fires, earthquakes, and invasions repeatedly levelled the buildings, and new ones were erected on their remains until the valley was covered by 50 feet of debris, earth, and ashes. Medieval Romans called it Campo Vaccino (Cow Field) and the abutting Capitoline Hill, Monte Caprino (Goat Hill). Excavation began late in the 19th century, and most of the accumulation has been dug away, down to the level at which Julius Caesar knew it.

The little stream cutting diagonally across the valley floor was, according to tradition, canalized as the Cloaca Maxima in the 6th century. Stratigraphic excavations again support the folktales and date the sewer construction at about 575 BC. Although later buildings perpetuated the name and roughly the position of the first halls and temples, they do not necessarily stand where earlier buildings stood, and many details of the earlier Forum are still the subject of scholarly speculation.

Janus and Saturn, both of whom have temples here, were among the gods of early Rome, and the Temple of Vesta, even in its last marble version (AD 191), retained the circular shape of a primitive clay-and-wattle hut. The forge of Vulcan, the Volcanal, had very early beginnings. The Regia, traditionally described as the residence of Numa Pompilius, the priest-king, became the administrative building for the pontifex *maximus*, who took on the monarchy's priestly duties. The Temple of Castor and Pollux was built at the establishment of the republic.

Building and rebuilding of the Forum

The oldest formally consecrated monument was the open space of the social Forum. A roughly trapezoidal stretch of ground about 125 by 70 feet (38 by 21 metres), it was bare save for three plants essential to Mediterranean agriculture: the grape, the fig, and the olive. Centuries later, when the basilicas were built behind the bordering shops, they served as a protective palisade for the Forum and a covered extension of its open space. At the wide end of the Forum and to one side was the Comitium, in which the Popular Assembly met. Between the



Ruins of the Roman Forum with the Colosseum (left background).
Authenticated News International

two clearings lay the orators' platform, the Rostra, decorated in 338 BC with the iron rams *(rostra)* taken as trophies from the warships of Antium.

At the other end of the Comitium stood the Curia, where the Senate met. When it was destroyed by fire, along with the Basilica Porcia (184 BC, the first of the basilicas), Julius Caesar built a new and greatly enlarged one that encroached on the open space of the Comitium. For the assembly, he built a meeting hall in the Campus Martius, outside the valley altogether. He built a new and much bigger Rostra, though across the wide end of the Forum. He supplanted the Basilica Sempronia (170 BC) on the western side of the Forum with his own Basilica Julia (54 BC), installing new shops in place of the old Tabernae Veteres. On the other side of the Forum already stood the shop-fronted Basilica Aemilia (179 BC), named for the censor who constructed the Tiber bridge now called the Ponte Rotto.

Caesar also carried his building program onto the flat ground just north of the valley between the Quirinal and Esquiline hills, making his own forum of shops and temple, alongside which Augustus, Trajan, Nerva, and Vespasian later constructed their forums. Pompey's theatre in the bed of the Tiber (55 BC) was followed by the Theatre of Marcellus (13 BC). The great baths, Agrippa's grand concourse in the Campus Martius, the circuses, and the Colosseum all drew the populace away to other centres of activity. The political attraction of the Forum, already vitiated in Caesar's day, continued to decline.

Nevertheless, the halls and temples of the Forum were assiduously rebuilt, ever grander, and more were added. Caesar, after his death, was made a god, and his temple was erected between the Forum proper and the Regia. Eventually, the sacred open space was defiled with honorary columns and an equestrian statue of Domitian. The last thing to be erected in the Forum was a column, raised by Phocas, a Byzantine usurper (608) to honour himself. Septimius Severus placed his arch over the Via Sacra. Other temples were rammed into empty places, and the whole became a forest of towering columns, gleaming walls, and ornate statuary. The dazzling marble mountain of the Palatine flowed down into the Forum as well, and the opposite rim glittered with the splendours of the imperial forums.

Today, the Forum is a confusing boneyard of history. Of the thousands of columns, not many more than 50 stand erect. Amid the ruins are Christian churches, thickets of trees and bushes, and hundreds upon hundreds of free-living cats.

THE RIVERLANDS

Along a 1%-mile stretch of the Tiber, around a big kangaroo-nosed bend, lie all the historic quarters of the river plain. On the left (east) bank are the Campus Martius, Circus Flaminius, Forum Boarium, and Forum Holitorium; on the right, the Castel Sant'Angelo, or Hadrian's Tomb, the entrance to Vatican City, and Trastevere. At the bottom of the bend is Tiber Island.

The palace-citadel-tomb on the Tiber

Castel Sant'Angelo and the bridges. Four of the 11 bridges along this part of the Tiber are of special interest. The Ponte Sant'Angelo, to which Bernini was asked to add angels, is in the main the Pons Aelius built in AD 134. A year later Hadrian began his tomb, just off the end of the bridge. A towering cyclinder 20 metres high on a square base, it was in size and form a typical imperial mausoleum. In 271 it was built into the Aurelian Wall and became a key fortress in the defense of Rome. In 587 Gregory the Great, leading a procession to pray for the end to a plague, allegedly had a vision of the archangel Michael atop the tomb. The epidemic ceased and the tomb-citadel became known as the Castel Sant'Angelo (Castle of the Holy Angel). In time it became a papal castle, with richly furnished and frescoed rooms, loggias for the view, a siege store of 5,800 gallons (22,000 litres) of oil and 770,000 pounds (3,500 quintals) of grain, a centrally heated bathroom, a prison that incarcerated Benvenuto Cellini, among others, and a still intact fortified passage from the Vatican to carry the pope to refuge there. It is now a state museum with an arboured terrace for outdoor drinking and dining.

At Tiber Island are two bridges. The Ponte Cestio, often rebuilt since the 1st century BC, leads to Trastevere, while the Ponte Fabricio (62 BC), the oldest in Rome, runs from the shore below the Capitoline. The island, 1,100 feet (350 metres) long and less than 330 feet (100 metres) wide at its widest, has been a place of healing since the Temple of Aesculapius was erected after the plague of 291 BC; the largest building there today is the

Fatebenefratelli hospital (also called the Hospital of S. Giovanni di Dio). Facing the hospital is another of Rome's towered medieval family fortresses, this one built by the Pierleone. The traffic howls along both banks, noisier and more voracious than the wolves of the Pierleone's anarchic Rome, but on the island peace prevails. Just downstream are the remains of the Ponte Rotto (Broken Bridge) of 179 BC and two bridges farther along. The modern Ponte Sublicio is named for the wooden bridge defended by Horatio and his comrades on this part of the river.

The lower east bank. On the shore by the Ponte Rotto is the site of the earliest cattle market (Forum Boarium) and vegetable market (Forum Holitorium), girt with temples, of which two remain: the elegant, circular Pentellic marble structure of the 1st century BC and a nicely proportioned, rectangular Ionic building, perhaps a few decades older. Their dedications are disputed, save that they are not, as they are popularly called, temples of Vesta and of Fortuna Virilis. In the 6th century the Church of Sta. Maria in Cosmedin was built into the antique grain-commission offices. Some of the Forum Boarium columns can still be seen on the interior of the church, and one of its drain lids, fixed to the outer wall, was carved to represent a face with a gaping mouth. This classical manhole cover became the dread Bocca della Verità, the mouth of truth, which allegedly would crunch down upon the hand of anyone telling a lie.

Nearby is the Theatre of Marcellus, begun by Caesar and completed in AD 13 by Augustus, who named it for a short-lived nephew. It owes its preservation to its conversion into a fortress for one of the quarrelsome clans of the Middle Ages. Converted into a palace for the Orsinis in the 16th century, it remains private property. The classical orders of the facade, adopted for the Collosseum, became the model for Renaissance architects.

From there northward to the tomb of Augustus and as far inland as the Via Flaminia (today's Corso), the river plain was a vast plantation of temples, baths, and sports grounds until the Middle Ages, when the remaining Romans took up residence there. Today, three major imperial monuments survive: the Pantheon, the reconstructed Ara Pacis Augustae (Altar of Augustan Peace), and Hadrian's Column. Here and there among the 40 palaces and 100 churches are remnants of what the emperors built.

The Jewish ghetto

The portion closest to the Tiber Island was once a major republican racing and sports ground, the Circus Flaminius (220 BC), which in the 16th century became the Jewish ghetto. Jews were not persecuted in Rome until Pope Paul IV (1555–59) herded them into a ghetto under curfew. Although Paul was so loathed that the Romans decapitated his statue when he died, other popes carried on his anti-Jewish program. Except for brief respites under Napoleon and the momentary Roman Republic of 1848, Jews until 1870 were debarred from all the professions, government service, and landownership. Until recently the neighbourhood retained a Jewish flavour, with some 3,000 Jews living there in the 1960s, but by 1971 only 300 remained, as the ghetto, like Trastevere, became ripe for conversion to luxurious flats. Nearby, the Largo Argentina, excavated 1926–29, contains four small temples of the 1st and 2nd centuries BC.

The crescent of buildings between the Piazza del Biscione, an open-air junk market, and the Piazza dei Satiri take their curved shape from having been built into and around Pompey's Theatre, the first stone theatre building in Rome. Inspired by the Greek theatre of Mitylene, in which Pompey had been so spectacularly entertained, it had a portico of 100 columns that was equipped to be a community centre almost as much as the baths. The Senate met there on the Ides of March in 44 BC, when Julius Caesar was stabbed 23 times and fell at the foot of Pompey's statue. For almost 400 years a piece of sculpture, unearthed nearby in 1550 and deposited in the Palazzo Spada, was erroneously believed to be the Pompey statue. A part of the threatre was fortified by the Orsinis in the 12th century and later converted into the Palazzo Righetti, or Pio.

The Campus Martius. The rest of the river bend northward was known as the Campus Martius. Marshy in places, with a few temples and public buildings, it was made into one of the grandeurs of Rome by Agrippa (died 12 BC), a landscape of lawns, baths, temples, and parks The swamp became a lake, the Stagnum Agrippae, where—according to Tacitus—Nero led one of his more elaborate orgies from a sumptuous raft.

Of all this splendour almost nothing remained after the fire of AD 80. Hadrian undertook to restore some of it. Among his works was the new Pantheon, one of the West's great buildings, extraordinary as architecture, remarkable as a feat of engineering. This "Temple of All the Gods," imperial property, survived because it became a church, the gift of the Byzantine emperor Phocas to Pope Boniface IV in 608. This protected the building from everyone but the pope: the bronze roof beams of the grandiose pedimental porch of 18 sixty-ton columns of Egyptian granite were stripped by Urban VIII, the Barberini pope, who took them as raw material for the baldachin in St. Peter's, provoking the celebrated anonymous comment, Quod non fecerunt barberi, fecerunt *Barberini*, "What was not done by the barbarians was done by the Barberinis."

The Pantheon

It has been suggested that the temple was designed by Hadrian himself, whose villa at Tivoli is another landmark in the development of architecture. The Pantheon was possibly the first monumental building of antiquity conceived as an interior. Evenly lighted from a single source—the open eye (*oculus*) in the centre of the dome—the enormous interior, circular and richly marbled, is almost unchanged from classical times. Until the 20th century the dome was the largest ever built, 141 feet in diameter, exactly the height of the building. Two things made its construction feasible: the magnificent quality of the mortar used in the concrete and the meticulous selection and grading of the aggregate, which became lighter in weight with increasing height. Roman concrete was essentially a hydraulic cement, deriving its unique strength from the properties of the dark volcanic ash (pozzolana) of the Roman subsoil that was substituted for sand. There is some brick ribbing in the lowest part of the dome and thrust-containing brick outer facing, but, in general, brick was not used by the Romans as a building material in itself. Brick and tile were used to help hold the concrete until it dried, making for a less brutal exterior. The stamped trademarks on the bricks from the big yards behind Vatican Hill and up the Tiber Valley help in determining chronology. The Pantheon, for example, bears the original dedicatory inscription of Agrippa, modestly replaced by Hadrian. The latter's name does not appear, but the stampings on the bricks show that construction does indeed date from Hadrian's reign. The original bronze doors are still in place. Italy's first two kings are buried in the Pantheon, as are many artists, of whom Raphael is the most notable. Nearby are fragments of Agrippa's baths, and the Rome stock exchange gains considerable dignity from the incorporation of some of the Temple of Hadrian.

The shattered drum of Augustus' tomb marks the spot where he was buried AD 14. The mausoleum became a 12th-century Colonna fortress, a 16th-century garden, a ring for Spanish bullfights in the 17th century, and then a concert hall until 1936, when it was scraped down to its impressive but mournful foundations by Mussolini, who may have planned to be buried there himself. Next to the tomb is the delicately beautiful white marble Ara Pacis Augustae (Altar of Augustan Peace, designed 13 BC, dedicated 9 BC). The altar, raised on steps, is enclosed in a sculptured screen. Bits of the friezes were discovered off the Corso in the 15th century, and the altar itself was dug up there in 1938 after 35 years of labour. The pieces unearthed earlier were bought back from museums, and the whole was reassembled to stand four streets away from its original location.

In the Campus Martius Italy's Chamber of Deputies sits in the Bemini-designed Palazzo di Montecitorio, its Senate in Palazzo Madama (17th century), and its Council of State in Palazzo Spada (c. 1540), the picture gallery of which is open to the public. The Museo di Roma, which illustrates the life of the city through the ages, is in Palazzo Braschi (18th century). The Brazilian embassy is in the Palazzo Pamphili, which has a gallery designed by Borromini and painted by Pietro da Cortona. The early-16th-century Palazzo di Firenze was the Florentine embassy until the union of Italy; it is now occupied by the Società Dante Alighieri. The Palazzo Capranica is a cinema, while the Palazzo della Sapienza, near the Senate, is now the National Archives but was from 1431 to 1935 the seat of the University of Rome (founded 1303).

Architecture of the palaces

The palaces. The three architecturally celebrated palaces in this palace-truffled quarter are the Cancelleria, the Farnese, and the Massimo alle Colonne. Because all the pertinent documents were destroyed in the Spanish sack of Rome in 1527, the architect of the Cancelleria remains unknown. Dated 1486–98, it was built by Cardinal Raffaelo Riario out of a night's winnings at the gaming table. Seized by the Medici Pope Leo X (1513–21), it has housed some portion of the Vatican chancellery ever since, except for Napoleonic and revolutionary interruptions. A square building with a rusticated ground floor, its upper stories are plain and rhythmically pilastered, while the columned inner court is noble and deeply harmonious. The city's first High Renaissance building, it could be said to symbolize Rome's displacement of Florence as art capital of the world—its artists drawn from north and south but not from Rome.

The Farnese, the most monumental of Rome's Renaissance palaces, was designed by Antonio da Sangallo the Younger, who was succeeded after his death by Michelangelo, Vignola, and Giacomo della Porta. Sangallo followed the Renaissance precepts regarding the architectural orders on the lower floors, but Michelangelo's top story uses the traditional elements in a willful way, capping it all with an overpowering cornice—a personal expression that foreshadowed Mannerism, a leaching of Renaissance ideals, and the subsequent theatrical self-expression of Baroque. Michelangelo's project to join this palace to the Farnesina by a bridge over the river was actually begun. It can be seen from the surviving arch over the Via Giulia, one of the city's most charming streets.

Mannerist architecture is typified by Baldassare Perruzi's Palazzo Massimo alle Colonne (1535), the name of which comes from a colonnaded palace on the site destroyed in the 1527 sack. It disregards all Renaissance canons, with its brooding entry and heavy cornice below a slightly bowed and airy facade punched with small windows. The Massimo family, who still live in the palace, gave shelter to Konrad Sweynheym and Arnold Pannartz, who produced Rome's first printed book in their house in 1467.

THE CHURCHES

Today, 25 of the original parish churches, or tituli, the first legal churches in Rome, still function. Most had been private houses in which the Christians illegally congregated, and some of these houses, as at SS. Giovanni e Paolo, are still preserved underneath the present church buildings. Since the 4th century the tituli priests have been cardinals who, over the centuries, have rebuilt, enlarged, and embellished their churches.

In the 4th century, basilicas were built to mark the burial places of martyrs. Most had been interred beyond the city walls in underground galleries, the catacombs. When later sieges of Rome laid waste the Campagna, saintly relics were removed to the safety of city churches. During the Middle Ages, when the prevalence of malaria and of tomb robbers—there was a brisk commerce in religious relics—made ventures beyond the walls risky, some of the oratories and basilicas fell almost to ruin and the location of some catacombs was forgotten.

The "great" basilicas. Among the basilicas, seven are designated as great (maggiore) : St. Peter's, S. Paolo Fuori le Mura (St. Paul's Outside the Walls), and S. Giovanni in Laterano, all built by Constantine, and those of S. Lorenzo Fuori le Mura (St. Lawrence Outside the Walls), Sta. Croce in Gerusalemme (Holy Cross in Jerusalem), S. Pietro in Vincoli (St. Peter in Chains), and Sta. Maria Maggiore. Under the 1929 concordat with Vatican City, the Italian government grants them extraterritorial privileges.

The basilicas established the model for Western ecclesiastical architecture for centuries to come. Basilica, a Greek word meaning royal, was used by the pre-Christian Romans to designate a public hall, but no surviving example of Roman basilica anywhere in the empire is the architectural predecessor of the Christian basilica.

Basic design of the basilicas

The basilical church has a nave higher than the aisles, from which it is separated by a colonnade on each side. It has either a cloistered court (atrium) or anteroom (narthex) or both at the west end and a semicircular projection (apse) at the east. The basilicas in Rome that are closest to the early Christian structures are the churches of competing cults, as strikingly exemplified by the Neo-Pythagorean-sect basilica of the Porta Maggiore, unearthed by the railroad viaduct in 1926.

Some early Christian churches were centrally rather than longitudinally organized, a plan dictated by the circular form of the imperial mausoleums into which they were built. A good example is Sta. Costanza (c. AD 320), which also has a superb series of 4th-century vault mosaics in pagan designs. Although churches of this type were few, they had a strong influence on the development of the centrally planned house of worship.

St. Peter's. Protected by the fortified Castel Sant'Angelo, St. Peter's Basilica and the Vatican Palace gained precedence over the cathedral church and Lateran Palace during the papacy's troubled centuries. St. Peter's was built over the traditional burial place of the Apostle from whom all popes claim succession. The spot was marked by a three-niched monument (aedicula) of AD 166–170. Excavations in 1940–49 revealed well-preserved catacombs, with both pagan and Christian graves dating from the period of St. Peter's burial.

Constantine enclosed the aedicula within a shrine and during the last 15 years of his life (died 337) built his basilica around it. The shrine was sheltered by a curved open canopy supported by four serpentine pillars that he brought from the Middle East. The design, enormously magnified, was followed in making the baiacnin (1623–33) over today's papal altar.

In spite of fires, depredations by invaders, and additions by various popes, the original basilica stood for 1,000 years much as it had been built, but in 1506 Julius II ordered it razed and a new St. Peter's built. His architect was Donato Bramante, a Florentine who in 1502 had completed the first great masterpiece of the High Renaissance, the Tempietto in the courtyard of S. Pietro in Montorio, a mile away on the Janiculum Hill. Built to mark the spot where, according to tradition, St. Peter had been crucified, the Tempietto is round, domed, and unadomed. Its outer face is a colonnade of bare Tuscan Doric, the earliest modem use of this order. Because of its proportions, the tiny temple has the majesty of a great monument.

Work of Bramante and Michelangelo on St. Peter's

Bramante's ground plan for St. Peter's was central: a Greek cross, all of the arms of which are equal, around a central dome. Both he and the Pope died before much could be built. Successive architects, including Raphael, drew fresh plans. The last of them, Antonio da Sangallo, died in 1546, and the 71-year-old Michelangelo was solicited to complete Sangallo's projects, which included St. Peter's, the Palazzo Farnese, and the Capitol. He accepted but refused payment for his work on the basilica.

Michelangelo adapted Bramante's original plan, the effect being more emotional and mighty, less classically serene. Of the exterior, only the back of the church, visible from the Vatican Gardens, and the dome are Michelangelo's. After his death Giacomo della Porta and Domenico Fontana, who executed the dome, altered the shape, making it taller and steeper than the original design.

The east end remained unfinished, and it was there that Carlo Maderna was ordered to construct a nave, the clergy having won its century-long battle to have a longitudinal church for liturgical reasons. Thus, St. Peter's orientation reverses the normal. Maderna added a Baroque facade in 1626. Then came Gian Lorenzo Bernini, busy from 1633 to 1677, both inside and outside the building. His pontifical crowd-funnelling colonnade in the shape of a keyhole around the piazza, a fountain for the piazza, the breathtaking baldachin, his several major pieces of sculpture, his interior arrangements for the church, and his dazzling Scala Regia (Royal Stair) to the Vatican exhibit his legendary technical brilliance and his masterful showman's flair. Before the lamentable assault in 1972 which damaged the sculptural masterpiece, one could enter the church and, in the first chapel at the right, see the "Pieta" (1499) of Michelangelo in the original splendour.

All the planning, plotting, labour, and faith of all the popes, priests, artists, and artisans produced a vast, gorgeous ceremonial chamber. Amid the gleam and glitter of gold and bronze and precious stones eddy throngs of awed, dwarfed humanity.

S. Giovanni in Laterano. When Borromini redid the interior of S. Giovanni in Laterano (St. John Lateran) in 1646–50, little of the original Constantinian fabric remained after destruction by the Vandals (5th century), damage by earthquake (9th), two devastating fires (14th), and four consequent rebuildings. The Emperor had built a five-aisled basilica over the remains of the barracks of the imperial guard, the Equites Singulares. The bronze doors come from the Curia (the Senate chamber in the Forum); the silver reliquaries containing the heads of SS. Peter and Paul are copies of the twice-stolen originals.

The octagonal 5th-century baptistery replaced that of the 4th, which had been built into the baths of the House of Fausta, Constantine's second wife. (Later, in another palace, she was strangled in the hot room of the bath, a conventional Roman device for suggesting accidental suffocation of an awkward relative.) Its chapels are decorated with mosaics of the period. The cloisters contain some of the finest examples of early 13th-century carved and inlaid decoration called Cosmatesque after the Cosmati, one of several families of traditional craftsmen. (The cloisters of S. Cosimato, S. Paolo Fuori le Mura, and SS. Quattro Coronati are notable examples of this work, which often was accomplished with porphyries and marbles robbed from classical buildings.)

On the exterior a 1732 facade is topped with 15 giant statues that were visible across the city. The piazza around which the Lateran buildings are grouped is decorated with another obelisk, the oldest and tallest in Rome (15th century BC), one of those erected by Sixtus V late in the 16th century. At the same time, he demolished the old patriarchate, from which the Sancta Sanctorum (the papal chapel) and the Scala Santa (Holy Stairs) were preserved. The Scala had been the principal ceremonial stairway of the palace, but about the 8th or 9th century it began to be identified popularly as having been brought from Jerusalem by St. Helena, Constantine's mother, reportedly as from Pilate's palace and thus the stair climbed by the Saviour. The steps are protected by a wooden cover, and believers mount on their knees. The Scala Santa is not mentioned, however, in ecclesiastic, imperial, or personal writings from the 4th, 5th, or 6th century.

The Holy Stairs

Sta. Croce in Gerusalemme. There is similar lack of record regarding St. Helena's acquisition of the True Cross, which is at Sta. Croce in Gerusalemme. This basilica was built into the palace in which St. Helena lived (317–322). At about this time a hall of the palace was converted into a church and two adjoining small rooms were converted into chapels. The rest of the palace continued to be lived in for centuries. The alleged relics of the cross, found in 1492 walled into a niche, are now in a modern chapel. The facade and narthex of the church are 1743 Rococo, the interior an earlier Baroque with a 12th-century Cosmatesque pavement, some antique columns, a few Renaissance details, and, somewhere within it all, part of a palace built around 180–211.

S. Lorenzo Fuori le Mura. Now in the midst of the Campo Verano cemetery, Rome's Catholic burying ground since 1830, S. Lorenzo Fuori le Mura (St. Lawrence Outside the Walls) dates from the 4th century. The nave is a 13th-century basilica built by Pope Honorius III, and the chancel is another basilica built by Pope Pelagius II in the late 6th century as a replacement for the 4th-century original. On the inner part of the triumphal arch between the two is a 6th-century mosaic, and along the walls are giant Corinthian columns of rare marble taken from a non-Christian building.

S. Paolo Fuori le Mura. A basilica built by Constantine over the Apostle's grave, S. Paolo Fuori le Mura (St. Paul's Outside the Walls), was replaced starting in 386 by a structure mammoth for its time, 328 by 170 feet (100 by 52 metres). It was faithfully restored after a fire in 1823 and thus remains an outstanding example of early basilica1 architecture. It has a single eastern apse, a lofty transept, and five majestic nave aisles. Before the Muslim rampage around the walls in 846, the approach to the basilica was a mile-long colonnade down the Ostian Way from the Porta S. Paolo. Today, after leaving the tomb of Gaius Cestius (died 12 BC), a 120-foot (37-metre) pyramid that has inspired many monument builders since, one-third of the route is fenced by gasworks on one side and warehouses on the other, the rest being given over to waste ground and decaying houses.

Sta. Maria Maggiore. Located on the Esquiline, Sta. Maria Maggiore was founded in 432, just after the Council of Ephesus, which raised the Virgin above all created things; it was thus the first great church of Mary in Rome. Behind its Neoclassic facade (1741–43), the original basilica has resisted change. Most of the mosaics date from the time it was built, lining the walls and bursting with blue and gold around the altar. When a new apse was added in the 13th century, it was also decorated with mosaics. Although the ceiling is Renaissance, the slabs of fine marble and the classical columns are pieces of original plunder from other buildings. The great treasure of the church is the Crib of Christ, five pieces of wood connected by bits of metal. Another pope, St. Liberius (352–366), built another church on the Esquiline in response to a vision of the Virgin, who told him to erect a church where snow fell on the night of August 5. In remembrance, it "snows" white flower petals from the roof of the Pope Paul V chapel in Sta. Maria Maggiore every August 5.

Other major churches. *Gesù.* The mother church of the Jesuit order, Gesù, was built 1568–84. Over the following four centuries, it supplied one of the most pervasively influential designs for church building. Michelangelo offered the new order plans for their first church but died before his plans could be acted upon. Building began under Giacomo Vignola (1507–73), very possibly following Michelangelo's ideas. The Jesuits, shock troops of the Counter-Reformation, proselytizers rather than liturgists, needed a new kind of church for their new approach. Vignola combined the central plan (for preaching) with the longitudinal plan (for ritual) by transforming the aisles into a series of chapels opening into the nave. The facade carried the classical orders upward, though only across the width of the tall nave, and the space above the lower aisles to either side was filled with a scroll. The ideas were not new in the history of architecture, but they were new to Rome and new to the age; and they spread with rapidity.

S. Pietro in Vincoli. Originally the Basilica Eudoxiana, S. Pietro in Vincoli (St. Peter in Chains) was built in 432–440 with money from the empress Eudoxia for the veneration of the chains of St. Peter's Jerusalem imprisonment. Later, his Roman chains were added. The chains became famous after they were mentioned at the Council of Ephesus (431). Michelangelo's thunderous Moses is on the tomb of Julius II. Behind the main altar is a 4th-century sarcophagus with seven compartments, brought to Rome from Antioch during the 6th century in the belief that it contained relics of the seven Maccabees.

Sta. Maria della Vittoria. Built 1605–26, Sta. Maria della Vittoria harbours an unfailing crowd pleaser, Bernini's "Ecstasy of St. Teresa" (1645–52). It is a chapel conceived entirely in theatrical terms, even to having the

The Trevi Fountain, designed by Nicolo Salvi in 1732.
Fenno Jacobs—Photo Researchers

Cornaro family (in marble) seated in opera boxes at the sides of the chapel. Their eyes are directed at the central group in a niche framed in columns, exactly like a proscenium arch, the back wall concealed by gilded metal beams of glory, the scene lighted from above and behind by a hidden yellow-paned window. Amid this setting the angel hovers above the swooning saint, who is—and the illusion is nigh to perfect—borne into the air at the moment of her ecstatic mystical union with Christ. Extraordinarily convincing and utterly voluptuous, it has been both praised as a masterwork of consummate spirituality and condemned as an impious, pornographic peepshow.

*S. **Agostino.*** Of the scores of churches in the Campus Martius of historical, architectural, and artistic interest, the S. Agostino (1479–83) is the most Roman, the church to which would-be mothers come and in which they have offered ex-votos when their prayers have been answered. The "Madonna and Child" (1521) by Jacopo Sansovino, obviously derived from a pagan Juno, is covered with gold and jewels given by the gratified. The church was constructed entirely of travertine looted from the Colosseum. Caravaggio painted the "Madonna with Pilgrims"; Raphael did the fresco of Isaiah. This was these artists' favourite church, and some of the more celebrated among them managed to be interred in it.

THE FOUNTAINS

Rome is as much a city of fountains as it is of churches or palaces, antiquities or urban problems. The more than 300 monumental fountains are an essential part of Rome's seductive powers. Part of the everyday, yet part of the daily surprise, they are points of personal, often sentimental attachment to the city. The Roman composer Ottorino Resphigi found in them inspiration for his orchestral tone poem ***Fontane di Rorna*** (1917; ***The Fountains of Rome***). In their ceaseless pouring forth, they also provide a sense of luxury: on her arrival Queen Christina, having watched the fountains in St. Peter's Square, gave her permission for them to be turned off and was impressed to learn that they flowed all the time.

Legends of the waters

Every fountain has its history and many have legends, the best known of which guarantees a return to Rome to those who toss coins into the Trevi Fountain. Restored after 1,000 years of silence by Pope Nicholas V in 1485, the fountain was renewed in the 17th century and then transformed from a handy source of household water into a scenic wonder. The huge fountain bulges into most of a tiny square and takes up the entire end of an abutting palace. Nicolò Salvi won a 1732 competition by designing a late Baroque marble mass of rocks and rills, rush and gush, beards and buttocks, all very allegorical and damp. It took 30 years to complete. Its water, from Acqua Vergine, was considered Rome's softest and best tasting; for centuries, barrels of it were taken every week to the Vatican and carried off by the jugful by expatriate English tea brewers. Declared nonpotable in 1961, the waters are now recycled by electric pumps.

Out of the Bernini–Borromini rivalry that so enriched the Roman cityscape arose a legend, still believed and recounted today. This explains that, on Bernini's allegorical Piazza Navona fountain, the statue of the Nile River, whose source was then unknown, hides its head to avoid seeing the Borromini facade on the church opposite, and that of the Rio de la Plata raises its arm in alarm to prevent the building from falling. The fountain was, in fact, unveiled in 1651, a year before the Church of S. Agnese was begun, two years before Borromini was called in, and 15 years before the facade was completed. The church is owned and maintained by the Doria-Pamphili family.

The oldest of the city's fountains is really a spring, the Lacus Juturnae in the Forum, restored in 1952 to the appearance it had in Augustan times. The newest fountain in the old city is one of the most admired. Inaugurated as simple jets of water in the Piazza Esedra (now the Piazza della Repubblica) by Pius IX just ten days before the troops of united Italy broke into the city, it was probably the last public work dedicated by a pope in his role of temporal magistrate of the city. In 1901 the nymphs frolicking with sea beasts were added.

The least-liked fountain figure in Rome, unpopular since it was installed in 1587, is on the triumphal arch fountain in the Piazza S. Bernardo, commissioned by Sixtus V. The figure is a pallid Moses, apparently in imitation of Michelangelo's, and its sculptor, Prospero Bresciano, is said to have been so hurt by the public's jeers that he died of a broken heart. (B.E.)

BIBLIOGRAPHY

General works: GEORGINA MASSON, ***The Companion Guide to Rome*** (1965); and ALEC RANDALL, ***Discovering Rome*** (1960), two good modern introductions and guides; MURRAY JAFFE (ed.), ***The Romans' Guide to Rome*** (1965), practical

information supplied by 34 residents of Rome; S.B. PLATNER and THOMAS ASHBY, *A Topographical Dictionary of Ancient Rome* (1929), detailed information on every monument of the ancient city; RICHARD R. and BARBARA G. MERTZ, *Two Thousand Years in Rome* (1968), a popular outline, including suggestions for walks and other tourist information.

History: FERDINAND GREGOROVIUS, *Geschichte der Stadt Ronz im Mittelalter,* 8 vol. (1859–72; reissued by w. KAMPF, 3 vol., 1953–57; Eng. trans. by ANNIE HAMILTON, *History of the City of Rome in the Middle Ages,* 13 vol., 1894–1902), a massive, indispensable reference work; RAYMOND BLOCH, *Les Origines de Rome,* 3rd ed. (1958; Eng. trans., *The Origins of Rome,* 1960), a well-written survey, archaeologically oriented, with illustrations; MAX CARY, *A History of Rome down to the Reigrz of Constantine,* 2nd ed. (1954), one of the best of the textbook surveys; PIO PASCHINI, *Roma nel Rinascimento* (1940); DIEGO ANGELI, *Storia romana di trent'anni 1770–1800* (1931); J.R. CLORNEY BOLTON, *Roman Century, 1870–1970* (1970). on life in Rome and the struggle between the "black" and the "white" aristocracy for ascendancy; H. and A. GELLER, *Jewish Rome* (1970), a pictorial history of the Jews in Rome from 161 BC, text with plates and bibliography.

Antiquities: RODOLFO LANCIANI, *Ancient Rome in the Light of Recent Discoveries* (1888, reprinted 1967), a fascinating account by one of the best of the old-school archaeologists; ERNEST NASH, *A Pictorial Dictionary of Ancient Rome,* 2nd rev. ed., 2 vol. (1968), for the archaeologist, art historian, and interested nonspecialist; DONALD REYNOLDS DUDLEY (ed. and trans.), *Urbs Roma* (1967), classical texts on the city and its monuments, relating the literature of the period to its art and architecture—perhaps the best single book on the ancient city for the general reader.

Art: RONALD BOTTRALL, *Rome* (1968), a detailed guide to the principal museums, galleries, and free-standing monuments; EMILE MALE, *Rome et ses vieilles églises* (1942; Eng. trans., *The Early Churches of Rome,* 1960), churches to the 13th century related to the history of their times; MARIANO ARMELLINI, *Le chiese di Roma dal secolo IV al XZX,* rev. ed. by CARLO CECCHELLI, 2 vol. (1942); ROBERT PAYNE, *The Horizon Book of Ancient Rome* (1966).

Daily life in ancient Rome: uco ENRICO PAOLI, *Vita rornana* (1940; Eng. trans., *Rome: Its People, Life and Custorns,* 1963); JEROME CARCOPINO, *La Vie quotidienne à Rome à l'apogée de l'empire* (1939; Eng. trans., *Daily Life in Ancient Rome,* new ed., 1962); and J.P.V.D. BALSDON, *Life and Leisure in Ancient Rome* (1969), three outstanding works on the subject.

Special topics: BERNARD WALL, *A City and a World* (1962), Rome seen in its religious setting and significance; S.G.A. LUFF, *The Christian's Guide to Rome* (1967); JOCELYN TOYNBEE and J.B. WARD-PERKINS, *The Shrirze of St. Peter and the Vatican Excavations* (1965), a general account of the excavations under St. Peter's; GABRIEL FAURE, *Les Jardins de Rome* (1959; Eng. trans., *Gardens of Rome,* 1960); H.V. MORTON, *The Waters of Rome* (1966), an account of the aqueducts of Rome and their principal fountains.

Personal views: WILLIAM WETMORE STORY, *Roba di Roma,* 8th ed., 2 vol. (1887); STENDHAL, *Promenades dans Rome,* 2 vol. (1829; Eng. trans., *A Roman Journal,* 1957); AUGUSTUS HARE, *Walks in Rome,* 2nd ed. (1878, reprinted 1925); ELEANOR CLARK, *Rome and a Villa* (1952), a personal account of living in Rome and its suburbs and countryside; H.V. MORTON, *Traveller in Rome* (1957); ELIZABETH BOWEN, *A Time in Rome* (1960); AUBREY MENEN, *Rome Revealed* (1960); MAURICE ROWDON, *A Roman Street* (1964).

Rome in photographs: MARTIN HURLIMANN, *Rome* (1954); WILLIAM KLEIN, *Rome: The City and Its People* (1960); R.S. MAGOWAN, *Rome* (1960).

(R.R.R./B.E.)

Rome, Ancient

From humble beginnings in the 8th century BC as a minor people settled on a few hills overlooking the Tiber River, the Romans progressively conquered the Italian peninsula, extended their dominion over the entire Mediterranean basin, and expanded their empire into continental Europe toward the Atlantic. Their conquests allowed them not only to dominate but also to civilize a large number of peoples, on whom they made a lasting impression. Modern historians can still distinguish those regions that were deeply penetrated by the Latin language and Roman law from those that, because of their remoteness, were untouched by it. In attempting to explain the extraordinary and lasting success of Roman culture, it is necessary to seek its sources not only in the vicissitudes of history but also in the psychology of the Roman people. Although the Romans were not gifted in the areas of artistic creativity, scientific research, or philosophic thought, they did possess a steadfast attachment to their native land and a profound sense of political and administrative organization. From a nation of peasants and soldiers, Rome was able to develop an urban civilization, on the pattern of the Greek *polis.* In the domain of art and thought, the Romans were the heirs of the Greeks, and their political genius, common sense, and exemplary piety—which gave them the constant conviction of being sustained and guided by the gods—made it possible for them to create an empire in which the conquered, instead of being kept at a distance, participated in the life and prosperity of the conqueror. No barriers were raised between the nations or races; the emperors themselves often came from outlying regions of the Roman world. The Romans' tenacity, civic feeling, and fundamental liberalism made it possible to establish a lasting peace among men and to confer the benefits of culture and material prosperity on a large number of peoples. Certainly, faults were not absent from a world that could not survive without slavery and that occupied its leisure hours with the unwholesome pleasures of the amphitheatre. The overall balance remains positive, however, and Rome appears to modern eyes as the great teacher of the Western world (see also ROMAN RELIGION; ROMAN LAW).

This article is divided into the following sections:

I. Rome from its origins to 264 BC

EARLY ROME TO THE 6TH CENTURY BC

Greco-Latin historiography

Historiographical problems. For the history of the later republic and the empire there is an abundance of documentary sources, epigraphic texts, and archaeological data that provide the historian with a sound basis for research. But for the founding of Rome and the early years of the republic, there is little evidence. The archaeological remains for this period are rare, and inscriptions, which do not begin to increase in number until the 1st century BC, are scarcer still. There are, however, numerous Greek and Latin writings dating from the end of the republic and the beginning of the empire that recall the birth and early years of the urbs. These later histories have for centuries enriched the literary traditions of the Western world, but their value as history remains problematical. Until recently, critical opinion has been suspicious of the highly colourful accounts of the Roman annalists and has seen in them only fantasy and lies intended to appease the vanity of great families and the patriotism of a proud people. The attitude of many current historians, however, is considerably different, for archaeology has brought indisputable corroboration to tradition. The tale of the founding of Rome by Romulus in the middle of the 8th century BC, for example, has to a certain extent been confirmed by the discovery after World War II of a village of huts on the Palatine effectively dating from about 750 BC. The power and prosperity accorded to the Rome of the Tarquins (Etruscan kings) by the annals are confirmed by discoveries resulting from excavations of ancient Rome. Archaeological and epigraphic discoveries, in fact, continue to demonstrate that the legendary guise of the traditional material actually masks a real foundation of authentic events.

Aeneas and Romulus

The myths of origin and early monarchy. The legend pertaining to the origins of Rome is twofold and distinguishes two events: the colonization of Latium, the territory surrounding Rome, in the 12th century BC by the Trojan Aeneas and the actual founding of the city of Rome 500 years later by the Latin Romulus. The arrival on the Latin plain of Aeneas and his troop of Trojans, who had escaped the destruction of their country, has until recently been viewed as pure legend. Current archaeology, however, has proven that in that remote epoch Italy was visited by Late Bronze Age navigators whose trading activities have been attested as far as southern Tuscany. The legend of Aeneas perhaps expresses the memory of these ancient contacts with the Orient. In any case, the legend of Aeneas can no longer be viewed as the creation of Roman writers of the last centuries of the republic. The figure of Aeneas was familiar to Etruscans from the 6th century BC, and the legend could have made its way to Rome during the period of Etruscan hegemony.

The second aspect of the legend concerns the founding of the town by its eponym Romulus, whose life is depicted in legend as a succession of marvellous and violent episodes. Having miraculously escaped death with his twin brother Remus, he and his brother asked for signs from the gods to designate which of them was to found a new city. Romulus was designated by the divine omens and killed his twin in a murderous quarrel. He then became the first of seven legendary kings who, according to tradition, ruled in Rome from c. 754 BC to the advent of the republic in 509 BC.

The early kings

This account was elaborated gradually and had received its definitive form by the 6th century BC. It rests on a fact now attested by archaeology: the establishment on the Palatine (one of the seven hills of Rome) in the 8th century BC of a primitive village of huts where shepherds of Latin origin had gathered. Much of the legend surrounding the early history of Rome is derived from the account of Livy (59 or 64 BC–AD 17) writing hundreds of years after the events he describes. Livy's first book of Roman history recounts in great detail the activity and achievements of the kings. Mythical, legendary, and strictly historical elements are blended into the text of this narrative, which was based on the annal material. According to Livy, Romulus was a military chief who possessed wide political and military powers. He set out to organize the Roman state and, in order to provide wives for his comrades, planned the rape of the Sabines, a neighbouring people. The war with the Sabines ended in the union of the two peoples, and, after the mysterious death of Romulus in 717, a Sabine, Numa Pompilius, reigned until 673. He was a pious sovereign who dedicated himself to giving Rome, founded by force and arms, its laws and morality. It was he who first gave form to the Roman religion. He was succeeded by a warrior king, Tullus Hostilius (traditional reign 672–641), who framed the military code. The famous episode of the Horatii and Curiatii, the champions of Rome and Alba Longa, respectively, occurred during a war Tullus Hostilius conducted against Alba Longa. The line of pre-Etruscan kings ended with Ancus Martius, a Sabine, who was the

Archaeological accounts

grandson of Numa Pompilius. He supposedly expanded Rome and founded Ostia. Beneath the surface of the traditional account rests the reality of the progressive growth of a town that initially was only a federation of villages perched on the Palatine and neighbouring hills. Recent archaeology has made it possible to trace from early times the course of the growth of Rome, which became a true city only with the arrival of the Etruscans. The Latin population settled on the hills overlooking the Tiber presumably brought with it its language and its scrupulous piety, expressed in a whole series of ancient rituals that were faithfully preserved until classical times.

The reign of Servius Tullius

The Etruscan hegemony over Rome. According to tradition, the end of the 7th century and the 6th century BC were taken up by the reigns of the Etruscan sovereigns Tarquinius Priscus and Tarquinius Superbus, between whom was interposed a man of obscure origin, Servius Tullius. The annalistic account gives many dramatic details of this period, during which Rome obeyed rulers of foreign origin who allegedly made the very name of royalty odious to the Romans. As legend has it, Tarquinius Priscus was a native of Tarquinii, where his father Demaratus, a Corinthian Greek, had taken refuge. This first of the Etruscan kings supposedly ruled from 616 to 579 BC. He was a good administrator and a great builder and gave Rome new lustre. Assassinated by the sons of Ancus Martius, who considered him a usurper, he was succeeded by Servius Tullius, whose name suggests a man of humble origin. Many historians believe that Servius was an interpolation by later annalists, who, for patriotic reasons, refused to ascribe the achievements of this period to the Tarquins. Servius was an enlightened king who was supposed to have established the organization of the city by census results and the distribution of the citizens into classes, in turn divided for military purposes into centuries (groups of 100). Servius is alleged to have divided the town into four territorial tribes and encircled it with a strong, unbroken wall. Assassinated in turn by the sons of Tarquinius Priscus, Servius was succeeded by one of them, Tarquinius Superbus, whom legend depicts as a violent and unjust tyrant. He successfully waged war against neighbouring peoples and continued his father's construction work. The Cloaca Maxima (a remarkable drainage system) and the completion of the temple of Jupiter Capitolinus are attributed to him. According to legend, the excesses of the Tarquins provoked an insurrection among the Roman citizens. Lucretia, an honorable Roman matron, supposedly committed suicide after being violated by one of the King's sons. Lucius Junius Brutus, a freedom-loving Roman, roused the outraged populace and condemned the tyrant and his family to exile. Thus was the Roman Republic born and power transferred to its legitimate custodian, the people themselves.

Etruscan influence on Rome

This fictionalized and patriotic account contains some elements of truth. Rome came under the ascendancy of its powerful neighbour Etruria in the 7th century BC. Somewhat later than tradition would have it, near *550,* it was ruled directly by the Etruscans, who thus obtained an invaluable passageway to Campania, which they controlled. As a result of the Etruscan seizure of power, Rome became a great and prosperous city with a preponderant influence over the rest of Latium. The Etruscan leaders encircled it with a continuous wall and adorned it with stone monuments, the most imposing of which was the Capitoline temple, which was dedicated to Jupiter, Juno, and Minerva. The gods were Roman, but their grouping in a triad reflects the Etruscan predilection for grouping things in threes. Rome was thus genuinely affected by the Etruscan religious and cultural influence. The Etruscans' more significant achievements included passing on to the Romans the alphabet, which they had themselves borrowed from the Greek world, and acquainting them with certain Hellenic religious and art forms.

It should not be concluded, however, that Rome owed a great debt to its 6th-century Etruscan masters. On the contrary, the limits of their influence should be emphasized. Although the Etruscans physically transformed and reorganized Rome, the city on the whole withstood any profound exertion of Etruscan influence. The Latin tongue was little influenced by the Etruscan language, which had a quite different structure. In addition, once the Etruscan occupation had ended, Etruscans and Romans followed widely divergent paths. Rome had changed considerably in outer appearance, hut its people, who, after a century of foreign domination, patiently and stubbornly continued their progress and history, remained the same.

The development of Roman institutions. *Social institutions.* At the very basis of the Roman city were the gentes, groups composed of all those linked by a common ancestor, who gave the gens its name and character. Each gens had its own cults and traditions that stubbornly withstood the ravages of time. In addition to those citizens bound together by the blood of the gentes, the city included a varying number of *clierztes,* men who owed allegiance to patrons in return for aid and assistance. The gens comprised subdivisions, the *familiae,* the more limited but fundamental units of Roman society. They were grouped around the head of the family, the *paterfamilias,* whose authority is often demonstrated in Roman history by striking examples.

As far back as the period of monarchy, a privileged class, the patriciate, gradually began to form, made up of members of the privileged gentes. Opposed to it was the plebs, which was excluded from the gens organization. The plebeians had neither a cult nor common ancestors. They initially formed an unorganized class made up of former clients of vanished gentes and foreigners who had settled along the Tiber as traders or artisans. Lacking civil, political, or religious rights, the plebs initially also had no civic obligations and was liable to neither military service nor taxation. Although the reform attributed to Servius Tullius is full of dating errors, the Etruscan monarchy, following the principle of organizing the city according to census figures, was responsible for re-enlisting the plebs into the army.

The priesthoods

Religious institutions. The intense religious life of ancient Rome had achieved in remote times the essential organization it was to maintain for centuries. Greek and Etrurian influences were assimilated but did not profoundly alter the traditional religious structure. Several priesthoods of varying importance were charged with administering the *jus divinum* ("divine law"), keeping order within the city, and maintaining good relations with the gods. In the period of the monarchy, the king possessed religious authority in addition to his political, judicial, and military roles. Under the republic, this authority was passed on to a priest appointed for life, the *rex sacrorum,* who resided in the Regia in the Roman Forum. But the real religious power lay elsewhere. *Flamines* ("priests"), of which there were three major and 12 minor in the classical epoch, were assigned to the cult of individual deities. The *flamines* of Jupiter, Mars, and Quirinus were of remote origin and guaranteed the benevolent presence of the ancient triad of gods. The augurs were a body of theologians responsible for preserving and applying the rules pertaining to the observation and interpretation of the auspices, divine omens revealed by, among other things, the flight of birds. The pontiffs, particularly the *pontifex maximus,* held wide powers. They were the trustees of sacred knowledge and saw that it was strictly applied. They supervised the venerable college of the Vestals, virgin priestesses appointed for 30 years, who were charged with keeping the sacred fire of the city in the Temple of Vesta. There were also ancient brotherhoods, such as the Arvales and Salii, who assured the survival of ancient magic rituals.

Political and military institutions. The political and military power of the kings was absolute. This power, called the *imperium,* developed out of the right of military command but also had a religious dimension. The ceremony of the king's investiture was, in fact, controlled by a prior examination of the auspices through the offices of the augur. According to annals, each accession to the throne was normally contingent on a popular vote and the Senate's confirmation. What is certain is that the Roman monarchy was not hereditary and that it functioned for at least two and one-half centuries.

Early political divisions

According to legend, the kings, beginning with Romulus, organized the Roman people politically. They supposedly were divided into three tribes and 30 *curiae.* The primitive tribes may have corresponded to various ethnic elements: Latins, Sabines, and Etruscans. Each tribe was divided into ten *curiae,* which resembled the gentes and were local in character. Each *curia* was headed by a *curio* and had common meals and sacrifices. Together they formed the assembly called the Comitia Curiata, which elected the king and possessed a certain number of legislative and judicial rights.

The mos majorurn: the Roman virtues. Such was the general structure that controlled the life of Rome during the early centuries of its existence. Simultaneously stable and flexible, it made possible the progressive development of a town that was self-governing and at the same time open to constructive influences from abroad. The cardinal Roman virtues, which were Rome's most steadfast support and the guarantor of its survival, had already asserted themselves. Rome was characterized by a profound patriotism that consistently subordinated the individual to the group and always resolved the conflicts of duty to the benefit of the state. The deep sense of the permanence of the state assured a continuing respect for ancestral custom, although it did not entail the rejection of innovation. But this continuity could be assured only by the exercise of a high level of morality, which gave first place to the civic virtues of unwavering piety and good faith. *Virtus, pietas, fides:* these guaranteed the survival and growth of Rome.

EARLY CENTURIES OF THE ROMAN REPUBLIC (6TH CENTURY BC–264 BC)

Dedication of the temple of Jupiter Capitolinus

The overthrow of the monarchy. The annals give a coherent and highly colourful account of the fall of the monarchy and the establishment of the republic. According to this account, the tyranny of Tarquinius Superbus became hateful to the Romans, who rose under the leadership of Lucius Junius Brutus after the rape of Lucretia. The Tarquins were banished from Rome; the popular assembly, grouped by centuries, named the first two magistrates in 509, Brutus and Tarquinius Collatinus. The king of Clusium, Lars Porsena, laid siege to Rome in an attempt to restore Tarquinius Superbus to power; but, according to legend, the heroism of Horatius Cocles and Mucius Scaevola forced him to retreat. The dedication of the temple of Jupiter Capitolinus by Marcus Horatius, who replaced Spurius Lucretius, Brutus' successor, as magistrate, also took place in 509 after the exile of Tarquinius Superbus and his family. In the same year was allegedly begun the custom of drawing up the list of eponymous magistrates, the consular *fasti,* the first and foremost document of Roman history.

Analysis of the various elements of the tradition shows that the annalists, for patriotic reasons, more or less consciously telescoped events that occurred over several years into that *annus mirabilis,* 509. The reliable basis for the chronology of the period is the dedication of the Capitoline temple of Jupiter, Juno, and Minerva, which took place in 509 BC. At that time began the annual rite of driving a nail into the wall of Minerva's cella in the Capitoline sanctuary. By counting the number of nails the true date of the dedication was easily obtained. But the Etruscan departure from Rome and the subsequent founding of a republic freed from foreign occupation has to be placed later, even as late as 470 BC. Archaeological data and the persistence of religious ceremonies show that until that time there was no break in Roman life. The Etruscans abandoned Rome and Latium only after they had been defeated by the Greeks in the waters of Cumae (474 BC), when they were already on the decline. The annals placed their departure in 509 BC because patriotism required it. The Capitoline temple, which was to become the symbol and centre of Rome and its empire, could not remain what it had in fact been, a purely Etruscan achievement. Its dedication at least had to be Roman. By arbitrarily placing the expulsion of the Tarquins in 509, the annalists could ascribe the glory of the dedication to Horatius, a true Roman.

Later, the expulsion of the Tarquins was regularly celebrated as the end of a monarchy that stood for the domination of an individual and the servitude of the citizens. In the folklore of later Romans it was at this point that *dominatio* and *servitus* vanished from Roman soil and the state became the property of its legitimate owner, the Roman people, who exercised power. Such was the foundation of Roman liberty (or *libertas),* the prestige of which remained such that even Augnstus, the first emperor, improperly laid claim to it by proclaiming himself *libertatis vindex* ("saviour of liberty").

The chief magistracy and the dictatorship. With the end of the monarchy, a republican constitution gradually began to take form. The Romans' pragmatism led them to create a whole series of powers designed to balance one another while providing the maximum effectiveness to the essential proceedings of the state. Power was divided among the Senate, the popular assemblies, and magistrates of varying degrees of importance. What gradually took shape was an essentially oligarchic republican constitution in which a governing class, the *nobilitas*, consistently held the foremost position.

The king's sovereign executive power appears to have passed initially to a magistrate with the title of *praetor maximus*, who no doubt headed a body of *praetnres,* who had perhaps been the immediate officers of the monarch. It was a *praetor maximus*, Marcus Horatius, who in 509 dedicated the Capitoline temple.

The consuls

The *praetor maximus*, however, soon gave way to a collegiate system in which two elected consuls shared the *imperium* for a single year. The principles of collegiality and of limitation of the term of office were characteristic of the republican constitution and effectively prevented the reappearance of tyrannical and absolute authority. It thus appeared safe to grant royal insignia to the magistrates and to have them preceded by 12 lictors bearing the fasces and the axe, hidden to Rome but visible on the other side of the *pomoerium*, the trench that defined the city limits.

Apparently, the plebeians initially were not excluded from the consulate, but the patriciate quickly closed ranks and made it extremely difficult for anyone not of its order to accede to the chief magistracy. In the middle of the 5th century, pressure from the plebs caused the patricians to create the military tribunate with magistral authority. a kind of promagistracy open to plebeians (see below). The consulate, however, remained open only to patricians until the mid-4th century BC, when a law was passed requiring that one of the two consuls be a plebeian. By that time the ancient distinction between the patriciate and the plebs had become blurred, and a new nobility occupied the forefront of the political scene.

During times of national emergency, the dual consulship was abandoned and a dictator established in their place. The dictator was, in fact, appointed by one of the consuls on the initiative of the Senate, and he was free of the restrictions imposed on the other magistrates. The dictator's *imperium* was unlimited but was entrusted to him for six months only, after which he surrendered his authority, as in the legend of Cincinnatus, who was appointed dictator in the war against the Aequi and afterward returned to his farm. The dictator was backed up by a cavalry commander (or *magister equitum*) of his own choosing, who had to resign at the same time as the dictator. This special form of magistracy was seldom used. It no doubt arose from an early institution of the Latin towns, which reserved special dictatorships for carrying out certain religious rites. The dictator disappeared from Rome after the Second Punic War and reappeared only during the struggle of the 1st century BC among ambitious men who used it to fight their way to personal power.

The praetorship

Other magistracies. Initially, the *imperium* was exercised only by the consuls, but in 366 BC, according to tradition, a praetor was created who also possessed *imperium* but of an order inferior to that of the consuls (*imperium minus).* The praetor was given the responsibility for the administration of justice in the city. Beginning in the second half of the 3rd century, the growth of Rome and its conquests entailed the creation of other praetors who were charged with administering the law in the provinces.

Tradition credits King Servius Tullius with having taken the first census of the citizens and their property. Under the republic, a magistracy, important but *sine imperio,* fulfilled this function. This was the censorship, created, according to tradition, in 433. Two censors were elected every five years, but at the end of 18 months, after they had completed their task, they resigned their posts. They made a census of persons and property and were able to bring disgrace upon a citizen for bad conduct. They thus exercised a true *cura morum,* and this concern for morality made them important and respected figures.

As administration became increasingly complex, more officials were created to aid all of the high magistrates. The quaestors specialized in financial questions and were the custodians of the public coffers. The college of four aediles (two patricians and two plebeians) was charged with maintaining public order in the town and provisioning it, along with the upkeep of public monuments and the organization of games.

Judicial institutions. From remotest times the Romans displayed a concern for legal rules, the religious character of which was striking. According to tradition, a religious brotherhood of 20 members (the *fetiales*) appeared as far back as the 8th century BC, in possession of the laws of warfare. War could not be declared nor treaties concluded without applying strict rules, of which they were the trustees. Rome recognized neither the rule of the strongest nor the right of the conqueror over the vanquished in warfare. The Romans' respect for legal obligations permitted them to conduct only a war that was *pium justumque* ("pious and just"), so that the gods would aid the Roman army to victory. Piety, good faith, and justice were interrelated and, in the mind of the Romans, were needed to guide their struggles and justify their conquests.

Law of the Twelve Tables

In the middle of the 5th century, the law of the Twelve Tables was framed and displayed in the Forum on 12 bronze tablets. It was Rome's first written code and was of extreme importance because it was viewed as the source of all law, public as well as private. The schoolchildren of Cicero's time still learned it by heart. The code survives only in fragments, but enough remains to enable modern historians to grasp its spirit.

The code seems to have been the creation of a special committee. There is an interruption in the consular *fasti* for 451 and 450, when the consuls were replaced by a board of ten men called decemvirs, who ruled with consular authority and who were charged with putting the laws into written form. The first college of decemvirs was appointed for one year and was replaced by a second college that was to complete the task; but the second college tried to retain power illegally. Military defeats and scandals in Roman life brought about its downfall and the consuls' return to power. During their brief rule, however, the decemvirs had sufficient time to fulfill their task of drafting a written code, to which anyone might refer. Until then the law had remained purely oral, and the plebs accused the magistrates of applying it unfairly. According to the ancient writers and from the fragments of extant laws, the code of the Twelve Tables seems to be the heir of a remote past, for many of its provisions present an essentially religious and sacred character. The code, however, also displays the distinguishing feature of judicial law: the assurance to every citizen of the exercise of his rights.

The Senate. In the political life of republican Rome, the Senate constituted a third element, in addition to the popular assemblies and magistracies. Its prestige was great and its influence of continuing importance. The republican Senate was heir to the Senate of the royal epoch, which, according to tradition, was created by Romulus and brought together the most influential of the *patres familiarum.* The royal Senate had been responsible for nominating the king and was, no doubt, his constant adviser, already enjoying an *auctoritas,* an effective power of moral force with religious overtones, which lent value to its acts and gave them unusual weight in the citizens' eyes. After the fall of the monarchy, the Senate was presided over by the acting consul and represented the most stable political authority of the republic.

The Senate was originally composed of heads of the patrician gentes but soon came to include all former magistrates of the city, whose names were entered on the roll and who thus became *conscripti.* This probably explains the name borne by the Senators under the republic —*patres conscripti.* This term denoted two groups of different origin and should be understood as having arisen from the formulation *patres et conscripti.* At first senators were appointed by the consuls, and later by the censors, so that changes in the Senate's composition occurred only every five years, when new censors were elected. Senators were seated for life, and their hierarchy corresponded to that of the magistracies that they had previously held.

The Senate's *auctoritas*

The Senate's *auctoritas* manifested itself in all areas. In the early years of the republic, the Senate approved the laws after they had been voted upon in the assemblies, but from the second half of the 4th century, Senate action no longer followed but, instead, preceded the popular vote. Senatorial assent was succeeded by the right of initiative in proposing laws. The Senate expressed its opinions in the form of a *senatus consultum* (a decree of the Senate), the phrasing of which followed an unvarying scheme. Although it was not obligatory, the magistrates were aware that they should abide by the opinion of the wisest and most respected men of the city. In time, the Senate's opinions gained the force of law.

In the later republic, the Senate also performed the important function of assigning magistrates to govern the provinces. Since the political career of a Roman statesman usually depended upon obtaining a lucrative provincial appointment, this became of great significance in the power struggle of the 1st century BC.

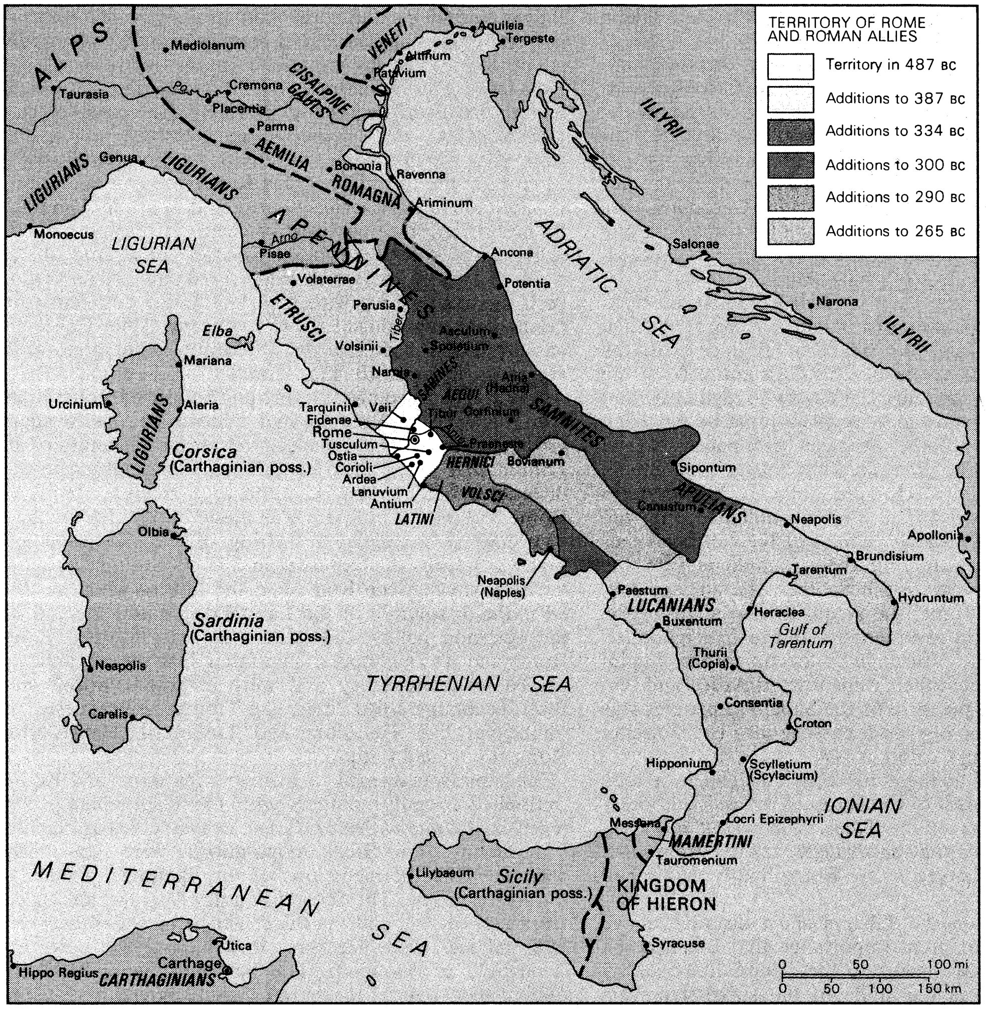
The Senate also had great authority in foreign affairs. The ambassadors of foreign powers presented themselves before it, and it chose the envoys to be sent abroad. With respect to religion, it upheld the ancestral rites and ceremonies, accepting foreign divinities when the need was felt and securing Rome against innovations dangerous to the *mos majorum.* Above all, it had supreme control over public finances since it managed the state treasury (the *aerarium Saturni)* and made final decisions on expenditures and taxation.

In the early centuries of the republic, the Senate thus had supreme influence, due to custom and to the respect in which it was held by the magistrates and popular assemblies. The Roman constitution, which appeared to be based on a combination of forces, was in fact essentially oligarchic and plutocratic. The senators, who were proprietors of great estates, were thrifty, competent administrators. Their real difficulties arose with conquest, which gave them not only a town and a country to govern but a true empire to administer as well.

Plebeian institutions. In a patrician state that wished to exclude all those who had not issued from its ranks, the plebeians, forced into the background, put all of their efforts into forcing recognition of their existence and rights. The first two centuries of the republic were taken up by their unceasingly renewed attempts to gain the political position that they claimed and by the progressive disintegration of the privileges to which the patricians clung tenaciously. The struggle took place on the economic as well as on the civic levels. The plebeians wanted the expanses of the public lands (the *ager publicus),* over which the great proprietors were extending their influence, opened to them. The public domain was expanding continually as a result of the gradual conquests of the Roman armies.

Struggle over public land

The long conflict between patriciate and plebs went through various changes, described by the annalists in accounts that are often unreliable and contain errors of dating. What is certain is that in the first decades of the republic the plebs was able to organize itself and form a state within the state. The unity of the republic was ulti-



Roman expansion in Italy, 487–265 BC.

From A. Boak and W. Sinnigen, *A History of Rome to A.D. 565*; Copyright © 1965 by The Macmillan Company

mately maintained, but not until serious crises had threatened the very foundations of its existence.

Plebeian religious institutions

The process of organization and defense of the plebeians was apparently as follows: around the beginning of the republic, 493–92, according to tradition, the plebs, returning armed from a military campaign, seceded to the Sacred Mount and forced the Senate to grant to it magistrates charged with defending the plebs under all circumstances. The tribunes of the plebs were inviolable and sacrosanct and could bring aid (*jus auxilii*) to plebeians threatened by the possible excesses of the executive power. They also had a right of veto, through which they could prevent the consuls, Senate, or assembly from making a decision, and they could intercede to prevent the carrying out of a decision if they judged it to be unfairly detrimental to the plebs. The tribunes' power, which they exerted only within the city, operated within the sphere of magic and religion. In the face of the regular Roman authorities, it appeared that the plebs could be protected only by figures from its own ranks, who were effectively protected by divine power and who themselves enjoyed a sacred power from the depths of antiquity. Immediately after the tribunates were established, the consul Spurius Cassius dedicated a tripartite temple at the foot of the Aventine, which was constructed on the model of the Capitoline sanctuary but devoted to Ceres, Liber, and Libera. This was to be the appointed sanctuary of the plebs. Ceres was the protectress of agriculture and kept the deadly famine away from the urban population. The couple Liber and Libera presided over the birth and growth of animals and people. Permanently established on the Aventine, this triad of fertility protected the plebs, whose land and herds guaranteed only an uncertain subsistence, and, to some extent, counterbalanced the Capitoline triad, the great protector of the Roman state and its regular authorities, who were patricians. The temple on the Aventine held the plebeian archives and, in the later republic, became the centre for wheat distributions to the plebs. Aediles were created to protect the temple, guard the archives, and keep order in the markets.

Completing the organization of the plebeian class was the assembly of the plebs, the Concilium Plebis, which was organized according to the citizens' domiciles instead of by property qualifications, as in the Comitia Centuriata.

The state was thus apparently divided in half. Rome was wise and fortunate in its ability, through successive measures, to bring the opposing orders to an understanding. A law of 456, which determined the sale of the lands on the Aventine, signalled the first extension of the ***ager publicus.***

THE EXPANSION OF ROME IN ITALY

Rome and its Latin neighbours. Roman historians provide a detailed picture of Rome's foreign policy and its conquests during the early centuries of the republic. There is splendour and grandeur in that account, which often takes on an epic rhythm, in which Rome, engaged with many powerful adversaries, fells them one by one, relying on allies, of course, but bearing the burden of the battles and struggles itself. There are many more or less conscious errors of dating in these highly colourful pages, whose coherence and frequent beauty filled the Romans of the classic era with satisfaction and pride in their ancestors. One important error of perspective makes Rome a capital among the other towns of central Italy as early as the 5th century BC and gives the city a pre-eminence over neighbouring cities that it certainly did not have. Historical research today makes it possible to take a more realistic view of the situation.

The Latin League

The Etruscan departure from Latium to a certain extent weakened Rome, for the Etruscans had fortified Rome as a base for their expeditions to the south of the peninsula. Rome joined the Latin League and had to adopt its policy since it certainly was incapable of imposing its will on it. The entire 5th century and the first part of the 4th was the era of the corzciliurn Latinorum. Decisions were made in common, and in time of war the Romans, along with everyone else, placed themselves under the command of the dictator elected by the league.

These common perils did not prevent Rome from breaking with its Latin neighbours. Rome's victory at Lake Regillus in 499 BC and the treaty of alliance with the Latin towns—known as the Foedus Cassianum from the name of the consul, Spurius Cassius, who signed it, and ascribed by tradition to the year 493—must be put back into the framework of the Etruscan occupation, which had not yet ended. The great conflicts between republican Rome and its Latin rivals for pre-eminence over Latium have to be placed later. The major problem during the 5th century was curbing the peoples of the central Apennines, who were multiplying their incursions into the western plains. These were the Sabines, the ancient enemies of monarchic Rome; the Aequi, whose territory extended from the Anio to Praeneste; and their neighbours, the Hernici and the Volsci, who occupied the Lepini mountains and directed their raids toward Ardea and Antium. Somehow the pressure of the Sahellian peoples was contained, and Rome opened a passage into the Volscian country by occupying Corioli in 444.

Rome found itself at the same time engaged in a long conflict with the nearest of the Etruscan cities, Veii, located on the right bank of the Tiber. Veii was a powerful and formidable rival, and the struggle was to be long and to undergo many changes before Rome finally triumphed over its opponent.

The siege of Veii

Fidenae, situated on the Tiber a short distance above Rome, was a port of great importance that blocked the salt route going from the sea to the mountains, used by the Etruscans. It was the stake in the successive wars between Rome and Veii, which firmly held Fidenae. Rome did not definitely take possession of it until 426 BC. Then, according to tradition, in 405 began the siege of Veii itself, which lasted ten years (like the siege of Troy) and ended in the city's collapse under the blows of the dictator Furius Camillus. Livy paints an epic picture of this siege and its dramatic outcome. The religious notes that sprinkle his narrative certainly derive in part from traditions that are unmistakably Etruscan. Fate and the oracles had announced the impending end of Veii, abandoned by the gods themselves. Camillus, for his part, was fate's moving spirit. He sent an envoy to the Delphic Apollo and promised the tithe of the spoils if he were successful. He was assured of support, but he was loath to attack the town without appealing to Juno, the goddess who protected it. He therefore practiced the ancient rite of evocatio, inviting the deity to leave the besieged town and come to live in Rome, where she would be received with honour. Juno accepted and the town was stormed and captured, after which the statue of the cult left the ruins of Veii and arrived on the Aventine, where it was given a temple suitable to Juno's dignity.

The Gallic invasion and further conquests. The capture of Veii was followed by a catastrophe for Rome that lived on in its history as a disastrous event, the Celtic invasion and the capture of the city by the Gauls. Book V of Livy and the pages dedicated by Polybius to the Celtic world make it possible to portray vividly the vicissitudes of this dramatic period. At the end of the 6th and the beginning of the 5th century, the first bands of invaders began to arrive in northern Italy. In the 4th century the trickle became a flood, and the barbarian hordes penetrated the peninsula in force. According to the annalists, the Gauls allegedly defeated the Etruscan army near Clusium and continued southward toward Rome and crushed the Roman forces on the banks of the River Allia, to the north of the city. About 390 (the traditional date) they succeeded in routing the Roman army, which was dismayed by the appearance and mode of combat of its new adversaries. This day remained in the memory of the Romans a dies *religiosus* (an "accursed day"). Caere, Rome's Etruscan neighbour, received the sacred objects (the sacra) rescued by the Vestals from the pillage and ruin, after which a common religious atmosphere seemed to envelop the two cities. Once Rome was out of danger, its gratitude to the Caerites prompted it to grant them public hospitality, and it soon gave them citizenship without suffrage (civitas sine *suffragio*). After the drama of the occupation of Rome, which engendered an acute feeling of peril revived by Livy with a sort of tragic grandeur, the Gauls were summoned back to the north. Other Celtic invasions followed during the 4th century, but none gave rise to the terror of the first assault.

Since the Latin confederation had abandoned Rome to its own devices, the Romans had confronted the invaders virtually alone. There ensued a period of anarchy in Latium, and Rome took advantage of it to subjugate the disordered Latin towns. A temporary agreement between Rome and the Latin towns was made in 358 BC.

The Samnite Wars

Through a complicated shifting of alliances, Rome's first war against the Samnites in 343 BC brought about a decisive war between Rome and the Latins, which resulted in the dissolution of the Latin League and the end of the common policy undertaken, at least in theory, by Rome and its neighbours. The latter lost the advantages gained by confederacy and were assigned various statuses, becoming allied cities, like Tibur and Praeneste, or municipia, like Tusculum and Lavinium. Henceforth, Rome dominated Latium.

Campania was opened to Roman expansion, and Rome multiplied friendly contacts with the Campanian privileged classes and with such Greek coastal cities as Naples. The Oscan and Greek populations were constantly threatened by the incursions of the Samnites, a martial mountain people. The Second Samnite War, which lasted from 326 to 304, and the third, which extended between 298 and 290, were necessary to subjugate the powerful stronghold of Samnium. The Samnites, who faithfully obeyed their federal magistrates, were moved by a fierce and bellicose spirit. Well-armed and good tacticians, they inflicted heavy losses on the Romans. In 321 they trapped the Roman army in the closed valley of Caudine Forks between Capua and Beneventum; the two consuls, who survived, accepted an agreement that was harsh for Rome, and the entire army surrendered. Rome did not ratify the Pax Caudina, but it had suffered a deep humiliation. Its tenacity enabled it to regain the advantage, and Bovianum, one of the Samnite capitals, was stormed in 306, and the peace signed. At the same time, Rome pressed into central Etruria, which was living out its last years of independence. The Third Samnite War broke out in 298, in which the Etruscans, Gauls, and Samnites united against the increasing threat of the Romans. In 295 the consuls Fabius Rullianus and Decius Mus won a great victory at Sentinum, north of the Umbria, over the Samnito–Gallic coalition, in which Decius, according to legend, heroically sacrificed his own life. The Gallic tide was stemmed, and the Ager Gallicus Romanized. The conquest of Etruria was completed with the fall of the last independent Etruscan town, Volsinii, in 265 BC.

Wars with Pyrrhus

The southward expansion of Rome, which was still on good terms with Carthage, transformed the republic into a Mediterranean power. The Romans came into contact with peoples such as the Lucanians and the Apulians, over whom they extended their influence, which engendered the hostility of the most powerful city of Magna Graecia, Tarentum. When a Roman garrison and ships appeared in areas of Tarentine authority, Tarentum, sensing danger, decided to put a stop to it and appealed to Pyrrhus, the king of Epirus, to bring the insolent Romans to their senses. Pyrrhus thought he could unify the Hellenic cities of southern Italy and the native populations of the interior. Although this ambition miscarried, his expeditionary corps, using formidable combat elephants, defeated the legions on several occasions between 280 and 276 BC. From Carthage, then friendly to Rome, he won back the

Greek cities of Sicily that had fallen under Punic hegemony. But Pyrrhus' victories were short-lived. He returned to Epirus, and Tarentum had to concede defeat in 271. Rome accorded Tarentum the status of an allied city.

The Roman mastery of Italy. Thus, by the beginning of the 3rd century BC, Rome had become mistress of Italy and was rapidly becoming a Mediterranean power. Tarentum had bequeathed to Rome its hegemony in the commerce of the Adriatic. The Greek world for the first time turned its gaze toward this latest arrival on the international chessboard, this city of the west whose dynamism asserted itself through a series of conquests that neither difficulties nor defeats had been able to check. In 306 the agreement with the Carthaginians was renewed once more, the last time before the outbreak of the hostilities that were to lead to great upheavals. This last treaty forbade all Carthaginian intervention in Italy and all Roman intervention in Sicily. Forty years later, however, Rome interfered in Sicilian affairs, and the former allies found themselves in opposition.

The Roman concept of imperialism

In its conquest of Italy, Rome had given proof of a flexible diplomacy, which treated the vanquished honorably and gave them a place, and rights that varied, in a country gradually being unified under Roman law. Roman imperialism sought constantly to justify itself morally, and the demands of Roman good faith (*fides Romana*) were not just empty words, even though such honesty was not always observed. Through deditio, the vanquished relied on the *fides* of Rome, which claimed to use the new power conferred on it for the good of all, in the best way possible. In point of fact, the defeated city usually lost its autonomy and incurred various obligations to Rome, so that after the dissolution of the Latin League, the status of Rome's former allies was diverse. Lanuvium and Tusculum, for example, had citizenship without suffrage (*civitas* sine *suffragio*), while Tibur and Praeneste remained allied cities (civitates foederatae), which did not prevent them from losing part of their territory. Rome also sent new colonists into the conquered regions, and the Roman colonies spread throughout the ever-expanding *ager publicus* were like extensions of the metropolis in a more or less distant land. The population of the Latin colonies was composed largely of proletarians who cleared the conquered soil and brought it under cultivation. Because of this vast extension of Rome's national territory, economic development of the city was rapid. The city soon had need of a currency of its own for its expanding and intensified commercial relations. The war with Pyrrhus gave rise to the first Roman silver coinage.

On the eve of the First Punic War, Rome had become a commercial city participating in the great economic currents of the Hellenistic world. The wars against Carthage forced Rome to become a naval power. Rome could then operate in the east and undertake the expeditions and conquests beyond the seas that ultimately would lead to the extension of its hegemony throughout the Mediterranean world. (R.Bl.)

II. The middle republic (264–133BC)

THE FIRST TWO PUNIC WARS

Roman invasion of Sicily

The First Punic War and the aftermath. The Romans invaded Sicily, in 264 BC, in response to an appeal by the Campanian mercenaries who had seized Messana by treachery in 289 BC. The Romans feared that if they did not act, Carthage, which had saved Messana from King Hieron of Syracuse, would remain in possession of it and stand at the doorstep of Italy. Land fighting broke out in Sicily, in which the Romans had Hieron for an ally after 263; but because Carthage was the first naval power of the western Mediterranean, decisive victory could only be won at sea. So the Romans, who had no naval experience, built a navy and turned their soldiers into sailors and their consuls into admirals. At first, the war went well for the Romans. The consul Marcus Atilius Regulus even invaded Africa in 256, after a naval victory, and the Carthaginians treated for peace; but Regulus was not satisfied. The Carthaginians rallied, however, under the Spartan mercenary captain Xanthippus, and in 255 Regulus was defeated and taken prisoner. After naval disasters in 255 and 253 and a long period of stalemate, the Romans built a new fleet, and with it Gaius Lutatius Catulus won the battle of the Aegates Islands (241) off western Sicily. Carthage could no longer provision its army in Sicily; so peace was made in 241, with the Carthaginians agreeing to pay an indemnity and to abandon Sicily, which, with the exception of Hieron's kingdom, became the first province of the Roman Empire.

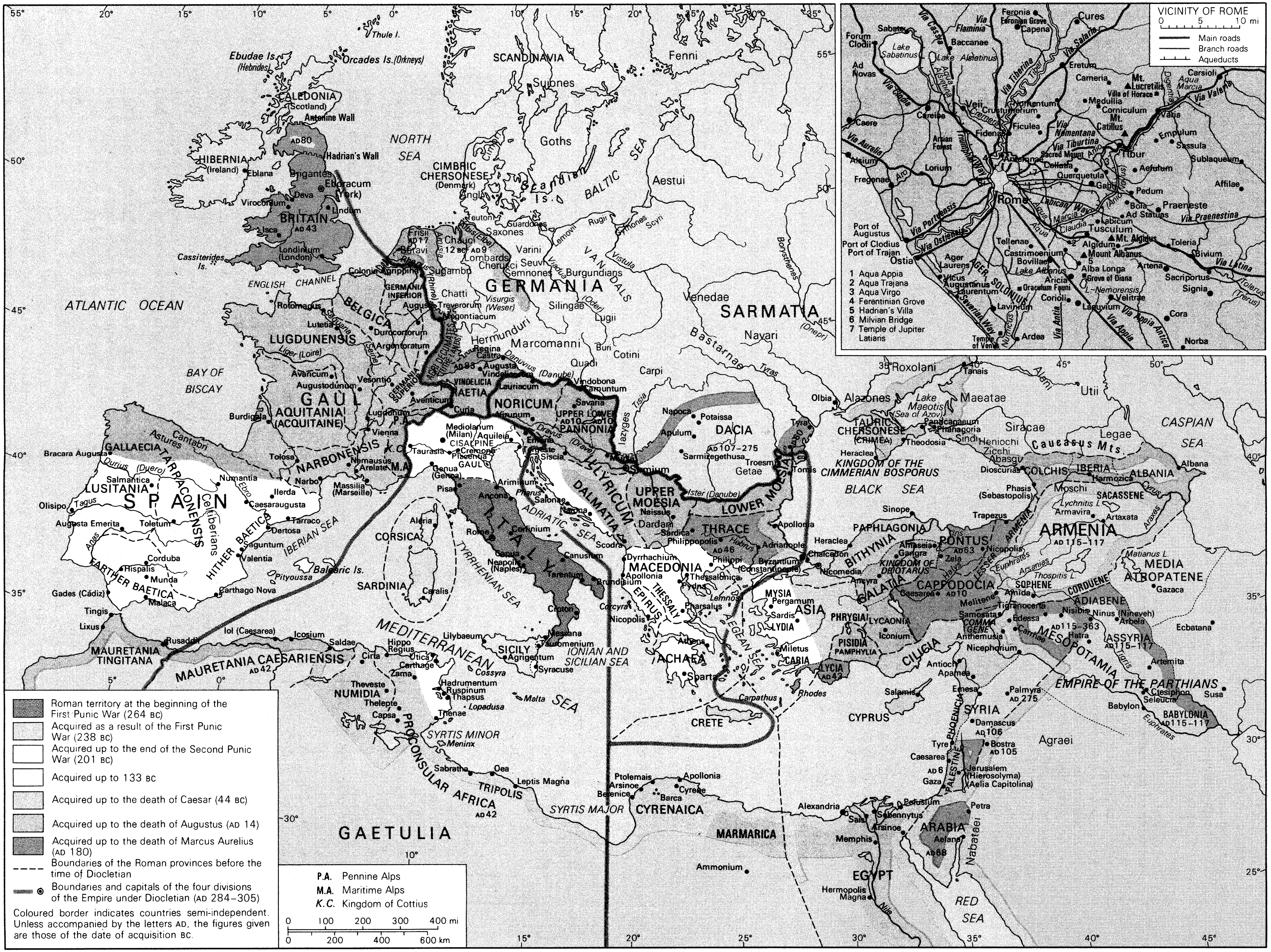
A serious and prolonged mutiny in the Carthaginian army (the "Truceless War"), caused by an attempt to underpay mercenary troops on their discharge, enabled the Romans, with no pretext of justice whatever, but again from fear of the Carthaginians as near neighbours, to take possession of Sardinia and Corsica in 238. This seizure exacerbated the Carthaginians and in particular their general, Hamilcar Barca, who still resented the necessity of capitulation in 241 because his own position at the time, on Mt. Eryx in western Sicily, had been, he thought, unassailable. He, therefore, now turned, with the support of the Carthaginian government, to extending Carthaginian power in Spain, with the intention of using its manpower and resources to mount a war of revenge against Rome. (On his death, in 229, Hamilcar Barca was succeeded as commander in Spain by his son-in-law Hasdrubal; Hamilcar's son Hannibal. who had served under them both, assumed command in Spain on Hasdrubal's murder in 221.) Prompted no doubt by its ally Massilia (Marseille), which was concerned by the growing Carthaginian threat to its own colonies in Spain, Rome protested to Hamilcar in 231 and made an agreement with Hasdrubal in 226, in which he pledged not to push Carthaginian expansion beyond the Ebro. Then, or earlier, Rome made an alliance with Saguntum, a strategically placed city near the coast that lay within the Carthaginian sphere of influence, no doubt with the thought of using it as a bridgehead for operations in Spain if war with Carthage should be resumed.

Even more serious to Rome than the implications of Carthaginian expansion in Spain was the threat of a new Gallic invasion of Italy from the north. There had been an alarm in 236, and in 225 the Boii, Taurini, Insubres, and Gaesati moved southward with more than 50,000 men; after an initial success, these forces were caught between two Roman armies and defeated at Telamon on the west coast, north of Cosa. The Romans followed up their success and in 218 established colonies at Placentia and Cremona, both on the Po. With Hannibal's invasion of Italy that year (for he had rightly counted on Gallic assistance), the work of pacification in the north was undone for the time being.

There was a further diversion east of the Adriatic. Attacks by state-protected Illyrian pirates on Italian shipping and the murder of a Roman ambassador provoked the Romans to declare war on the Illyrian queen Teuta in 229, and a swift victory was won. The Romans then assumed a protectorate over 120 miles of the Illyrian coast, including the ports of Apollonia and Dyrrhachium (modern Durazzo), from Lissos in the north to Corcyra in the south. Greece was delighted to see Illyrian piracy suppressed, and the Aetolian and Achaean leagues, Athens, and Corinth politely received Roman envoys. In 219 the terms of the 229 agreement were violated by the widespread naval depredations of Demetrius of Pharos, an ally of Rome in the first war; and, perhaps with the intention of avoiding the danger of a stab in the back by Illyria if war with Carthage broke out, the Romans went to war again in 219 and again won conclusive success—Pharos was captured and Demetrius fled. That same year Hannibal attacked and took Saguntum; and Rome, with this Illyrian distraction on its hands, did nothing to help its ally.

Hannibal's invasion of Italy

The Second Punic War. The war that the Romans had planned—an invasion of Carthage and a simultaneous landing in Spain to form a Roman bridgehead at Saguntum—never took place. Hannibal assumed the initiative by invading Italy through Gaul and over the Alps, exploiting his superior military genius and experience to defeat a succession of sanguine but inexperienced Roman generals and untrained Roman armies at the Ticinus and



Roman territorial expansion.

Adapted from W. Shepherd, *Historical Atlas*, Harper and Row, Publishers (Barnes & Noble Books), New York, Copyright © 1964 by Barnes & Noble, Inc.

Trebia rivers in 218, at Trasimene in 217, and at Cannae in 216. To all appearances, the war was now over; and the greedy Philip V of Macedon broke off his war against the Aetolians, allied himself to Hannibal, and declared war on Rome. Capua joined Hannibal, expecting to become Italy's new capital city; Syracuse seceded from Rome in 215, and Hannibal secured Tarentum in 213. Yet from 215 Hannibal's chances of success dwindled. Historians have questioned his not marching on Rome when it was demoralized after Cannae and wondered if the luxury of Capua, where he wintered, sapped the vigour of his troops.

There are more positive explanations of the turn of the tide. The first is the fact that the Roman Senate never panicked, fortified by the unruffled good sense of Quintus Fabius Maximus, consul in 215 and 214, who by refusing to risk a further pitched battle with Hannibal earned his nickname, "Cunctator." Second is the fact that Rome's Latin allies remained loyal and that Hannibal's army—without reinforcements either from Philip, who with no fleet could not cross the Adriatic, or from Carthage, whose government and navy showed none of Hannibal's enterprise—was a declining and, eventually, a spent force. The army's strength was cavalry; and cavalry cannot besiege cities. Directly after the invasion the Romans decided to keep open a second front in Spain in the hope of preventing reinforcements from reaching Hannibal by land, a plan followed even after a disastrous defeat in 211 and the deaths of two great generals—the elder Publius Cornelius Scipio and his brother.

The Roman recapture of Capua and the fall of Syracuse in 211 were straws in the wind; and the Aetolian League in Greece now allied itself to Rome and relieved Rome of the burden of the war against Philip. Young Publius Cornelius Scipio, who had been sent, at the age of 26, to take over his father's and his uncle's command, failed to prevent reinforcements for Hannibal from leaving Spain under Hasdrubal Barca, Hannibal's brother, in 208. But these forces were disastrously defeated in Italy at the battle of the Metaurus a year later, and Hannibal's last hope was shattered. He remained ineffectively in southern Italy.

The Aetolians, complaining of inadequate support from the Romans, made peace with Philip in 206; and the Romans, concentrating on bigger things, followed suit in 205. Scipio completely defeated the Carthaginian armies in Spain at Ilipa in 207.

With a surge of optimism, the Roman people became determined to take the offensive and end the war by invading Carthage. Scipio was the obvious man to command the invasion; so, though under the legal age, he was elected consul for 205. Despite opposition from Fabius and his supporters, who thought the defeat of Hannibal in Italy should have top priority, Scipio crossed to Africa in 204 with an army of about 25,000. Hannibal was recalled, and Scipio, with the valuable assistance of Masinissa, the Numidian king, defeated Hannibal at Zama in 202 and won for himself the name of "Africanus."

By the peace settlement, Carthage was to pay 10,000 talents in 50 years and to keep no elephants and only ten warships. It was forbidden to wage war outside Africa, or in Africa itself without Rome's consent. Masinissa's power was greatly strengthened on his western border. The Romans assumed direct responsibility for the administration of Spain and established two provinces—Farther and Nearer Spain (see further PUNIC WARS).

THE ESTABLISHMENT OF ROMAN HEGEMONY IN THE HELLENISTIC WORLD

Some scholars believe that Rome made peace with Philip V of Macedonia in 205 in order to concentrate on defeating Carthage but that it intended to settle accounts with him later; others (and this was evidently what Philip himself believed), that Rome intended the settlement to be final. In the latter case, Rome was provoked to declare war against Macedonia (the Second Macedonian War) in 200 by the news brought in 201 by Attalus I of Pergamum (an ally in the first war) and by the Rhodians that Philip had made a secret compact with Antiochus III, king of Syria, to capture and share the external possessions of Egypt, whose new king, Ptolemy V Epiphanes, was a boy.

Provocation for war against Macedonia

Wars against Philip V, Antiochus III, and the Aetolians. Rome's declared object was to force Philip to withdraw Macedonian troops from the three key positions he had garrisoned in Greece—Acrocorinth (since 224), Chalcis in Euboea, and Demetrias on the Gulf of Pagasae. Rome secured the alliance of the Aetolian League; and, with some heart searching because of recent friendly relations with Macedonia, the Achaean League in the Peloponnese; and also of Attalus and Rhodes, whom Philip had provoked in Asia Minor. From the Adriatic, the Romans forced Philip's strongly held position on the Aous River and advanced into Thessaly under Titus Quinctius Flamininus in 198. In 197 they defeated the Macedonian army at the Battle of Cynoscephalae in Thessaly, in which the Roman legion showed its superiority over the Macedonian phalanx.

Peace was made with Philip on generous terms; but the Aetolians, exasperated by being denied the acquisition of towns in Thessaly for which they had hoped, maligned Rome by suggesting that, rather than freeing Greece, the Romans proposed to garrison the country themselves. Flamininus gave the lie to this claim when he proclaimed the liberty of all Greek cities at the Isthmian Games of 196, and the last Roman soldiers left Greece two years later. Rome had kept its word; but some Romans already doubted the wisdom of this move because, with as yet undeclared intentions, Antiochus, who had possessions on the European side of the Hellespont (Lysimacheia, Aenus, and Maroneia) and who had accepted Hannibal as an adviser, was already moving into western Asia Minor.

Negotiations between Antiochus and Rome foundered over the former's refusal to relinquish his possessions in Europe, and the dissatisfied Aetolians solved Antiochus' problem by inviting him, in 192, to save them from the Romans. Antiochus crossed to Greece with inadequate forces and with inadequate supplies to join the Aetolians, who now declared war on Rome. He had no help from Philip, who honoured his peace with Rome. The Achaean League stood by Rome, as did Eumenes 11 (who had succeeded Attalus as king of Pergamum in 197) and the Rhodians, both of whom expected to profit from Antiochus' defeat.

The Romans, under Acilius Glabrio, defeated Antiochus and the Aetolians at Thermopylae in 191. Antiochus retired to Asia; his navy was soundly defeated in two battles by Roman fleets in cooperation with the Rhodians; and in 190 Lucius Scipio (with his brother, the great Africanus, on his staff) crossed the Hellespont and defeated him at Magnesia. Peace was made, Antiochus being compelled to abandon his possessions in Europe and his territory in Asia west of the Taurus range and retaining only a token navy of ten ships. The abandoned Aetolians were forced to capitulate on humiliating terms in 189, becoming subject-allies of Rome. That same year the Romans, under Manlius Vulso, conquered the Galatians in Asia Minor.

Rome's policy in the East was fraught with ambiguity. In the Second Macedonian War, the Romans had crusaded for the "freedom of the Greeks," and, by their own withdrawal from Greece in 194, they had given the appearance of honest crusaders. But although Rome removed Macedonian garrisons from cities in Greece and Asia Minor, Aegina remained the property of Pergamum, having been sold to Attalus in the First Macedonian War, and the Roman peace commission of 197 had even considered giving Eretria in Euboea to Pergamum. After Magnesia, Rome had no compunction about giving a number of Greek cities in Asia Minor to Pergamum and Rhodes.

Establishment of a protectorate over Greece. Considerations of its own interest drove Rome to curtail the aggressive potentiality of Philip, Antiochus, and the Aetolians, with all of whom, after their defeats, Rome was bound by treaty and, so, was justified in acting if terms of these treaties were infringed. Relations with the

Relations with the Achaean League

Achaean League were less clear-cut, and the problem came to a head in 195. The prime object of Achaean policy was to include the entire Peloponnese by the acquisition of Elis, Messene, and in particular Sparta; Sparta's communist regime, established by King Cleomenes III in the previous century, had been revived by Nabis and now evoked envious admiration among the proletariat of the Achaean cities, thus appearing to the governing class in those cities a dangerous contagion that if not stifled might spread. Apprehensive of the slaughter and chaos that might follow an Achaean conquest of Sparta, Flamininus accepted the submission of Sparta in 195 and insisted on its remaining independent; and indeed Rome made a treaty with it.

In 194 Philopoemen, a headstrong Achaean patriot who had separated from Flamininus because of mutual jealousy and antagonism, returned after six years from Crete and controlled the league's policy. In 192 he outwitted Flamininus (now back in Greece with a roving commission) and brought Sparta into the league with the promise of no interference with its government. Then, in 188, he settled the issue once and for all with such acts as eliminating the enemies of the league, pulling down the walls of Sparta, abolishing communism, and enslaving the helots once more and putting them up for sale. In 183 Messene seceded from the league in a short-lived revolt, and Philopoemen, seeking to recover it, was captured and killed.

The year 181 was, in the Greek historian Polybius' view, decisive for the future relationship of Rome and the league. Philopoemen was dead, and the Achaean Callicrates persuaded the Senate at Rome that its interest lay in giving uncritical support in the league to the unequivocal supporters, like himself, of the Roman point of view.

Worsening relations with Macedonia

The Roman government carried a considerable responsibility for the worsening of relations with Macedonia and the eventual outbreak of the Third Macedonian War. Philip's refusal to join Antiochus was of untold value, and he was rewarded by the Romans' return of his younger son Demetrius, a hostage in Rome since 197, and the remission of his outstanding indemnity. But Philip felt himself cheated when he was not allowed to keep the Thessalian cities he had taken from the Aetolians when cooperating with Glabrio in 191. Nor was he allowed to recover Aenus and Maroneia on the Thracian coast, which had once belonged to him and were now given up by Antiochus, even though King Eumenes of Pergamum had been granted Lysimacheia and the Thracian Chersonese, which Antiochus was also forced to abandon. Still worse, Rome made an unprincipled attempt to split the Macedonian royal house. Demetrius had won great popularity in Rome when a hostage and, when he returned to Rome in 184 as Philip's representative to answer a variety of charges against the King, the Romans agreed to accept Philip's defense only because Demetrius was his mouthpiece. Neither Philip nor his elder son Perseus could be blamed for thinking that the Romans wished to upset the natural succession. When it was suspected that Demetrius planned to escape from Macedonia to Rome, and when Macedonian envoys returning from Rome in 181 produced a letter from Flamininus that, whether or not it was forged, suggested that Demetrius was conspiring for power, Philip had him killed. By this time Philip had decided on war with Rome; but he died a year later (179), having by his campaigns in the north in the latter part of his reign greatly increased the resources of his kingdom both in wealth and manpower. The Romans believed these campaigns to be exploratory—the prelude to an attempt to invade Italy from the northeast. Philip's son Perseus turned from the north to a policy of reconciliation with the Greeks and, though he was not a soldier of his father's calibre, roused increasing suspicion at Rome, which Eumenes adroitly fostered. Marriage connections suggested that Perseus might become the nucleus of a formidable anti-Roman combination—his own marriage to the daughter of Seleucus IV of Syria and his sister's marriage to Prusias II of Rithynia.

The Third Macedonian War

The issue was decided when Eumenes, himself a probable sufferer if Macedonia re-emerged as a great power, visited Rome in 172 and persuaded the Senate of Perseus' aggressive plans. The Third Macedonian War was sparked off by a couple of bizarre episodes—an amateurish attempt to kill Eumenes on his way to Delphi while returning from Rome, thought to be instigated by Perseus, and the "evidence" of a prominent citizen of Brundisium that Perseus had invited him to poison eminent Romans who had lodged with him on their way to and from the East. It was a difficult war; the Roman army had to land at Apollonia and cross the mountains into Thessaly. Perseus at the start had the advantage of numbers; indeed he won the first battle, in Thessaly in 171, and after it he offered to treat for peace. The offer was refused; and when, in 170 and 169, the Romans failed to bring Perseus to battle, the belief grew, not only in openly anti-Roman circles, that Perseus might not be defeated after all. The Illyrian chief Genthius joined Perseus. Secret talks were held between Eumenes and Perseus, the full truth of which was never revealed, with the object that, for a suitable price, Eumenes should endeavour to secure a negotiated peace. And the Rhodians, whether or not they had misunderstood Marcius Philippus, the Roman commander in 169, thought that they had been encouraged by him to try their hand at peace negotiations. On June 22, 168 BC, however, the Roman commander Lucius Aemilius Paullus brought Perseus to battle at Pydna and defeated him decisively. Perseus capitulated.

Division of Macedonia and Illyria. Perseus lost his throne and survived in exile for two more years at Alba Fucens, near Rome. Macedonia was divided into four independent republics. Genthius had already been defeated in Illyria and was also condemned to exile in Italy. Illyria, like Macedonia, was given its independence; it was divided into three units, but not before 150,000 of its men had been taken and sold into slavery.

Treatment of Eumenes II and the Achaean League. The effects of the Roman victory over Perseus were momentous. Eumenes, the Achaean League, and Rhodes were all suspect. Although Attalus II, brother of King Eumenes, was received in Rome in 167, Eumenes himself was met by a quaestor at Brundisium and told that the Senate would receive no kings in Rome—a snub that was reinforced by the audience already given to his enemy, King Prusias of Bithynia, who in Rome startled the Senate by doing obeisance to them and addressing them as saviour gods. The Achaeans were required to send 1,000 hostages to Rome; they were relegated to country towns in Italy until 151, when the 300 survivors were allowed to go home. One of the hostages, Polybius, did well, living as a favoured guest of the Scipios and laying the foundations of his great history.

Reduction of Rhodes. Rhodes fared worst of all. Already in 177 it had discovered the ambivalence of Roman policy: having been "given" Lycia and Caria on the mainland by the Romans after Magnesia and having proceeded to their direct administration, it was informed by Rome, in consequence of a Lycian appeal, that it had not been given Lycia and Caria at all but had simply been hound to them in friendship and alliance. Now, in 168, Rhodes's delegates, sent to propose a peaceful settlement of the issue with Perseus, arrived in Rome to learn that the war was over. A Roman praetor in 167 proposed that Rome should declare war on Rhodes, but the censor Cato pointed out that whatever the Rhodians' sympathies in the recent war might have been, they had performed no act of overt hostility. So the Rhodians received and forcibly obeyed Rome's harsh commands—to abandon Caunus and Stratoniceia on the mainland, neither of which had been received from Rome, and to withdraw from Caria and Lycia. Delos was declared a free port, and Rhodes consequently lost about six-sevenths of its income from harbour dues. Then, having reduced Rhodes to a second-class power, Rome consented, in 165, to make a formal alliance with her.

Piracy, based largely on Crete, was always a menace, and Rhodes, with Syria, had been the policeman of the eastern Mediterranean. With the Syrian navy already reduced to a token force and Rhodes now weakened, piracy flourished unchecked; and in the following century the Romans were themselves the victims of their own shortsighted policy.

Establishment of provinces in Africa and the East. Further east, as part of their policy of preserving a balance of power between Syria and Egypt, the Romans in 168 stopped the invasion of Egypt by the Seleucid ruler Antiochus Epiphanes. A Roman legate, Gaius Popillius Laenas, caught up with him and ordered him home, and the King did as he was told. Between 151 and 146 Rome found itself at war with Carthage, Macedonia, and the Achaean League, as a result of which all three powers lost their independence; and the provincial empire of Rome, previously confined to the west, was extended into Africa and into the Balkans.

The *Third* Punic War. In Africa, Carthage studiously respected the peace treaty of 201 despite constant filching of territory by its Numidian neighbours under King Masinissa and by the hostile decisions of successive Roman boundary commissions; but in 151 Carthaginian patience was exhausted and it declared war on Masinissa. The octogenarian Cato—the very man who had scotched the proposal to make war on Rhodes in 167—was fanatically determined that Carthage must be eliminated. Cato was frightened, it seems, by the evidence of Carthage's economic prosperity; for he cannot have thought it strong enough militarily to fight Rome with any prospect of success. In Rome, Scipio Nasica opposed a declaration of war on the grounds that Rome would be in danger of relaxing its military preparedness if Carthage, its historical rival, disappeared. War was declared, however, and both consuls of 149 were sent to conduct it. The Carthaginians were prepared to accept almost any terms—to give up 300 noble hostages, to surrender all their arms—but not Rome's final demand that they leave their city and settle ten miles inland; abandon the sea, which was their livelihood; and turn to farming.

Destruction of Carthage. Carthage was besieged and, despite the surrender of arms, defended with incredible resolution until, in 146, Scipio Aemilianus (son of Aemilius Paullus and adopted by a son of the great Scipio Africanus) captured it and burnt it down. Its site, one of the finest harbours in the Mediterranean, was cursed and abandoned. Carthaginian territory, with its capital at Utica, became the Roman province of Africa.

Subjugation of Macedonia and Greece. Macedonia too, with Illyria and Epirus and with Achaea as an appendage, became a Roman province in 146. The Macedonians had missed their monarchy, and when, in 149, a certain Andriscus claimed that he was the son of Perseus, he found followers and was soon in possession of the whole country. Andriscus was defeated in 148 by Quintus Caecilius Metellus (who won the name "Macedonicus") and was brought to Rome and executed.

Appropriately enough, given past history, the independence of Sparta was the issue that brought an end to the Achaean League. Callicrates had died, and control of league policy passed to fanatical opponents of Rome—Diaeus and Critolaus. Sparta had seceded, and in 147 a Roman commission announced its decision that Corinth and Argos should be detached from the league as well. Negotiation was impossible in the consequent atmosphere of insensate fury. The league was ready for war. The Roman leader Metellus Macedonicus defeated Critolaus north of the isthmus; Lucius Mummius, consul of 146, arrived and defeated Diaeus. Corinth was pillaged, its wealth of artistic treasures plundered and dispatched to Rome. The city was destroyed and abandoned.

The independent kingdom of Pergamum survived until 133, when its last king, Attalus III, died and, having no heir, bequeathed his territory to Rome. With the title of "Asia," it became Rome's seventh province.

Development of an empire

Beginnings of Roman provincial government. The Roman Empire was founded and developed as Rome successively assumed direct responsibility for the government and administration of conquered territories (provinces): of Sicily in 241, of Sardinia and Corsica (always administered as a single province) in 238, of Spain (the two provinces of Nearer and Farther Spain) in 197. It was at the start a western empire. Then, after half a century, came Macedonia (with Illyria and Achaea) and Africa in 146, and finally Pergamum (the province of Asia) in 133. Each province had a separate written constitution (a lex *provinciae*) with detailed regulations made by the conquering general in association with a commission of ten senators (*decem* legati) acting within broad regulations laid down by the Senate. This constitution named the cities within the province that retained their independence, uncontrolled by the governor's jurisdiction (civitates liberae); they were usually cities that had assisted Rome at the time of the conquest and that in some cases were already bound to Rome by treaty (*civitates foederatae*). The constitution also specified the nature of the taxation, whether direct (*stipendium*) or indirect (tithes [*decumae*] on agricultural production, which were the rule in Sicily and Sardinia).

Provincial *administrator*. At the head of the Roman administration were the annually appointed governor and the financial secretary (one of the quaestors of the year). The governor's function was expected to be predominantly administrative (he would constitute the province's high court of justice) and comparable to that of a praetor in Rome. Because the governor was normally a praetor, the number of praetors was, therefore, increased to four in 227 and to six in 197, lots being cast each year to determine the postings. But because the Hannibalic war and the great wars in the East made other demands on praetors (*e.g.*, naval command) it was frequently necessary to extend a governor's provincial command for more than a year. His power (*imperium*) remained the same, but after his magisterial year he was a propraetor or, more properly, a proconsul.

Only when there was a major war in a province, or the prospect of one, was a consul sent out to be its governor. This happened ten times in the case of Nearer Spain and four times in Farther Spain in the period 153–134 BC, and it happened in Sicily after the outbreak of a serious slave revolt in 134 and 133.

Central authority. Consciousness of imperial responsibility, which rested with the Senate, grew slowly and was, perhaps, never fully grasped (except in rare cases, such as by the elder Cato and the elder Tiberius Gracchus) until, under the empire of the Caesars, the responsibility was assumed by the emperor; and only when faced with undeniable evidence of misgovernment and corruption did the Senate take corrective measures. In 149, after a scandalous acquittal in a public trial of Servius Sulpicius Galba, a criminally corrupt governor of Farther Spain, a permanent court (the quaestio repetundarurn) was established in Rome, with senators as jurors, before which provincial subjects—provided that a Roman would act as their patron—could impeach Roman officials for extortionate practices.

The Roman economy was quick to feel the benefits, however unethical, of the great wars and of the new provinces. Apart from the wealth brought back to Rome in official and private loot, there were the indemnities from Carthage, Philip, and Antiochus. The grain paid in taxation in Sicily and Sardinia was sold in Rome at a price with which homegrown grain could not compete, with disastrous effects on large sections of Roman agriculture. Rome also gained wealth from the Spanish mines. In 167 the property tax for Roman citizens in Italy was abolished—evidence of imperialism's dividends.

Apart from brigandage, which had always been endemic in Sardinia, the island provinces, until the great Sicilian revolt of 134, caused little anxiety to the Roman government. Spain, however, was a different matter. Fighting in the mountains was difficult, the policy of resettling communities in cities in the plains was hard to enforce, and the peninsula was never completely conquered and pacified until the time of Augustus. From 154 there was continuous war against both the Lusitani in Farther Spain and against the Celtiberi in Nearer Spain—a war that produced a Spanish resistance leader of genius, Viriathus. After his murder in 139, the war continued until at last Scipio Aemilianus captured the stronghold of Numantia in 133. The war had its repercussions at home. Military service in Spain was exceedingly unpopular, and in both 151 and 138 tribunes imprisoned the consuls, so as to prevent the levies (behaviour that was to point the way to the Gracchi).

Securing the empire. Of the 70 consuls between 200 and 166 BC, 55 were engaged for part of their year of office in campaigning with armies in the north of Italy against the Gauls, particularly against the Ligurians in the northwest and the Boii in the northeast, with the object of clearing the northern Apennines and bringing the peninsula up to the Alps under firm control. Once a district was sufficiently pacified, Latin colonies were founded at strategic points, manned by Latins (and also, sometimes, by Romans) who were attracted by the prospect of free land even in potentially dangerous areas. And great roads were built—always a vital element in Romanization—for communication and, in particular, for the rapid movement of troops.

Establishment of colonial outposts

Particularly in the northeast, fighting was sometimes critical, sometimes a mere matter of triumph hunting (a "Ligurian Triumph" became an expression of mockery), and the serious fighting ended after Marcus Baebius Tamphilus, consul in 181, removed 40,000 Apuani with their families to the neighbourhood of Beneventum in southern Italy. A Latin colony was settled at Luca in 180, and a Roman colony at the port of Luna in 177. In the northeast a Latin colony was sent to Bononia (Bologna) in 189, and another, with unusually large allotments of land, to Aquileia in 181.

The names of the two consuls of 187 are perpetuated in the Via Flaminia from Arretium (Arezzo) to Bononia and in its continuation, the Via Aemilia, from Bononia to Placentia. In 175 Bononia was connected by road with Aquileia, and in 148 a road was driven diagonally across the peninsula from the Mediterranean through the Po Valley to the head of the Adriatic—the Via Postumia, which ran from Genna (Genoa) through Placentia, Cremona, and Verona to Aquileia.

In southern Italy a number of small citizen colonies, each of 300 settlers (coast-guard colonies), were dispatched to vital points on the coast after the Hannibalic war to strengthen coastal defense, particularly in view of fears of a possible invasion of Italy by Antiochus' fleet under the command of Hannibal. These were founded in 194: Puteoli, Salernum, Volturnum, Liternum, Sipontum, Buxentum, Croton, and Tempsa.

ROMAN GOVERNMENT, ECONOMY, AND CULTURE

Polybius, deeply experienced in the political life of the Hellenistic world, completed, during and after the time he was a hostage in Rome, a vast history of the years 264 to 146 BC that shows how Carthage and Macedonia during this period were absorbed into Rome's empire and how the rest of the Hellenistic world fell under the dominating influence of Rome. He tried to determine the secret of Rome's uninterrupted success. He might, as did Philip V (in a letter known from an inscription), have emphasized the liberality with which, in contrast to Greek city-states, Rome increased the number of its citizens through granting citizenship to slaves and aliens, and the manner in which it often resettled, instead of enslaving, conquered peoples. Or he could have stressed the system of clientela (the dependence of the poor and unprivileged on a rich and distinguished patron, which was reflected in public life when they recorded their votes), including foreign clientelae (the protection given to the interests of provincial subjects and foreign kings by particular Roman families of prominence). But he did not. He emphasized the binding force of Roman public religion as inducing in Roman public servants, through the sacrosanctity of oaths, an integrity that had no parallel in contemporary Greece. He was impressed by Roman loyalty to the state and, within the great families, loyalty to their own distinguished traditions. He stressed the unrivalled efficiency of the Roman army as a fighting force.

Elements of government. Above all, Polybius was struck by the manner in which the three elements—of monarchy (the consuls), aristocracy (the Senate), and democracy (the popular assemblies)—struck a balance of power that guaranteed a stable equilibrium; this is not an entirely attractive theory, however, because it greatly exaggerates the independent power and importance of the two annual consuls, whose position was anything but monarchical.

The Roman sovereign body was the Senatus Populusque Romanus (Senate and the Roman People). The Senate's membership depended on the two censors—very senior senators elected normally at five-year intervals. They admitted to the Senate those who had held the quaestorship within the preceding five years (quinquennium), and then they went through the list of existing senators, expelling those who had acquired any public notoriety (*infamia*), as Cato expelled the brother of Titus Quinctius Flamininus in 184 for scandalous behavior when campaigning in northern Italy. The revised list of senators was then published, the first name on the list being that of the princeps *Senatus*, who was invited to speak first in any debate at which he was present. Publius Scipio Africanus was *princeps Senatus* from 199 until his death, in 183; Marcus Aemilius Lepidus, from 179 until his death, in 152.

Qualifications for public service

In the early 2nd century rules were made governing the succession of offices in a public career (*cursus* honorum). The first such rule stated that no one could be elected consul if he had not already been praetor. In 180 minimum ages were established for holding office—for plebeians 36 was the minimum for the curule aedileship, 39 for the praetorship, and 42 for the consulship; patricians, it seems, could hold the magistracies two years earlier. The quaestorship could be held after the ten years' cavalry service done by youths of the upper classes (as against 16 years' infantry service to which others were liable). This service started at age 17; so the quaestorship could be held at age 27, though there are recorded cases when it was held at 25. Whether the quaestorship was an essential part of the *cursus* is uncertain. Tenure in the curule aedileship and the tribunate of the plebs were not compulsory; but the ambitious rarely omitted them because they offered an opportunity to win popularity, which was likely to be reflected in the voting when candidates stood for the praetorship. (The curule aediles were responsible for the public games, and the tribunes of the plebs had the opportunity to champion popular discontents; for instance, during the calling-up for the wars in Spain in 151 and 138, the tribunes stopped the enrollment of soldiers by imprisoning the consuls.) Beginning in 267 there were eight quaestors, whose duties were largely concerned with public finance; there were four aediles (two plebeian, two curule), and ten tribunes. Before 227 there were two praetors (the second, praetor peregrinus, since 242), who were wholly occupied with judicial duties in Rome; after 227 there were four, and after 197 there were six, the additional praetors being required as provincial governors.

The consuls. There were two consuls; therefore, only one in four of those elected to the quaestorship could eventually reach the consulship and so ennoble his family, if it was not noble already (the *nobiles* were families that could number a consul among their ancestors).

If both consuls were in Rome, they held the fasces (badge of authority) and presided in the Senate on alternate months, the consul who had headed the poll presiding in the first month of the year (March until 153 and January after that). In the period of the great wars, however, the consuls generally left Rome early in the year, after they had discharged their necessary formal obligations (for instance, the celebration of the Latin games on Mons Albanus), in order to command armies on campaign; they normally returned at the end of the campaigning season. In the interval, the urban praetor presided over meetings of the Senate.

One of the consuls was made responsible for the holding of the elections, which took place late in the year. If the consul was unable to return to Rome for this purpose, he could nominate a dictator—a man who held office for three nundinae (probably 17 days) and had no other function to perform. If both consuls died before the elections or if the year ended without the elections taking place, there was an interregnum (sovereignty reverting to the patrician members of the Senate, as had happened in early days after the death of a king). An interrex then held the election.

A consul or praetor, on the termination of his magistracy, could retain his imperium with the title of promagistrate (proconsul or propraetor); such extension of *imperium* and use of promagistrates developed quickly during the Second Punic War because the number of necessary military commands exceeded the number of available magistrates. The former consuls Publius Cornelius Scipio and his brother, for instance, commanded in Spain as proconsuls until they were killed in 211. This practice was continued and extended to meet the exigencies of provincial government in the second and first centuries.

Political debates in the Senate

The Senate. Although there were public meetings (*contiones*), it was in the Senate that political debates on the highest issues of state took place. Members were invited to speak in order of seniority, a fact that explains the great weight and importance of the former consuls (the *consulares*) in the Senate. In the middle republic these were members of a limited number of the great families whose influence, exerted through their clientes, assured their election—consuls and praetors were elected by the Comitia Centuriata, in which the rich exercised a predominant influence. The results of the elections show clearly that the presiding magistrate was able to exert great influence in favour of candidates who were related to him or who were his political associates.

In the Senate the consulares did not speak with a single voice; and there was strong infighting, certain families—for example, the Cornelii Scipiones and the Aemilii, and the Claudii and the Fulvii—being closely linked in political friendship (*amicitia*) with one another and in opposition (inimicitiae) to other groups. Such groupings, however, were not rigid, for groups dissolved and re-formed around particular prominent political figures and the policies favoured by those politicians. The Scipios, for example, favoured considerate and friendly relations with the Hellenistic kings and welcomed the genial influence of Greek civilization on Roman life, which Cato opposed. In the years around 170 a new ruthlessness marked Roman foreign policy, and this seems to have been largely due to the influence of the Fulvian group, including Marcius Philippus and Popillius Laenas, who were then in a controlling position in Roman politics.

Some of the personal antagonism of leading members of the great families was shown in the opposition of Fabius Maximus to Scipio Africanus, and it assumed great virulence early in the 2nd century. There was an attempt to deny Acilius Glabrio, consul of 191 (a new man and a protégé of the Scipios), a triumph for his great success in Greece, and soon afterward there came an attack on the integrity of Africanus and his brother Lucius in their dealings with Antiochus. In consequence, both Acilius and Lucius failed to be elected to the censorship, and Africanus spent his last years at Liternum in Campania in a kind of banishment. Similar attacks were launched against Manlius Vulso after his campaign in Asia Minor in 189. This kind of personal hostility could harm Roman interests if a commander in the field was prepared to make peace with an enemy on disadvantageous terms rather than allow his successor the credit of victory, or if a commander tried to overturn a peace so as to continue a war under his own command: in 201 the consul Gnaeus Cornelius Lentulus tried, without success, to overturn the peace with Carthage made after Zama.

Thanks to its cool-headed direction of the Second Punic War, the power of the Senate increased insensibly. In the alarm caused by the exposure of the Bacchanalian conspiracy in 186 (see below Religion), nobody objected to the Senate's behaviour in empowering the consuls to put even citizens to death, though by law a citizen could only be executed after trial and condemnation by the People. Similar behaviour by the Senate in the case of Tiberius Gracchus' supporters in 132 was to be exposed as unconstitutional by Gaius Gracchus nine years later.

The Senate's role in war

The declaration of war and the making of peace were still the formal responsibility of the People, as was shown particularly in the declaration of the First Punic and Second Macedonian wars; but the administrative arrangements that followed peace were in the Senate's hands. If the Senate decided that a conquered land should be made a Roman province, the Senate itself decided the general lines on which the new province should be organized and, as has been seen, sent out ten of its members to collaborate with the victorious general on the spot in giving detailed effect to its general resolution, as well as to issue the charter of the province.

Popular assemblies. The oldest of Rome's effective assemblies, the Comitia Centuriata, was not a democratic body. Of the 193 voting units (after the reform of 241), 70 belonged to the highest of the five property classes and 18 to the aristocratic cavalry (equites equo publico), so that the rich came near to having an absolute majority of votes. A junior century of the top property class (centuria praerogativa) voted first; and its vote, when announced, influenced the voting of the other centuries. The legislative Comitia Tributa, which also elected the quaestors, was more democratic because the voting unit was the tribe; but in the 3rd century BC few members of the 31 country tribes, apart from the rich, could spare the time to come to Rome to vote, so the result was the same. During the 2nd century, however, a big change took place; the population of Rome grew rapidly on account of the drift to the city of the ruined peasantry, who had lost their farms through the spread of large grazing estates (latifundia). These immigrants remained members of their original country tribes, so that by the middle of the 2nd century the Comitia Tributa (and likewise the Concilium Plebis) became representative bodies of the city dwellers of Rome itself.

In the matter of legislation, the Comitia Centuriata and the Comitia Tributa could vote only on measures that had already been approved by the Senate. The Concilium Plebis, on the other hand, under the presidency of the tribunes, had received in 287 the right to legislate for the whole Roman people without consultation of the Senate; and the way seemed open to extreme democracy. This development was arrested by the invasion of Pyrrhus (280) and the succession of great wars. A land bill of Gaius Flaminius in 232, carried against the wishes of the Senate, was a rarity. Normally, if conservative interests were threatened by a revolutionary, or even a reforming, tribune, it was not difficult to persuade one of his nine colleagues, with promises of support in his political future, to veto a proposal. This was the simple means by which Tiberius Gracchus' opponents expected to kill his land bill in 133; and the bill would have failed if Tiberius Gracchus had not been prepared to act in violation of the Roman constitution by persuading the plebs to unseat the dissenting Octavius.

As long as there was an open system of voting, without ballots, the rich and powerful, acting through their agents, were able by intimidation to control the voting of their clients. There was, therefore, strong conservative opposition to the introduction of voting by ballot, with its attendant secrecy, for elections in 139 (Lex Gabinia Tabellaria) and for legislation in 131 (Lex Papiria Tabellaria).

By very early Roman law, constantly re-enacted, no Roman citizen could be executed except after condemnation by the People; and so the People in the Comitia constituted the supreme criminal court of Rome. Impeachment could be by a tribune or an aedile; this popular function was supplemented, and in the end replaced, by the creation of permanent criminal courts, sitting under a praetor, on the model of the quaestio *repetundarum*, the permanent court to handle extortion cases, established by a bill of the tribune Calpurnius Piso in 149.

There was another field in which the People were not altogether voiceless or ineffective. It was at a meeting of the Comitia Tributa under the presidency of tribunes that, despite his youth and inexperience, young Publius Cornelius Scipio was sent to command in Spain in 210; and it was as a result of popular outcry in favour of disregarding legal requirements concerning the minimum age for the consulship that even the Comitia Centuriata in 148 decided that Scipio Aemilianus, a candidate for the aedileship, should instead be elected consul and sent to conduct the war against Carthage. Thus there were 3rd-

and 2nd-century precedents for the extraordinary popular commands of the last decades of the republic.

Economy. By the Lex Claudia of 218, senators and their sons were effectively precluded from trading by being forbidden to own vessels larger than those necessary for shipping produce from their estates to Rome; but it would be a mistake to think of them as utterly distinct from the business and trading class, for they might have relatives who were in business, and even the senators themselves might carry on enterprises in the name of one of their freedmen. Nonetheless, the Senate seemed to show a complete lack of concern with imperial commercial expansion, as evidenced by the closure of the Macedonian mines in 167 and the destruction 21 years later of two of the finest harbours in the Mediterranean—Corinth and Carthage.

Private enterprise

At the same time there were large business interests operated by a class of men that later developed into the equestrian order. As has been seen, during this period many of the great roads of Italy were constructed; and, indeed, the first of the great imperial roads—the Via Egnatia from Dyrrhachium and Apollonia to Thessalonica—was built not long after 146. Tenders for such work were submitted to the censors who, after scrutiny, allotted the contracts. In Spain the state soon abandoned its first attempt to manage the mines and handed the management over to private enterprise, the contractors paying a rent to the state. There were contracts for the arming and equipping of the Roman armies, which must have required particularly large quantities of leather. After Cannae the companies (societates) of contractors agreed to defer payment on condition of being fully insured against losses by storm or war.

The greatest economic change in Italy was the rapid development (in the early 2nd century) of ranching as a major industry, particularly in Campania and south Italy, which had been devastated and depopulated as a result of Hannibal's occupation. Rich men extended their holdings by encroaching on public land (ager publicus), partly by buying up at a low price the small farms of those who, largely through extended periods of military service (particularly in Spain), were no longer able to work their holdings at a profit. Thus, large estates (*saltus*, pascua) came into existence, partly in hill country for summer grazing, partly in the plains for winter grazing. In the south, sheep were raised, farther north, oxen, the hides of which produced the leather that was so badly needed.

These enterprises were assisted by the ready availability of cheap slaves, captives from the wars in the East. The slaves were tough men whose labour was ruthlessly exploited; they were often chained, and they slept in slave prisons (*ergastula*). Large numbers were required—to fight off wild beasts and to prevent cattle rustling.

These great grazing estates, the latifundia that, in the elder Pliny's famous expression, "were the destruction of Italy," had nothing in common with the medium-sized and large arable farms, on the economy and management of which the elder Cato wrote his De agricultura. Cato, indeed, discussing means of making a fortune, dismissed commerce as risky and moneylending as being disreputable in itself and said that more money could be made by ownership of even a poor ranch than by agriculture.

Demographically, the swelling of the population of the big towns, particularly Rome, by immigration of dispossessed small farmers was a striking feature of the first half of the 2nd century and was the movement which Tiberius Gracchus later (133) tried to reverse. In the country towns the immigrants made what money they could as casual labourers, and it is doubtful that there was much unemployment among the expanding population of Rome. Grain production in the vicinity of Rome and the large coastal towns of Italy dropped heavily as a result of the import of tax grain from Sicily and Sardinia; but in the north agriculture flourished in the rich plain of the Po Valley (Gallia Cisalpina).

The rapid growth of luxury imports and also of the number of slaves from the eastern Mediterranean brought Roman and Italian businessmen in large numbers to the East, particularly to Delos (made a free port in 166), where a large Roman and Italian business community came into existence; many thousands of persons a day are said to have passed through the Delos slave market. There were numbers of Roman settlers in the western provinces, too. Wounded Roman soldiers were left behind at Italica in Spain by Scipio Africanus in the Second Punic War. In 171 about 6,000 sons of Roman soldiers who had married Spanish women established the first Latin colony outside Italy, at Carteia, near Gibraltar.

Culture. From the 5th century onward Rome had received mention by Greek historians; after its defeat of Pyrrhus, when Ptolemy II made diplomatic overtures to the city, its formidable potential was evidently recognized in the Hellenistic world, and soon after that Timaeus wrote at length of Italy and Rome in his Sicilian history. By 217 mainland Greece was frighteningly aware of "the cloud in the West"—the victor, whichever it should be, of the Second Punic War. Both of the belligerents were "barbarians" in the technical Greek sense of the word; Rome's very name—"rhōmē"—was a Greek word and meant "brute force."

Hellenizing influences on Rome

In sharp contrast was the impact the Greek world made on the Romans, first in Campania in the 4th century, then in Sicily and Greece. There were magnificent temples, centuries old, permanent stone theatres, public baths (hot baths at that), and luxurious private houses. There were performances of classical and modern tragedy and comedy in the theatres. There were exercises and athletics in the palaestrae of the baths and at public festivals, with young men competing naked—"Greek games," by which Romans were always a little shocked. There was the Greek language, infinitely more delicate and flexible than contemporary Latin; there was the treasure-house of Greek literature, from Homer onward; and there was the deeply embedded Greek interest in philosophy, in argument, in questioning the truth of what, to the Romans, were fundamentals—traditions of religion and government. Rome had to come to terms with this newly discovered dynamic and to determine if it was to enrich Roman life or to serve as a catalyst.

Language. The children of upper class Roman families now learned to speak Greek—often from a Greek slave nurse—at a very early age; and Greek literature was taught, either in school or by a family tutor, a Greek freedman or slave. As a result, the educated Roman of the 2nd century BC was bilingual and was fluent in both languages. It has been suggested that in the family of the younger Scipio, Greek, not Latin, was the language of general conversation.

Literature. Greek literature offered a challenge that was accepted. The Roman poet and dramatist Lucius Livius Andronicus (c. 284–c. 204) published a Latin version of Homer's Odyssey; Gnaeus Naevius (*c.* 270–c. 199) wrote a Latin epic on the First Punic War, set against Rome's early background; and Quintus Ennius (239–169), a Calabrian, wrote an epic history of Rome to 171 BC—the Annales. Roman drama made its debut as an adaptation of Greek, though it sought to be distinctively Roman. Livius Andronicus was the first playwright. Naevius, Marcus Pacuvius (c. 220–*c.* 130), who came from Brundisium, and Pacuvius' uncle Ennius were prolific authors of tragedies. Ennius and Gaius Lucilius (*c.* 180–103/102), respectively, gave a start to the distinctively Roman genres of affable and of hard-biting satire. There were comedies, too—Roman adaptations of Greek Middle Comedy but with a distinctively Roman veneer—in this genre Plautus (born c. 254), an Umbrian, and Terence (c. 186/185–159), who was of Libyan stock, were preeminent (21 comedies of Plautus survive, six of Terence). Plays were acted at Roman festivals on improvised platforms in front of temples, with temporary seats erected for the audience. (There was no permanent stone theatre at Rome until 55 BC.)

Historiography. While, with the exception of Lucilius, Rome's earliest literary figures were men of obscure origins, the first writing of history by Romans was done by politicians (the earliest of them, Quintus Fabius Pictor, flourished at the end of the 3rd century). And so a lasting Roman tradition was established that historical writing was a field for men with experience in public affairs. Based on briefly annotated lists of magistrates preserved by the pontiffs and on the vainglorious traditions of the great families, early Roman history before the invasion of Pyrrhus was largely fictional. The first histories were written in Greek, partly to counter hostile accounts of Rome by contemporary Greeks, including accounts of the Punic Wars written from the Carthaginian standpoint. Cato's history (*Origines*), however, was written in Latin and set the fashion for the future.

Religion. Greek religion both startled and frightened the Romans; unlike Roman religion, it was anthropomorphic, and the Greek world was full of statues—often of the highest art—of Greek gods and goddesses in the image of men and women. At the same time, contemporary Greek skepticism shocked the fundamental piety of the Romans. On the instruction of the Oracle at Delphi, a black stone from Pessinus in Asia Minor was sent to Rome in 205–204 and, worshipped as the Great Mother, became an object of Roman cult, though with Oriental, not Roman, priests. New orgiastic cults seeped in; and in 186 there was panic on the discovery of the existence of a widely diffused cult of Bacchus, which seemingly involved not only unspeakable immorality but also a plot to overthrow the established government. A witch hunt and a spate of executions followed before the cult was transformed to one of closely supervised respectability, not only in Rome but throughout Italy.

Dangers to the traditional life

Philosophical currents. The firm stability of Roman traditional life was endangered, conservatives thought, not only by such orgiastic cults and by the increasing popularity among the wealthy of luxurious homes (with hot baths) but also by the acute logic of Greek philosophers who used reason to destroy belief and by the persuasive arts of Greek rhetoricians. In Rome they shocked the old, but they fascinated the young. Thus, when three distinguished Greek philosophers arrived in Rome in 155 to urge the return of the surviving Achaean hostages and conducted widely attended public discussions, Cato had them bundled out of the city. Greek ideas of social equality were utterly un-Roman—ideas by which, it would be claimed, Tiberius Gracchus was corrupted. On the other hand, Stoic ethics were supremely adapted to suit the best Roman mores.

Two schools of thought existed from the start. Both Scipios were patrons of the new writers—Scipio Aemilianus had such Greeks as Polybius and the Rhodian philosopher and historian Panaetius as his friends; Cato opposed them, himself tutoring his son rather than employing a Greek tutor. But Cato was not a fool; he learned Greek as did everybody else. Yet, morbidly fascinated by the spectre of their own decline, later Romans accepted the commonplace that Eastern conquest had weakened the tough Roman fibre, in particular because it introduced new and extravagant tastes.

Law. Only in law did Greece exercise little influence on Rome. Based on the Twelve Tables that schoolboys had to learn by heart, Roman law was amplified by successive laws and senatorial resolutions (jus civile), by the interpretations of successive praetors (jus *honorarium*), and by the teaching of the jurists—the earliest of them Tiberius Coruncanius, consul of 280 BC and the first plebeian pontifer *maximus* ("chief priest"). (J.P.V.D.B.)

III. The late republic (133–31 BC)

The fall of Carthage and Corinth did not mark even a temporary end to warfare. For a few years the interminable wars in Spain dragged on—against the Lusitanian leader Viriathus (who had rallied that tribe and won some important victories) in the south and against the Celtiberi based on Numantia in the centre of the peninsula. The sorry tale of Roman defeats, treaties concluded and broken, perfidy, and genocide continued to mark these wars as no others in republican history. But Roman methods finally triumphed. Viriathus was assassinated, and resistance in the south soon ended. The Numantines forced a Roman consular army under Gaius Hostilius Mancinus to surrender in order to escape destruction, and Mancinus promised them peace (137 BC)—an act that inevitably was disavowed by the Senate. Finally, in 134 Publius Cornelius Scipio Aemilianus—again elected consul—enrolled an army consisting chiefly of private clients from all over the Roman world (thus avoiding the internal troubles caused by major levies for the unpopular Spanish fiont) and finished the war with the capture and destruction of Numantia, whose inhabitants were killed or enslaved (133). The long wars were over at last, and although the pictuie of peaceful development in Spain must not be exaggerated (five Spanish triumphs are recorded in the 50 years between the time of Scipio and that of Sulla), Spanish wars were no longer a running sore in Roman political life.

End of the wars in Spain

THE AFTERMATH OF THE VICTORIES

War and military glory were an essential part of the Roman aristociatic ethos and, hence, of Roman political life. Apart from major wars still to come, small wars on the frontiers of Roman power—never precisely fixed—continue to provide an essential motive in Roman history: in Spain, Sardinia, Illyria, and Macedonia, barbarians could be defeated and triumphs won. Thus the limits of Roman power were gradually extended, and the territories within them pacified, while men of noble stock rivalled the virtus (courage and qualities of leadership) of their ancestors, and new men staked their own competing claims, winning glory essential to political advancement and sharing the booty with their officers and soldiers. Cicero could still depict it as a major disgrace for Lucius Piso (consul 58 BC) that he had won no triumph in the traditionally "triumphal" province of Macedonia. Nonetheless, the coincidence of the capture of Corinth and Carthage was even in antiquity regarded as a turning point in Roman history: it was the end (for the time being) of warfare against civilized powers, in which the danger was felt to be greater and the glory and the booty were out of proportion to those won against barbarian tribes.

Provincial administration. The first immediate effect was on the administration of the empire. The military basis of provincial administration remained: the governor (as he is called) was in Roman eyes a commander with absolute and unappealable powers (*imperium*) over all except Roman citizens, within the limits (his provincia) assigned to him (normally) by the Senate. He was always prepared—and in some provinces expected—to fight and win. But it had been found that those unlimited powers were often abused and that Senate control could not easily be asserted at increasing distances from Rome. For political and perhaps for moral reasons, excessive abuse without hope of a remedy could not be permitted. Hence, when the decision to annex Carthage and Macedonia had been taken in principle (149 BC), a permanent court (the quaestio repetundarum) was established at Rome to hear complaints against former commanders and, where necessary, to assure repayment of illegal exactions. No penalty for offenders was provided, and there was no derogation from the commander's powers during his tenure; nevertheless, the step was a landmark in the recognition of imperial responsibility, and it was also to have important effects on Roman politics. Another result of the new conquests was a major administrative departure. When Africa and Macedonia became *provinciae* to be regularly assigned to commanders, it was decided to break with precedent by not increasing the number of senior magistrates (praetors). Instead, prorogation—the device of leaving a magistrate in office pro *magistratu* ("in place of a magistrate") after his term had expired, which had hitherto been freely used when emergencies led to shortages of regular commanders—was established as part of the administrative system: thenceforth, every year at least two praetors would have to be retained as promagistrates. This was the beginning of that dissociation between urban magistracy and foreign command that was to become a cardinal principle of the system of Sulla and of the developed Roman Empire.

Social and economic ills. It is not clear to what extent the temporary end of the age of major wars helped to produce the crisis of the Roman Republic. The general view of thinking Romans was that the relaxation of external pressures led to internal disintegration. This has happened in other states and the view is not to be lightly dismissed. Moreover, the end of large-scale booty led to economic recession in Rome, thus intensifying poverty and discontent. But the underlying crisis had been building up over a long period. In view of the inadequacy of the source material and the still limited results of archaeological research, its nature cannot be confidently defined. But some facts are certain. Large tracts of southern Italy had been devastated in the Hannibalic war and, partly for ecological reasons, never recovered. Little successful resettlement took place on the land confiscated there by Rome from rebellious allies after victory, and much of it was turned over to grazing, chiefly by transhumance methods, hence chiefly exploited by wealthy Romans. Nearer Rome, imports of provincial wheat destroyed local wheat growing, with the result that peasants were driven off the land and flocked into the city, while their holdings were bought up by wealthy neighbours and often consolidated into medium-sized mixed farms. It was for this class of men that Cato wrote his book De agricultura. Farther north, causes and effects alike were less catastrophic; but long wars overseas and rigorous conscription led to the decline of peasant holdings and the accumulation of land. A population drift to the cities (both Rome and the country towns) is noticeable by mid-century, with the towns apparently growing and prospering and, no doubt, attracting peasants, who, by selling out and moving there, would cease to be landowners and liable to the draft. Their divorce from the land, however, was not absolute: seasonal work in agriculture always helped to provide a livelihood for former peasants and their families.

Intensified poverty and discontent

Rich landowners probably put tenants in the place of the peasants they had bought out; if progressive, they might change over to Cato's kind of mixed farming, based on slave labour and supplemented by seasonal free labour as required. Rich men owned many medium-sized estates, rounded off by occupation of public land; but there is little evidence of consolidated plantation-type estates such as were found in some provinces and later in Italy as well. Archaeological evidence, even quite near Rome, predominantly shows small- and medium-scale farming. This, however, might well go with the substitution of slaves and tenants (coloni) for free owners and the drift to the towns on the part of the latter.

THE REFORM MOVEMENT OF THE GRACCHI (133–121 BC)

From the state's point of view, the chief effect was a decline in military manpower. The minimum property qualification for service was lowered, the minimum age (17) ignored; resistance became frequent, especially to the distant and unending guerrilla war in Spain.

The program and career of Tiberius Sempronius Gracchus. Tiberius Gracchus, grandson of Scipio Africanus and son of the Gracchus who had conquered the Celtiberi and treated them well, was quaestor in Mancinus' army when it faced annihilation; on the strength of his family name, he personally negotiated the peace that saved it. When the Senate—on the motion of his cousin Scipio Aemilianus, who later finished the war—renounced the peace, Tiberius felt aggrieved and joined a group of senior senators hostile to Aemilianus and with ideas on reform. Elected tribune for 133, in Scipio's absence, Tiberius attempted to find a solution for the social and military crisis, with the political credit to go to himself and his backers. Tiberius had no intention of touching private property; his idea was to enforce the legal but widely ignored limit of about 300 acres (120 hectares) on occupation of public land and to use the land thus retrieved for settling landless citizens, who would both regain a secure living and be liable for service. A slave war in Sicily, lasting several years and threatening to spread to Italy, had underlined both the danger of using large numbers of slaves on the land and the need for a major increase in military citizen manpower.

Plan to settle landless citizens

Tiberius' proposal was bound to meet with opposition in the Senate, which consisted of large landowners. On the advice of his eminent backers, he took his bill—which made various concessions to those asked to obey the law and hand back excess public land—straight to the Assembly of the Plebs, where it found wide support. This was not revolutionary; bills directly concerning the People appear to have been frequently passed in this way. But his opponents persuaded another aristocratic tribune, Marcus Octavius, to veto the bill. Tiberius tried the constitutional riposte: an appeal to the Senate for arbitration. But the Senate was unwilling to help, and Octavius was unwilling to negotiate over his veto—an action apparently unprecedented, though not (strictly speaking) unconstitutional. Tiberius had to improvise a way out of the impasse. He met Octavius' action with a similarly unprecedented retort, and had Octavius deposed by the Assembly. He then passed his bill in a less conciliatory form and had himself, his father-in-law, and his brother appointed commissioners with powers to determine boundaries of public land, confiscate excess acreage, and divide it in inalienable allotments among landless citizens. As it happened, envoys from Pergamum had arrived to inform the Senate that Attalus III had died and made the Roman People his heirs (provided the cities of his kingdom were left free). Tiberius, at whose house the envoys were lodging, anticipated Senate debate and had the inheritance accepted by the People and the money used to finance his agrarian schemes.

Tiberius' opponents now charged him with aiming at tyranny, a charge that many may well have believed: redistribution of land was connected with demagogic tyranny in Hellenistic states, and Tiberius' subsequent actions had been high-handed and beyond the flexible borderline of what was regarded as mos *majorum* (constitutional custom). Fearing prosecution after he left office, he now began to canvass for a second tribunate—another unprecedented act, bound to reinforce fears of tyranny. The elections took place in an atmosphere of violence, with nearly all his tribunician colleagues now opposed to him. When the consul Publius Scaevola, on strict legal grounds, refused to act against him, Publius Scipio Nasica, the chief pontiff, led a number of senators and their clients to the Assembly, and Tiberius was killed in a resulting scuffle. Widespread and bloody repression followed in 132. Thus political murder and political martyrdom were introduced into Roman politics.

Introduction of political murder and martyrdom into Roman politics

The land commission, however, was allowed to continue because it could not easily be stopped. Some evidence of its activities survives. By **129,** perhaps running out of available land held by citizens, it began to apply the Gracchan law to public land held by Italian individuals or communities. This had probably not been envisaged by Tiberius, just as he did not include noncitizens among the beneficiaries of distributions. The Senate, on the motion of Scipio Aemilianus, upheld the Italians' protests, transferring decisions concerning Italian-held land from the commission to a consul. This seriously hampered the commission's activities. Marcus Fulvius Flaccus, chairman of the commission and consul in **125,** tried to solve the problem by offering the Italians the citizenship (or alternatively the right to appeal against Roman executive acts to the Roman People) in return for bringing their holdings of public land under the Gracchan law. This aroused fears of uncontrollable political repercussions. Flaccus was ordered by the Senate to fight a war in southern France (where he gained a triumph) and had to abandon his proposal. There is no sign of widespread Italian interest in it at this time, though the revolt of the Latin colony Fregellae (destroyed **125)** may be connected with its failure.

The program and career of Gaius Sempronius Gracchus. In 123 Gaius Gracchus, a younger brother of Tiberius, became tribune. He had served on Tiberius' land commission and had supported Flaccus' plan. Making the most of his martyred brother's name, Gaius embarked on a scheme of general reform in which, for the first time in Rome, Greek theoretical influences may be traced. Among many reforms—including provision for a stable and cheap wheat price and for the foundation of colonies (one on the site of Carthage), to which Italians were admitted—two major ideas stand out: to increase public revenues (both from the empire and from taxes) and pass the benefit on to the people; and to raise the wealthiest nonsenators (particularly the equites, holders of the "public horse"—who received state financial aid for the purchase and upkeep of their horses—and next to senators in social standing) to a position from which, without actually taking part in the process of government, they could watch over senatorial administration and make it more responsible. The idea was evoked by Tiberius' death. As early as 129 a law compelled senators to surrender the "public horse" (which hitherto they had also held) and possibly in other ways enhanced the class-consciousness and privileges of the equites. Gaius put the *publicani* (public contractors, hitherto chiefly concerned with army and building contracts and with farming minor taxes) in charge of the main tax of Asia—a rich province formed out of Attalus' inheritance, which would henceforth provide Rome with the major part of its public revenues. This was expected both to reduce senatorial corruption and to improve efficiency. Gaius also put eminent nonsenators (probably defined by wealth, but perhaps limited to the equites, or equestrian class) in charge of the *quaestio repetundarum*, whose senatorial members had shown too much leniency to their colleagues, and he imposed severe penalties on senators convicted by that court. Finally, in a second tribunate, he hoped to give citizenship to Latins and Latin rights to other Italians, with the help of Flaccus who, though a distinguished former consul, took the unique step of becoming tribune. But a consul and a tribune of 122 combined to persuade the citizen voters that it was against their interests to share the privileges of citizenship: the bill was defeated, and Gaius failed in his attempt to be re-elected once more. In 121, preparing (as private citizens) to use force to oppose the cancellation of some of their laws, Gaius and Flaccus were killed in a riot, and many of their followers were executed.

During the next decade the measures benefitting the people were largely abolished, though the Gracchan land distributions, converted into private property, did temporarily strengthen the Roman citizen peasantry. The provisions giving power to wealthy nonsenators could not be touched, for political reasons, and they survived as the chief effect of Gaius' tribunates. The court seems to have worked better than before, and, during the next generation, several other standing criminal courts were instituted, as were occasional ad hoc tribunals, always with the same class of jurors. In 106 a law adding senators to the juries was passed, but it remained in force for only a short time.

THE REPUBLIC (C. 121–91 BC)

Struggles of the aristocratic families

War against Jugurtha. Since Roman historians were no more interested in politics than (on the whole) in social or economic developments, the struggles of the aristocratic families must be pieced together from chance information. It would be mere paradox to deny the importance in republican Rome, as in better known aristocratic republics, of family feuds, alliances, and policies, and parts of the picture are known: *e.g.*, the central importance of the family of the Metelli, prominent in politics for a generation after the Gracchi and dominant for part of that time. In foreign affairs the client kingdom of Numidia—loyal ever since its institution by Scipio Africanus—assumed quite unwarranted importance when a succession crisis developed there soon after 120, as a bastard, Jugurtha, relying on superior ability and aristocratic Roman connections, sought to oust his two legitimate brothers from their shares of the divided kingdom. Rome's usual diplomatic methods failed to stop Jugurtha, who disposed of his brothers, but the massacre of Italian settlers at Cirta by his soldiers forced the Senate to declare war (112). The war was waged reluctantly and ineffectively, with the result that charges of bribery were freely bandied about by demagogic tribunes taking advantage of suspicion of aristocratic political behaviour that had smouldered ever since the Gracchan crisis. Significantly, some eminent men, hated from those days, were now convicted of conuption. The Metelli, however, emerged unscathed, and Quintus Metellus, consul in 109, was entrusted with the war in Africa. He waged it with obvious competence but failed to finish it and thus gave Gaius Marius, a senior officer, his chance.

The career of Gaius Marius. Marius, born of an equestrian family at Arpinum, had attracted the attention of Scipio Aemilianus as a young soldier and, by shrewd political opportunism, had risen to the praetorship and married into the patrician family of the Julii Caesares. Though he had deeply offended the Metelli, once his patrons, Marius' military talents were considerable, and Quintus Metellus had taken him to Africa as a *legatus.* Marius intrigued against his commander in order to gain a consulship; he was elected (chiefly with the help of the equites and anti-aristocratic tribunes) for 107 and was given charge of the war by special vote of the People. He did little better than Metellus had, but in 105 his quaestor Lucius Sulla, in delicate and dangerous negotiations, brought about the capture of Jugurtha, opportunely winning the war for Marius and Rome.

Roman sphere of influence in southern France

During the preceding decade a serious threat to Italy had developed in the north. Starting in 125, several Roman commanders (Marcus Flaccus has been noted) had fought against Ligurian and Gallic tribes in southern France and had finally established a Roman sphere of influence there: a road had been built linking Italy with Spain, and some garrison posts probably secured it; finally, a colony was settled at Narbonne, one of the most important road junctions in the region (*c.* 118). But unwilling to extend administrative responsibilities, the Senate had refused to establish a regular *urovincia.* Then some migrating German tribes, chief of them the Cimbri, after defeating a Roman consul, invaded southern France, attracting native sympathy and finding little effective Roman opposition. Two more consular armies suffered defeat, and in October 105 a consul and proconsul with their forces were destroyed at Orange. There was panic in Rome, allayed only by the firm action of the other consul, Publius Rutilius Rufus.

At this moment news of Marius' success in Africa arrived, and he was at once dispensed from legal restrictions and again elected consul for 104. After a brilliant triumph, restoring Roman morale, he took over the army prepared and trained by Rutilius and was reelected consul year after year, while the German tribes delayed attacking Italy. Finally, in 102–101, he annihilated them at Aquae Sextiae (Aix-les-Bains) and, with his colleague Quintus Catulus, on the Campi Raudii (near the Po Delta). Another triumph and a sixth consulship (in 101) were his reward.

In his first consulship, Marius had taken a step of great (and probably unrecognized) importance: aware of the difficulties long endemic in the traditional system of recruitment, he had ignored property qualifications in enrolling his army and, as a result, had recruited ample volunteers from among men who had nothing to lose. This radical solution was thenceforth generally imitated, and conscription became confined to emergencies (like the Social and Civil wars). He also enhanced the importance of the legionary eagle (standard), thus beginning the process that led to each legion's having a continuing corporate identity. At the same time, Rutilius introduced arms drill and reformed the selection of senior officers. Various tactical reforms in due course led to the increasing prominence of the cohort (one tenth of a legion) as a tactical unit and the total reliance on non-Roman auxiliaries for light-armed and cavalry service. The precise development of these reforms cannot be traced, but they culminated in the much more effective armies of Pompey and Caesar.

Marius' African army had been unwilling to engage in another war, and Marius preferred to use newly levied soldiers (no longer difficult to find). But neither he nor the Senate seemed aware of any responsibilities to the veterans. In 103 a tribune, Lucius Saturninus, offered to pass a law providing land in Africa for them in return for Marius' support for some anti-oligarchic activities of his own. Marius agreed, and the large lots distributed to his veterans (both Roman and Italian) turned out to be the beginning of the Romanization of Africa. In 100, with the German wars ended, Saturninus again proved a welcome ally, arranging for the settlement of Marius' veterans in Gaul. An incidental effect was the departure of Marius' old commander and subsequent enemy, Quintus Metellus, who refused to recognize the validity of Saturninus' law and, choosing martyrdom, went into exile. But this time Saturninus exacted a high price. With his ally, the praetor Gaius Glaucia, he introduced laws to gain the favour of plebs and equites and proceeded to provide for the settlement of veterans of wars in Macedonia and Sicily in the same way as for those of Marius' war. He planned to seek re-election for 99, with Glaucia illegally gaining the consulship. Violence and even murder were freely used to accomplish these aims.

Marius' opposition to Saturninus and Glaucia

Marius now had to make a choice. Saturninus and Glaucia might secure him the continuing favour of the plebs and perhaps the equites, though they might also steal it for themselves. But as the saviour of his country, six times consul, he now hoped to become an elder statesman (*princeps*), accepted and honoured by those who had once looked down on him as an upstart. To this end he had long laboured, dealing out favours to aristocrats who might make useful allies. This was the only accepted reward for achievement; Marius never thought of revolution or tyranny. Hence, when called on to save the state from his revolutionary allies, he could not refuse. He imprisoned them and their armed adherents and did not prevent their being lynched. Yet, having saved the oligarchy from revolution, he received little reward; he lost the favour of the plebs, while the oligarchs, in view of both his birth and his earlier unscrupulous ambition, refused to accept him as their equal. Metellus was recalled; this was a bitter blow to Marius' prestige, and he preferred to leave Rome and visit Asia.

Before long a face-saving compromise was found, and Marius returned; but in the 90s he played no major part. Though he held his own when his friends and clients were attacked in the courts, his old aristocratic protégés now found more promising allies. Sulla is typical: closely associated with Marius in his early career, he was by 91 ready to take the lead in attacking Marius and (significantly) found eager support. The oligarchy could not forgive Marius.

WARS AND DICTATORSHIP (C. 91–80 BC)

Events in Asia. In foreign affairs the 90s were dominated by Asia, Rome's chief source of income. Mithradates VI, king of Pontus, had built a large empire around the Black Sea and was probing and intriguing in the Roman sphere of influence. Marius had met him and had given him a firm warning, temporarily effective: Mithradates had proper respect for Roman power. Scheming to annex Cappadocia, he had been thwarted by the Senate's instructing Sulla, as proconsul, to install a pro-Roman king there in 96–95. (It was on this occasion that Sulla received a Parthian embassy—the first contact between the two powers.) But dissatisfaction in the Roman province gave new hope to Mithradates. Ineffectively organized after annexation and corrupt in its cities' internal administration, Asia was soon overrun with Italian businessmen and Roman tax collectors. When the Senate realized the danger, it sent its most distinguished jurist, Quintus Mucius Scaevola (consul in 95 and pontifex *maxi*mus), on an unprecedented mission to reorganize Asia (94). He took Publius Rutilius Rufus—jurist, Stoic philosopher, and former consul—with him as his senior officer, and after Scaevola's return Rutilius remained behind, firmly applying the new principles. This caused an outcry from businessmen, whose profits he had kept within bounds: he was prosecuted for "extortion" in 92 and convicted after a trial in which Roman publicani and businessmen unscrupulously used their power among the class that provided criminal juries. The verdict revealed the breakdown of Gaius Gracchus' system: the class he had raised to watch over the Senate now held irresponsible power, making orderly administration impossible and endangering the empire. Various leading senators were at once vexatiously prosecuted, and political chaos threatened.

Developments in Italy. The 90s also saw dangerous developments in Italy. In the 2nd century BC, Italians as a whole had shown little desire for Roman citizenship and had been remarkably submissive under exploitation and ill-treatment. The most active of their governing class flourished in overseas business, and the more traditionally minded were content to have their oligarchic rule supported by Rome. Their admission to citizenship had been proposed as a by-product of the Gracchan reforms. By 122, it had become clear that the Roman people agreed with the oligarchy in rejecting it. The sacrifices demanded of Italy in the Numidian and German wars probably increased dissatisfaction among Italians with their patent inferiority. Marius gave citizenship to some as a reward for military distinction—illegally, but his standing (*auctoritas*) sufficed to defend his actions. Saturninus admitted Italians to veteran settlements and tried to gain citizenship for some by full admission to Roman colonies. The censors of 97–96, aristocrats connected with Marius, shared his ideas and freely placed eminent Italians on the citizen registers. This might have allayed dissatisfaction, but the consuls of 95 passed a law purging the rolls and providing penalties for those guilty of fraudulent arrogation. The result was insecurity and danger for many leading Italians. By 92 there was talk of violence and conspiracy among desperate men.

Insecurity and danger for leading Italians

It was in these circumstances that the eminent young noble, Marcus Livius Drusus, became tribune for 91 and hoped to solve the menacing accumulation of problems by means of a major scheme of reforms. He attracted the support of the poor by agrarian and colonial legislation and tried to have all Italians admitted to citizenship and to solve the jury problem by a compromise: the courts would be transferred to the Senate, and 300 equites would be admitted to it. (To cope with the increase in business it would need this increase in size.) Some leading senators, frightened at the dangerous situation that had developed, gave weighty support. Had Drusus succeeded, the poor and the Italians might have been satisfied; the equites, deprived of their most ambitious element by promotion, might have acquiesced; and the Senate, always governed by the prestige of the noble principes rather than by votes and divisions, could have returned, little changed by the infusion of new blood, to its leading position in the process of government. But Drusus failed. Some members of each class affected were more conscious of the loss than of the gain; and an active consul, Lucius Philippus, provided leadership for their disparate opposition. After much violence, Drusus' laws were declared invalid. Finally he himself was assassinated. The Italians now rose in revolt (the Social War), and in Rome a special tribunal, manned by the Gracchan jury class, convicted many of Drusus' supporters until the Senate succeeded in suspending its sittings because of the military danger.

The Social War

The first year of the Social War (90) was dangerous: the tribes of central and southern Italy, traditionally among the best soldiers in Rome's wars, organized in a confederacy for the struggle that had been forced upon them. Fortunately all but one of the Latin cities—related to Rome by blood and tradition and specially favoured by Roman law—remained loyal: their governing class had for some time had the privilege of automatically acquiring Roman citizenship by holding local office. Moreover, Rome now showed its old ability to act quickly and wisely in emergencies: the consul Lucius Caesar passed a law giving citizenship to all Italians who wanted it. The measure came in time to head off major revolts in Umbria and Etruria, which accepted at once.

Civil war and the rule of Lucius Sulla. In 89 the war in central Italy was won, and Gnaeus Pompeius Strabo celebrated a triumph. Attention now turned to the East, where Mithradates had taken advantage of Rome's troubles to expel the kings of Cappadocia and Bithynia. A Roman embassy restored them, and he withdrew; but, when the envoys incited Bithynian incursions into his territory, Mithradates launched a major offensive, overrunning the two kingdoms and invading Roman territory, where he attracted the sympathy of the natives by executing thousands of Italians and defeated and captured the Roman commanders in the area.

Sulla's occupation of Rome

Sulla's rise to power. In Rome, various men, including Marius, had hoped for the Eastern command. But it went to Sulla, elected consul for 88 after distinguished service in the Social War. Publius Sulpicius, a tribune in that year and an old friend of Drusus, tried to continue the latter's policy of justice to the Italians by abolishing the gerrymandering that in practice deprived the new citizens of an effective vote. Finding the oligarchy firmly opposed, he gained the support of Marius (who still commanded much loyalty) for his plans by having the Eastern command transferred to him. After much street-fighting, the consuls escaped from Rome, and Sulpicius' bills were passed. Sulla's response was totally unforeseen: appealing to the army he had led in the Social War, still engaged in mopping-up operations in Campania, he persuaded them to march on Rome, which he occupied, executing Sulpicius; Marius and others escaped. Significantly, Sulla's officers left him. It was the first time a private army of citizens had occupied Rome—an effect of Marius' army reform, which had ended by creating a "client army" loyal chiefly to its commander, and of the Social War, which had made the use of force within Italy seem commonplace. The end of the republic was foreshadowed.

Having cowed Rome into acquiescence and passed some legislation, Sulla left for the East. Cinna, one of the consuls of 87, at once called for the overthrow of Sulla's measures. Resisted by his colleague Octavius, he left Rome to collect an army and, with the help of Marius, occupied the city after a siege. Several leading men were killed or condemned to death, Sulla and his supporters were outlawed, and (after Marius' death early in 86) another commander was sent to Asia. The policy now changed to one of reconciliation: the Social War was wound up, and the government gained wide acceptance until Cinna was killed by mutinous soldiers (85).

Dictatorship of Sulla. Sulla meanwhile easily defeated Mithradates' forces in two battles in Boeotia, took Athens, which under a revolutionary regime had declared for Mithradates, and cleared the King's army out of Greece. While negotiating with Cinna's government, Sulla also entered upon negotiations with Mithradates and, when he heard of Cinna's death, quickly made peace and an alliance with Mithradates, driving the government's commander in Asia to suicide. After wintering his troops in the rich cities of Asia, Sulla crossed to Greece and then to Italy, where his veteran army broke all resistance and occupied Rome (82). Sulla was elected dictator and, while Italy and all the provinces except Spain were quickly reduced. began a reign of terror (the "proscriptions"), in which hundreds of his enemies or those of his adherents were killed without trial, while their property went to enrich him and his friends. Wherever he had met resistance in Italy, land was expropriated and given to his soldiers for settlement.

While the terror was going on, Sulla used his powers to put through a comprehensive program of reform (81). Although he had twice taken Rome with a private proletarian army, he had earlier had connections with the inner circles of the oligarchy, and after Cinna's death some eminent men who had refused to collaborate with Cinna joined Sulla. By the time Sulla's success seemed certain, most of those who had collaborated were also on his side, and he was acclaimed as the defender of the nobility who had defeated an illegal revolutionary regime. His reforms aimed chiefly at stabilizing Senate authority by removing alternative centres of power. The tribunate was emasculated; the censors' powers were reduced; provincial governors were subjected to stricter Senate control; and the equites, who had been purged of Sulla's opponents by the proscriptions, were deprived of some symbols of dignity and made leaderless by the inclusion of 300 of Sulla's chief supporters in the Senate. The jury reform of Gaius Gracchus, seen by some leading senators as the prime cause of political disintegration, could now be undone and the criminal courts could once more become a monopoly of senators.

Sulla's attempt to restore stable government

Sulla's measures were by no means merely reactionary. His program was basically that of Marcus Drusus. His overriding aim was the restoration of stable government, and this could only come from the Senate, directed by the *principes* (former consuls and those they chose to consult). He accepted and even extended recent developments where they seemed useful: the Italians retained full citizenship; the system of standing criminal courts was expanded to include six or seven major crimes; the practice of praetors normally spending their year of office in Rome and then going to provinces for a second year was extended to consuls and became an integral part of his system. To prevent long command of armies (which might lead to careers like his own), he increased the number of praetors so that, in principle and in normal circumstances, each province might have a new governor every year. As for the overriding problem of poverty, his contribution to solving it was to settle tens of thousands of his veterans on land confiscated from enemies in Italy; they would be turned into "haves" and would be ready (using their military training) to defend the social order, in which they now had a stake, against "have-riots"--both old and newly created by proscription and confiscation.

The end of Sulla's rule. At the beginning of 80, Sulla laid down his dictatorship and became merely consul, with the senior Metellus (Quintus Metellus Pius), a relative of his wife, as his colleague. The state of emergency was officially ended. At the end of the year, Sulla retired to Campania as a private citizen, after seeing to the election of two reliable consuls, hoping that the restored oligarchy would learn to govern the state he had handed over to them. For 78 Marcus Lepidus, an ambitious patrician whom Sulla disliked and distrusted, was elected consul. Sulla did not intervene. Within a few months, Sulla was dead. Lepidus at once attacked his system, using the grievances of the expropriated as a rallying cry and his province of Gaul as a base. But he was easily defeated by his former colleague Quintus Catulus, assisted by young Gnaeus Pompeius (Pompey).

THE ROMAN STATE IN THE TWO DECADES AFTER SULLA (79–60 BC)

The early career of Pompey. Pompey was the son of Gnaeus Pompeius Strabo, who had triumphed after the Social War but had incurred general hatred because of cold-blooded duplicity during the troubles of 88 and 87. After Strabo's death, young Pompey, who had served under him and inherited his dubiously won wealth, was protected by Cinna's government against his father's enemies. Following in his father's footsteps, he deserted the government after Cinna's death, raised a force among his father's veterans in central Italy, and helped to conquer Italy and, in a lightning campaign, Sicily and the province of Africa for Sulla. Though not old enough to hold any regular magistracy (he was born in 106), he had. from these military bases, blackmailed Sulla into granting him a triumph (81) and had married into the core of the Sullan oligarchy. Out of pique against Sulla, he had supported Lepidus' election for 78, but he had too great a stake in the Sullan system to permit Lepidus to overthrow it.

Revolt of Sertorius

Meanwhile a more serious challenge to the system had arisen in Iberia. Quintus Sertorius, a former praetor of tough Sabine gentry stock, had refused to follow most of his social betters in joining the invader of Italy; he had left for Spain, where he claimed to represent the legitimate government. Although acting throughout as a Roman proconsul, with a "counter-Senate" of eminent Roman citizens, Sertorius won the enthusiastic support of the natives by his fairness, honesty, and charisma, and he soon held most of the lberian Peninsula, defending it successfully even against a large force under Quintus Metellus Pius. The consuls of 77 would have nothing to do with this war, and so Pompey was entrusted by the Senate, through the efforts of his eminent friends and sponsors, with the task of assisting Metellus. The war dragged on for years, with little glory for the Roman commanders. Sertorius had many sympathizers in Italy; but superior numbers and resources finally wore him down, and he was assassinated by a Roman officer. Pompey easily defeated the remnants of his forces in 72.

Meanwhile, the death of Nicomedes IV of Bithynia (74) led to another major war. Like Attalus of Pergamum, Nicomedes left his kingdom to Rome, and this provoked Mithradates, who was in contact with Sertorius and knew of Rome's difficulties, to challenge Rome again. The Eastern command again led to intrigues in Rome, where pressure against the Sullan system had brought a relaxation of the restrictions on the tribunate; amid increasing agitation, the armies in Spain (especially Pompey's) had come to appear as future arbiters of Roman politics. The command finally went to Lucius Lucullus, a relative of Sulla and consul in 74, who hoped to build up a countervailing power in the East.

At the same time, Marcus Antonius, father of the later Triumvir, was given a command against the pirates in the eastern Mediterranean (whom his father had already fought in 102–loo), partly, perhaps, as further re-insurance against Pompey. With Italian manpower heavily committed, a minor slave rising led by Spartacus (73) assumed threatening dimensions, until Marcus Crassus (an old Sullan and profiteer in the proscriptions) volunteered to accept a special command and defeated the slaves. At this point (71), Pornpey arrived back from Spain with his army, crucified the remnants of the slave army, and claimed credit for the victory.

Pompey and Crassus. He and Crassus now confronted each other, each demanding the consulship for 70, though Pompey had held no regular magistracy and was not a senator. Agreeing to join forces, both secured it.

Repeal of the Sullan system. During their consulship, the political, though not the administrative, part of the Sullan settlement was repealed. The tribunes' powers were fully restored, criminal juries were divided between senators and wealthy nonsenators, and, for the first time since Sulla, censors—both supporters of Pompey—were elected, purged the Senate, and in compiling the registers at last fully implemented the Italians' citizenship. The year 70 also saw the prosecution of Verres (son of a "new man" and Sullan profiteer), who had surpassed the very liberal Roman conventions in exploiting his province of Sicily, relying for future impunity on his aristocratic connections (especially the Metelli and their friends), his fortune, and the known corruptibility of the Sullan senatorial juries. But Verres was unlucky. First, he had ill-treated some of Pompey's important Sicilian clients, thus incurring Pompey's displeasure; next, his case coincided with the anti-Sullan reaction of 70; finally, the Sicilians succeeded in persuading Cicero—an ambitious young "new man" from Arpinum, hoping to imitate the success of his fellow citizen Marius by means of his rhetorical ability—to undertake the prosecution. Despite obstruction from Verres' friends, Cicero collected massive evidence against him, presented his case to fit into the political context of the year, and obtained Verres' conviction as an act of expiation for the shortcomings of the Sullan order.

Loss of control by the Sullan establishment

The year 70 thus marks the loss of control by the Sullan establishment. The nobility (families descended from consuls) continued to gain most of the consulships, with the old patriciate (revived by Sulla after a long decline) stronger than for generations; the Senate still supervised administration and took ordinary political decisions; the system continued to rely essentially on *mos majorum* (constitutional custom) and *auctoritas* (prestige)—potent forces in the status society of the Roman Republic. But the solid bases of law and power that Sulla had tried to give it had been surrendered. The demagogue—tribune or consul—could use the legal machinery of the popular assembly (hence such men are called *populares),* while the commander could rely on his army, in the pursuit of private ambition. The situation that Sulla had tried to remedy now returned, made worse by his intervention. His massacres and proscriptions had weeded out the defenders of lawful government, and his rewards had gone to the timeservers and the unscrupulous. The large infusion of equites into the Senate had intensified the effect. While eliminating the serious friction between the two classes, which had made the state ungovernable by 91, it had filled the Senate with men whose tradition was the opposite of that sense of mission and public service that had animated the best of the aristocracy. Few men in the new ruling class saw beyond self-interest and self-indulgence.

One result was that massive bribery and civil disorder in the service of ambition became endemic. Laws were repeatedly passed to stop them, but they remained ineffective because few found it in their interest to enforce them. Exploitation of the provinces did not decrease after Verres: governors (still with unlimited powers) feathered their own nests and were expected to provide for all their friends. Extortion cases became a political ritual, with convictions impossible to obtain. Cicero, thenceforth usually counsel for the defence, presents hair-raising behaviour as commonplace and claims it as acceptable. The Senate's traditional opposition to annexation faded out. Pompey made Syria into a province and added a large part of Pontus to Bithynia (inherited in 74 and occupied in 70); the demagogue Clodius annexed Cyprus—driving its king to suicide—to pay for his massive grain distributions in Rome; Caesar, finally, conquered Gaul by open aggression and genocide and bled it white for the benefit of his friends and his ambitions. Crassus would have done the same with Parthia, had he succeeded. Opposition to all this in the Senate, where it appeared, was based on personal or political antagonism. If the robber barons were attacked on moral grounds, it was because of the use they made of their power in Rome.

Potnpey's extraordinary commands. Politically, the 60s lay under the shadow of Pompey. Refusing to take an ordinary province in 69, he waited for his chance. It came in 67, when his adherent Gabinius, as tribune, secured him, against the opposition of all important men, an extraordinary command, with unprecedented powers, to deal with the pirates. Pompey succeeded within a few months where Antonius and others had failed. The equites and the people were delighted because trade, including Rome's food imports, would now be secure. Meanwhile Lucullus had driven Mithradates out of Asia Minor and into Armenia; but he had offended Roman businessmen by strict control and his own soldiers and officers by strict discipline. Faced with mutinies, he suffered a reverse and became vulnerable to attacks in Rome. In 66 another tribunician law appointed Pompey, fresh from his naval victories, to take over supreme command in the East, which he did at once, studiously insulting his predecessor. He quickly defeated Mithradates and procured his death, then spent some years in a total reorganization of the East, where Asia (the chief source of revenue) was protected by three further provinces and a ring of client states beyond the frontier. The whole of the East now stood in his clientela, and most of it owed him money as well. He returned as by far the wealthiest man in Rome.

Defeat and death of Mithradates VI

Political suspicion and violence. Meanwhile Roman politics had been full of suspicion and violence, much of it stirred up by Crassus who, remembering 71, feared Pompey's return and tried to make his own power impregnable. There was much material for revolution, with poverty (especially in the country, among families dispossessed by Sulla) and debt (among both the poor and the dissolute rich) providing suitable issues for unscrupulous *populares.*

Conspiracy of Catiline. One such man, the patrician Catiline, after twice failing to gain the consulship by traditional bribery and intrigue, put himself at the head of a movement planning a coup d'état in Rome to coincide with an armed rising in Italy (late 63). Cicero, as consul, defeated these efforts and, relying on the doubtful legality of a Senate vote in support, had Catiline's eminent Roman associates executed. Catiline himself fell in a desperate battle.

Isolation of Cicero

For Cicero—the "new man" who had made his way to the top by his own oratorical and political skill, obliging everyone by unstinting service, representing Pompey's interests in Rome while avoiding offense to Pompey's enemies—this was the climax of his life. Like his compatriot Marius, he had saved the state for its rulers: he had taken resolute action when those rulers were weak and vacillating; and, like Marius, he got small thanks for it. Pompey was miffed at having to share his fame with a municipal upstart, and eminent gentlemen could not forgive that upstart for having driven patricians to their death.

Pompey's loss of power. Pompey's return was peaceful. Like Marius, he wanted recognition, not tyranny. He dismissed his army, to the surprise of Crassus and others, and basked in the glory of his triumph and the honours voted to him. But having given up power, he found himself caught in a net of constitutional obstruction woven by his politically experienced enemies and was unable to obtain either of his principal demands: land for his veterans and the ratification of his arrangements in the East. It was at this point that Caesar returned from Spain.

The rise of Caesar. Gaius Julius Caesar, descended (as he insisted) from kings and gods, had shown talent and ambition in his youth: he opposed Sulla, but without inviting punishment; married into the oligarchy, but advocated popular causes; vocally defended Pompey's interests, while aiding Crassus in his intrigues and borrowing a fortune from him; flirted with Catiline, but refused to dabble in revolution, then worked to save those whom Cicero executed. In 63 he won a startling success: defeating two distinguished ***principes,*** he, who had not yet been praetor, was elected ***pontifex*** *maximus* —a post of supreme dignity, power, and patronage. The Roman aristocracy had long ceased to believe in the state religion; but its ceremonial was kept up for popular effect and as a recognized means of political manipulation, so that priesthoods could give more lasting power than magistracies, in addition to the cachet of social success. Young Caesar was now head of the hierarchy. After his praetorship (62), Caesar successfully governed Spain, clearing a surplus sufficient to pay off his debts. On returning to Rome, he naturally hoped for the consulship of 59; but his enemies, by legal chicanery, forced him to choose between standing for office and celebrating a triumph. He gave up the triumph and easily became consul.

THE FINAL COLLAPSE OF THE ROMAN REPUBLIC (59–44 BC)

Alliance of Caesar, Pornpey, and Crassus

Caesar, Pompey, and Crassus. For his consulship Caesar fashioned an improbable alliance: his skill in having won the trust of both Crassus and Pompey enabled him to unite these two enemies in his support. Crassus had the connections, Pompey had the soldiers' vote, and Caesar was consul and ***pontifex*** *maximus.* The combination (often misleadingly called the "first Triumvirate") was invincible, especially since the consul Caesar had no scruples about countering legal obstruction with open force. Pompey got what he wanted, and so did Crassus (whose immediate need was a concession to the Asian tax farmers, in whose companies he probably had much of his capital). In return, Caesar got a special command in Cisalpine Gaul and Illyricum for five years by vote of the People; the Senate itself, on Pompey's motion, extended it to Transalpine Gaul. Marriage alliances sealed the compact, chief of them Pompey's marriage to Caesar's daughter Julia.

Caesar left for Gaul, but Rome was never the same; the shadow of the alliance hung over it, making the old-style politics impossible. In 58 Publius Clodius, another aristocratic demagogue, was tribune and defended Caesar's interests. Cicero had incurred Clodius' enmity and was now sacrificed to him: he was driven into exile as having unlawfully executed citizens in 63. By 57 Caesar's allies had drifted back into rivalry. Pompey secured Cicero's return, and Cicero at once tried to break up the alliance by attracting Pompey to the Senate's side. Just when he seemed about to succeed, the three dynasts secretly met and revived their compact (56). Rome had to bow once more. In 55 Pompey and Crassus were consuls, and the contents of their secret agreement were slowly revealed. Caesar, whom his enemies had made efforts to recall, was prolonged in his command for five years and (it later appeared) had been promised another consulship straight after, to secure him against prosecution and give him a chance of another army command. Pompey was given a special command over all of Spain, which he exercised through deputies while he himself remained just outside Rome to keep an eye on the city. Crassus, who now needed glory and new wealth to equal those of his allies, was to attack Parthia with a large army. Thus the three dynasts would practically monopolize military power for the foreseeable future.

Cicero, among others, had to submit and was thenceforth their loyal spokesman. After his achievement of 63 he had dreamt of leading a coalition of all "right-thinking" men in Italy in defending the traditional oligarchy, but he had found little support among the oligarchy and now used this fact to rationalize his surrender. His brother took service in Gaul under Caesar.

The dynasts' pact did not even bring peace. Clodius, as tribune, had created a private army, and there was no state force to counter it. Pompey could have done it by calling his soldiers in, but the Senate did not trust him enough to request this, and Pompey did not wish to parade himself as an unashamed tyrant. Other men formed private armies in opposition to Clodius, and one Milo at last managed to have him killed after a scuffle (52). By then, however, Roman politics had radically and unexpectedly changed.

End of the pact

Political manoeuvres. Julia died in 54, breaking the ties between Caesar and Pompey. Caesar pressed Pompey to renew them, but Pompey held off, preserving his freedom of action. Crassus' Parthian campaign ended in disaster and in Crassus' death (53). By 52 Pompey and Caesar stood face to face, still nominally friends but with no personal link between them and no common interests. Caesar, by conquering the whole of Gaul, had almost equalled Pompey's prestige and, by his utterly ruthless way of waging war, Pompey's wealth. Unlike Pompey, he used his wealth to dispense patronage and buy useful friends. At this point Pompey cautiously offered the oligarchy his support. It had much to give him that he wanted—control of the administrative machine, respectability, and the seal of public approval. Its leaders (even the intransigent young Cato, who had led opposition to the three individually long before their alliance and to their joint oppression of the state ever since) now recognized that acceptance of Pompey's terms and surrender to his protection was their only chance of survival. Pompey at once turned firmly against Milo, who presented a political threat: if Milo could use the force that had killed Clodius to keep firm control of Rome, he—an ambitious man of known conservative views—might in due course offer an alternative and more trustworthy champion to the oligarchy. But he was not yet ready. Pompey forced them to make their choice at once, and they chose Pompey in preference. He was made sole consul and had Milo convicted by an intimidated court. Meanwhile he had made a marriage alliance with the noblest man in Rome, Quintus Metellus Scipio, who became his colleague in the consulship. The state had captured Pompey (or vice versa), and Caesar stood alone in opposition to both of them. The next two years saw a series of manoeuvres: the Senate leaders, with Pompey's silent support, worked for Caesar's recall, which would have meant his instantly sharing the fate of Milo; while Caesar and his agents in Rome tried to strike some bargain that would ensure his safety and his future in politics. Finally, Pompey declared himself, and, early in 49, the Senate voted to outlaw Caesar. Two tribunes supporting him (one of them Mark Antony) had to flee. By the time they reached him, Caesar had already crossed the Rubicon: he now had a cause.

Civil war. Pompey had exuded confidence over the outcome if it came to war. In fact, however, Caesar's veterans were unbeatable, and both men knew it. To the disgust of his followers, Pompey evacuated Rome, then Italy. His plan was to bottle Caesar up there and starve him out. But Caesar, in a lightning sweep, seized Massilia and Spain from Pompey's commanders, then crossed to Greece, where a short campaign ended in Pompey's decisive defeat at Pharsalus (48). Pompey fled to Egypt, where he was assassinated by a man hoping thus to curry

Caesar's favour. This was by no means the end of the war. Almost at once, Caesar was nearly trapped at Alexandria, where he had intervened in a succession dispute; but he escaped and installed Cleopatra on the throne, for personal as well as political reasons. In Africa the Pompeian forces and their native allies were not defeated until Caesar himself moved against them and annihilated them at Thapsus. Cato, disdaining the victor's pardon, committed suicide at Utica (46). In Spain, where Pompey's name was still powerful, his sons organized a major rising, which Caesar himself again had to defeat at Munda (45) in the bloodiest battle of the war. By the time he returned, he had only a few months to live.

The dictatorship and assassination of Caesar. In Rome the administrative machine had inevitably been disrupted, and Caesar had always remained in control, as consul or as dictator. Those who had feared proscriptions, or hoped for them, were proved wrong. Some of Caesar's enemies had their property confiscated, but it was sold at fair value; most were pardoned and suffered no loss. One of these was Cicero, who, after much soul-searching, had followed his conscience by joining Pompey before Pharsalus. Poverty and indebtedness were alleviated, but there was no wholesale cancellation of debts or redistribution of property, and many of Caesar's adherents were disappointed. Nor was there a general reform of the republic. (Caesar's only major reform was of the calendar: indeed, the Julian calendar proved adequate for centuries.) The number of senators and magistrates was increased, the citizenship was more freely given, the province of Asia was relieved of some of its tax burden. But Caesar had no plan for reforming the system—not even to the extent that Sulla had tried to do, for Sulla had at least planned for his own retirement. For a time, honorable men, such as Cicero, hoped that the "Dictator for settling the Constitution" (as Caesar called himself) would produce a real constitution—some return to free institutions. By late 45 that hope was dead. Caesar was everywhere, doing everything to an almost superhuman degree. He had no solution to the crisis of the republic except to embody it in himself and none at all for the hatred of his peers, which he knew this was causing. He began to accept more and more of the honours that a subservient Senate invidiously offered, until finally he reached a position perilously close to kingship (an accursed term in Rome) and even deification. Whether he passed those hazy boundary lines is much debated and not very important. He had put himself in a position in which no Roman ought to have been and which no Roman aristocrat could tolerate. As a loyal friend of his was later to say: "With all his genius, he saw no way out." To escape the problem or postpone it, he prepared for a Parthian war to avenge Crassus—a project most likely to have ended in similar disaster. Before he could start on it, about 60 men—former friends and old enemies, honorable patriots and men with grievances—struck him down in the Senate on March 15, 44 BC.

Changes
under
Caesar

THE TRIUMVIRATE AND OCTAVIAN'S ACHIEVEMENT OF SOLE POWER

Brutus and Cassius, the organizers of the conspiracy, expected all Romans to rejoice with them in the rebirth of "freedom." But to the Roman people the freedom of the governing class had never meant very much; the armies (especially in the west) were attached to Caesar; and the Senate was full of Caesarians at all levels, cowed but biding their time. Mark Antony, the surviving consul, whom Brutus had been too scrupulous to assassinate with his master, gradually gained control of the city and the official machinery, and the "liberators" withdrew to the East. But a challenger for the position of leader of the Caesarians soon appeared in the person of Octavian, Caesar's son by adoption and now his heir. Though not yet 20, Octavian proved an accomplished politician, attracting loyalty as a Caesarian while cooperating against Antony with the Senate, which, under Cicero's vigorous leadership, now turned against the consul. Cicero hoped to fragment and thus defeat the Caesarian party, with the help of Brutus and Cassius, who were making good progress in seizing control of the eastern provinces and armies. In 43 the two consuls (both old Caesarian officers) and Octavian defeated Antony at Mutina, and success seemed imminent. But the consuls died, and Octavian demanded and, by armed force, obtained the consulship; and the armies of Italy, Spain, and Gaul soon showed that they would not fight against one another. Octavian, Antony, and Lepidus (the senior Caesarian with an army) now had themselves appointed "Triumvirs for settling the Constitution" for five years and secured control of Italy by massive proscriptions and confiscations (Cicero, Antony's chief enemy, was among the first to die). They then defeated and killed Brutus and Cassius at Philippi (42) and divided the Roman world among themselves, with Lepidus, a weak man accidentally thrust into prominence, getting the smallest share. Octavian, who was to control Italy, met armed opposition from Antony's brother and wife, but they got no help from Antony and were defeated at Perusia (41). Octavian and Antony sealed their alliance with a marriage compact: Antony married Octavia, Octavian's sister. Octavian then confronted Pompey's son Sextus Pompeius, who had seized control of the islands off Italy. After much diplomatic manoeuvring (including another meeting with Antony), Octavian attacked and defeated Sextus; when Lepidus tried to reassert himself, Octavian crushed him and stripped him of his office of Triumvir (while with conspicuous piety leaving him the chief pontificate, now powerless). Octavian now controlled the West and Antony the East, still officially as Triumvirs (their term of office had been extended), even though Lepidus had been eliminated in 36.

Formation
of the
Triumvirate

Each of the two leaders embarked on campaigns and reorganization in his half—Octavian in Illyricum, Antony particularly on the Parthian frontier. But Antony now married Cleopatra and tried to make Egypt his military and political base. In a war of propaganda, Octavian gradually convinced the western provinces, Italy, and most of the Roman upper class that Antony was sacrificing Roman interests, and trying to become a Hellenistic king in Alexandria, and planning to rule the Roman world from there with Cleopatra. In 32, though he now held no legal position, Octavian intimidated most of Antony's remaining aristocratic friends into joining him, made the whole West swear allegiance to himself, and in 31, as consul, crossed to Greece to attack Antony. On September 2 he defeated Antony and Cleopatra in a naval battle at Actium. Though in itself not a major victory, it was followed by the disintegration of Antony's forces, and Antony and Cleopatra finally committed suicide in Alexandria (30). (E.Ba.)

IV. The early Roman Empire (31 BC–AD 193)

THE CONSOLIDATION OF THE EMPIRE UNDER THF JULIO-CLAUDIANS

The establishment of the principate under Augustus. Actium left Octavian the master of the Roman world. This supremacy, unremittingly maintained until his death more than 40 years later, made him the first of the Roman emperors. Suicide removed Antony and Cleopatra and their potential menace in 30 BC, and the annexation of Egypt with its Ptolemaic treasure brought financial independence. With these reassurances Octavian could begin the task of reconstruction.

The first
of the
Roman
emperors

Law and order had vanished from the Roman state when its ruling aristocrats refused to curb their individual ambitions and thus transformed a republic based on disciplined liberty into a turbulent cockpit of murderous rivalries. Good government depended on limits being set to unrestrained aspirations, and Octavian was in a position to impose them. But, although his military might in 31 BC could guarantee orderly political processes, it was itself incompatible with them; nor did he relish the role of military despot. The fate of Julius Caesar, **an** eagerness to acquire political respectability, and his own esteem for ancestral custom combined to dissuade Octavian from it. He wished to be, in his own words, "the author of the best civilian government possible." His problem was to regularize his own position so as to make it generally acceptable, without simultaneously reopening the door to violent lawlessness. His pragmatic solution not only ensured stability and continuity, but it also respected republican forms and traditions so far as possible.

Large-scale demobilization allayed men's fears; regular consular elections raised their hopes. In 29–28 BC he carried out, with Agrippa, the first census of the Roman people since 70; and this involved drawing up an electoral roll for the Centuriate Assembly. Elections followed, and Octavian was inevitably chosen consul. Then, on January 13, 27 BC, he offered to lay down his powers. The Roman Senate rejected this proposal, charging him instead to administer (besides Egypt) Spain, Gaul, and Syria for the next 10 years, while it itself supervised the rest of the empire. (It is uncertain whether the Senate named him proconsul for this provincia in 27 or expected him to superintend it by virtue of being consul.) Three days later, among other honours, it bestowed upon him the name by which he has ever since been known, Augustus.

Therewith the Roman state ceased to be the plaything of dictators and triumvirs and became a res publica again. Indeed, the Latin expression for this settlement of 27 is res publica restituta, the usual translation of which, "restoration of the republic," is misleading, because the Latin res publica means a state governed by constitutional methods, whether a republic or not.

Augustus' powers

As most of the troops still under arms were in the regions entrusted to Augustus' charge, the arrangements of 27 BC hardly affected his military strength. Moreover, so long as he was consul (he was re-elected every year until 23 BC) he was civilian head of government as well. In other words, he was still pre-eminent and all-powerful, even if he had, in his own words, placed the res publica at the disposal of the Senate and the Roman People. Augustus particularly wished to conciliate the senatorial class, without whose experienced help civilian government was impossible. But his monopolization of the consulship offended the Senate, so that some different arrangement was clearly necessary. Accordingly, in 23 Augustus made a change; he vacated the consulship and never held it again (except momentarily in 5 BC and again in 2 BC, for a limited, specific purpose). In its place he received the tribunician power (tribunicia potestas). He could not become an actual plebeian tribune, as Julius Caesar's action in making him a patrician had disqualified him for the office. But he could acquire the rights and privileges pertaining to the office; and they were conferred upon him, apparently by the Senate, whose action was then ratified by the popular assembly. He had already been enjoying some of a tribune's privileges since 36; but he now acquired them all and even some extra ones, such as the right to convene the Senate whenever he chose and to enjoy priority in bringing business before it. Through his tribunician power he could also summon the popular assembly and participate fully in its proceedings. Clearly, although no longer consul, he still retained the legal right to authority in civilian affairs.

The arrangement of 23 entailed an additional advantage. The power of the plebeian tribune was traditionally associated with the protection of citizens, and Augustus' acquisition of it was therefore unlikely to rouse resentment. Indeed, Augustus thenceforth shrewdly propagated the notion that, if his position in the state was exceptional (which it clearly was), then this was precisely because of his tribunician power. Although he held it for only one year at a time, it was indefinitely renewable and was pronounced his for life. Thus, it was both annual and perpetual and was a suitable vehicle for numbering the years of his supremacy. His era (and this is true also of later emperors) was counted officially from the year when he acquired the tribunician power.

The year 23 likewise clarified the legal basis for Augustus' control of his provincia (the region under his jurisdiction) and its armed forces. The Senate invested him with an imperium proconsulare (governorship and command of an empire); this had a time limit, but it was automatically renewed whenever it lapsed (usually every ten years). This proconsular imperium, furthermore, was pronounced valid inside Italy, even inside Rome and the *pomerium* (the boundary within which only Roman gods could be worshipped and civil magistrates rule), and superior (majus) to the irnperium of any other proconsul. Thus, Augustus could intervene quite legally in any province, even in one entrusted to someone else.

It is clearly this imperium proconsulare *majus* that made Augustus unassailable; but, because it belonged more to military than to civilian life, he preferred not to draw attention to it. In fact, he de-emphasized it: it is not reflected in either his official titles or his autobiography, the tribunician power being paraded as the expression of his supreme authority.

Augustus' honours

After 23 no fundamental change in Augustus' position occurred. He felt no need to hold offices that in republican times would have conferred exceptional power (*e.g.*, dictatorship, lifetime censorship, or regular consulship), even though these were offered him. Honours, of course, came his way: in 19 BC he received some consular rights and prerogatives, presumably to ensure that his *imperium* was in no particular inferior to a consul's; in 12, when Lepidus died, he became *pontifex maximus* (he had long since been elected into all of the priestly colleges); in 8 BC, the eighth month of the year was named after him; in 2 BC, he was designated pater patriae ("father of his country"), a distinction which he particularly esteemed because it suggested that he was to all Romans what a paterfamilias was to his own household. He also accepted special commissions from time to time: *e.g.*, to superintend the supply of grain and water, the maintenance of public buildings (including temples), the regulation of the Tiber, the police and fire-fighting services, and the upkeep of Italy's roads. Such behaviour advertised his concern for public well-being and also his legal correctness; but it hardly altered the fact that his tribunician power and proconsular *imperium*, as defined in 23, were the real bases of his strength and the twin legal supports upon which his position and authority rested. It was these that made him a charismatic leader of unrivalled prestige (auctoritas), whose merest suggestions were binding; they ensured absolute imperial predominance for the future.

Like an ordinary Roman, he contented himself with three names. His, however, Imperator Caesar Augustus, were absolutely unique, with a magic all their own that caused all later emperors to appropriate them, at first selectively but after AD 69 in their entirety. Thereby they became titles, reserved for the emperor (or, in the case of the name Caesar, for his heir apparent); from them derive the titles emperor, kaiser, and tsar. Yet, as used by Augustus and his first four successors, the words Imperator Caesar Augustus were names, not titles—that is, respectively, praenomen, nomen (in effect), and cognomen. One title that Augustus did have was *princeps* ("prince"); this, however, was unofficial—a mere popular label, meaning Rome's first citizen—and government documents (inscriptions, coins, etc.) do not apply it to Augustus. But because of it the system of government he devised is called the principate.

Augustus' relationship with the Senate

The Roman Senate and the urban magistracies. Augustus regarded the Senate, whose leading member (*princeps senatus*) he had become in 28, as a body with important functions; it heard fewer overseas embassies than formerly, but otherwise its dignity and authority seemed unimpaired; its members filled the highest offices; its decrees, although not formally called laws, were just as binding; it soon became a high court, whose verdicts were unappealable; it supervised the older provinces and nominally the state finances as well, and it also in effect elected the urban magistrates; formally, even the emperor's powers derived from the Senate. Nevertheless, it lacked real power. Its provinces contained few troops (and by AD 40 it had ceased to control even these few). Hence, it could hardly dispute Augustus' wishes. Moreover, it was he, in fact, who superintended state finances. Above all, he controlled its membership, every senator's career depending upon his goodwill. But he valued the Senate as the repository of the true Roman spirit and traditions and as the body representing public opinion. He was considerate toward it, shrewdly anticipated its reactions, and generally avoided contention with it. He regularly kept it informed about his activities; and the imperial council (Concilium Principis), which he consulted on matters of policy, in the manner of a republican magistrate seeking the opinion of his advisory committee, consisted of the consuls, certain other magistrates, and 15 senators—not handpicked by him but chosen by lot every six months.

To rid the Senate of unworthy elements he reduced its numbers by successive reviews to about 600 (from the triumveral 1,000 or more). Sons of senators and men of good repute and substance who had served in the army and the vigintivirate (a minor magistracy) could become members by being elected, at age 25 or over, to the quaestorship. Their subsequent rank in the Senate depended on what other magistracies they managed to win; these were, in ascending order, the aedileship (or plebeian tribunate), the praetorship, and the consulship. No one disliked by Augustus could expect to reach any of them, while anyone whom he nominated or endorsed was sure of election. Despite the Emperor's control, there were enough candidates for keen contests. By AD 5 destinatio seems to have been the practice: that is, a special panel of senators and equites selected the praetors and consuls, and the Centuriate Assembly automatically ratified their choice. In c. AD 5, likewise, the consulship was shortened to six months. This not only gratified senators and increased the number of high-ranking qualified officials but also showed that the consuls' duties were becoming largely ceremonial. This was also true, but to a far lesser degree, of the other unpaid magistrates. A senator really made his mark in between his magistracies, when he served in important salaried posts, military or civilian or both, sometimes far from Rome.

Rise of equites to a privileged status

The equestrian order. Senators, however, were either too proud or too few to fill all the posts. Some posts were considered menial and went to the Emperor's freedmen or slaves. Others were entrusted to equites, and the equestrian order soon developed into one of the great institutions of the empire. Augustus decided that membership in the order should be open to Roman citizens of means and reputation but not necessarily of good birth. Ultimately, there were thousands of equites throughout the empire. Although this was a lower aristocracy, a good career was available to them. After tours of duty as an army officer (the so-called *militiae* equestres), an aspiring eques might serve as the emperor's agent (procurator) in various capacities and eventually become one of the powerful prefects (of the fleet, of the vigiles, or fire brigade, of the grain supply, of Egypt, or of the Praetorian Guard). This kind of an equestrian career became standardized only under Claudius I; but Augustus began the system and, by his use of equites in responsible posts, founded the imperial civil service, which later was staffed chiefly by them. The equites also performed another function: the senatorial order had difficulty in maintaining its numbers from its own ranks and depended on recruitment from below; this meant from the equestrian order, and, as it was not confined to Rome or even to Italy, the Senate gradually acquired a non-Italian element. The western provinces were already supplying senators under Augustus.

Administration of Rome and Italy. Ordinary Roman citizens who were neither senators nor equites were of lesser consequence. Although still used, the old formula Senatus Populusque Romanus had changed its meaning: in effect, its Populusque Romanus portion now meant "the emperor." The "Roman People" had become the "Italian People," and it was embodied in the person of Augustus, himself the native of an Italian town. To reduce the risk of popular demonstrations in Rome, the Emperor provided grain doles, occasional donatives, and various entertainments; but he allowed the populace no real power. After AD 5 the Roman People's participation in public life consisted in the formality of holding occasional assemblies to ratify decisions made elsewhere. Ultimately, this caused the distinction between the Roman citizens of Italy and the provincial inhabitants of the overseas empire to disappear; under Augustus, however, the primacy of Italy was insistently emphasized.

The primacy of Italy

Indeed, Italy and justice for its inhabitants were Augustus' first cares. Arbitrary triumviral legislation was pronounced invalid after 29 BC, and ordinary Roman citizens everywhere had access to Augustus' own court of appeal (his appellate jurisdiction dated from 30 BC and in effect replaced the republican appeal to the People). His praetorian and urban cohorts provided physical security; his officials assured grain supplies; and he himself, with help from such aides as Agrippa, monumentalized Italian towns. The numerous Augustan structures in Italy and Rome (a city of brick before his time and of marble afterward) have mostly perished, but impressive ruins survive (*e.g.*, aqueduct, forum, and mausoleum in Rome; bridge at Narni; arch at Fano; gate at Perugia). Doubtless their construction alleviated unemployment, especially among the proletariat at Rome. But economic considerations did not influence Augustus' policies much (customs tariffs, for instance, were for fiscal, not protective, purposes), nor did he build harbour works at Ostia. Italian commerce and industry nevertheless flourished in the conditions he created. Public finances, mints, and coinage issues, chaotic before him, were placed on a sound basis, partly by the introduction of a sales tax and of a new levy (death duties) on Roman citizens—who hitherto had been subject only to harbour dues and manumission (freeing of slaves) charges—and partly by means of repeated subventions to the public treasury (aerarium *Saturni*) from Augustus' own enormous private resources (*patrimonium* Caesaris). His many highways also contributed to Italy's economic betterment.

Augustus' great achievement in Italy, however, was to restore morale and unify the country. The violence and self-aggrandizement of the 1st century BC had bred apathy and corruption. To reawaken a sense of responsibility, especially in official and administrative circles, Augustus reaffirmed traditional Italian virtues (by outlawing adultery, strengthening family ties, and stimulating the birthrate) and revived ancestral religion (by repairing temples, building new shrines, and reactivating moribund cults and rituals). To infuse fresh blood and energy into disillusioned Roman society, he promoted the assimilation of Italy: the flower of its municipal towns entered the Roman Senate, and Italy became firmly one with Rome. To keep the citizen body pure, he made manumission of slaves difficult, and from those irregularly manumitted he withheld the citizenship.

Two types of provinces

Administration of the provinces. Sharply distinguished from Italy were the provinces of the empire. From 27 BC on they were of two types. The Senate supervised the long-established ones, the so-called public provinces: their governors were chosen by lot, usually served for a year, commanded no troops, and were called proconsuls (although only those superintending Asia and Africa were in fact former consuls, the others being former praetors). The Emperor supervised all other provinces, and collectively they made up his provincia: he appointed their governors, and these served at his pleasure, none with the title of proconsul because in his own *provincia* proconsular *imperium* was wielded by him alone. These imperial provinces might be "unarmed," but many of them were garrisoned, some quite heavily. Those containing more than one legion were entrusted to former consuls and those with a legion or less to former praetors; in both cases their governors were called *legati Augusti* pro praetore ("legates of Augustus with authority of a praetor"). 'There were also some imperial provinces governed not by senators but by equites (usually styled procurators but sometimes prefects); Judaea at the time of Christ's crucifixion was such an equestrian province, Pontius Pilate being its governor. An entirely exceptional imperial province was Egypt, so jealously guarded that no senator could visit it without express permission; its prefect was unique in being an equestrian in command of legions.

Taxation. The provinces paid tribute, which helped to pay for the armed services, the growing civil service, and the public-works programs. Periodical censuses, carefully listing provincial resources, provided the basis for the two direct taxes: *tributum* soli, exacted from occupiers of provincial soil, and *tributum capitis*, paid on other forms of property (it was not a poll tax, except in Egypt and in certain backward areas). In addition, the provinces paid indirect taxes, such as harbour dues. In imperial provinces the direct taxes (tributa) were paid to the Emperor's procurator, an equestrian official largely independent of the governor. In senatorial provinces, quaestors supervised the finances; but, increasingly, imperial procurators also appeared. The indirect taxes (*vectigalia*) were still collected by publicani, but they were now much more rigorously controlled and were being gradually replaced by imperial civil servants.

Autonomous communities

Development of urban life. As under the republic, each province was a collection of autonomous communities (civitates), roughly resembling modern counties, and it was through them that Rome governed its empire; without them provincial administration would have simply broken down. The autonomous communities, apart from a privileged few, all paid taxes to Rome. Most of them were peregrine; that is, their burgesses were not Roman citizens; but these non-Roman communities tended to decrease in number because they gradually adopted Rome's language and ways and thereby became qualified to receive "Latin rights" (*Latium*, or *jus Latii*)—a status that prepared them for the eventual bestowal of full Roman citizenship. Peregrine communities upgraded in this way were called municipia.

Whole communities with Roman citizenship in the provinces were virtually a new development, because there had been hardly any before Julius Caesar's dictatorship. Those called coloniae (see below) ranked highest, being regarded as actual extensions of Rome itself: some of them were even exempted from payment of tribute. There were, however, few coloniae and practically no *municipia* in the East.

In both senatorial and imperial provinces the civitates enjoyed local self-government. The Roman and Latin communities were city-states, resembling the towns of Italy; they regularly had an assembly of adult males and a local council (*ordo*), and their chief officials—collegial, annual, and unpaid—were either duoviri (in coloniae) or *quattuorviri* (in municipia). The peregrine communities might or might not be city-states—many were organized on tribal, not municipal, lines; moreover, the city-states among them showed great variety, some of the Greek-speaking ones having constitutions that were centuries old. Although Rome regularly ensured control by the propertied classes, it rarely imposed Roman-style constitutions. It did perhaps favour urbanized communities, rural areas being subordinated to city-states for administrative purposes until ready for municipal institutions of their own. Naturally, too, peregrine communities tended to imitate nearby Roman towns, so that even some tribally organized civitntes boasted duoviri and an *ordo*.

The provinces were generally better off under the empire. The governing power now felt some responsibility foı the governed and sent out salaried and experienced officials. These were closely supervised. During the first two centuries, provincial communities flourished on the whole; in many, a sense of civic liberty and healthy local patriotism induced wealthy burgesses to build baths, theatres, and the like at their own expense. Above all, there was peace, the Pax Romana.

Emperor worship. For this priceless gift of peace, many individuals and even whole communities, in Italy and elsewhere, expressed their thanks spontaneously by worshipping Augustus and his family. Emperor worship was also encouraged officially, however, as a focus of common loyalty for the polyglot empire. In the provinces, to emphasize the superiority of Italy, the official cult was dedicated to Roma et Augustus; to celebrate it, representatives from provincial communities or groups of communities met in an assembly (Concilium Provinciae), which incidentally might air grievances as well as satisfactions. (This system began in the Greek-speaking provinces, long used to wooing their rulers with divine honours. It penetrated the west only slowly; but from 12 BC an assembly for the three imperial Gallic provinces existed at Lugdunum.) In Italy the official cult was to the Genius Augusti (the life spirit of his family); it was coupled in Rome with the Lares Cornpitales (the spirits of his ancestors). Its principal custodians (seviri *Augus*tales) were normally freedmen. Calculation always entered into emperor worship; it was no act of piety but a political gesture, which could be continued after the emperor's death if the Senate pronounced him worthy of posthumous deification.

Allegiance of the army to the Emperor

The army. It was Augustus' soldiers, however, not his worshippers, that made him all-powerful. Their allegiance, like the name Caesar, was inherited from his "father," the deified Julius. The allegiance was to the Emperor personally, through a military oath taken in his name every January 1; and the soldiers owed it after his death to his son or chosen successor. This preference of theirs for legitimacy could not be ignored, because they were now a standing army, something that the republic had lacked. Demobilization reduced the 60 legions of Actium to 28, a number hardly sufficient but all that Augustus' prudence or economy would countenance. These became permanent formations, each with its own number and name; the soldiers serving in them were called legionaries. Besides the legionaries there were auxiliaries (or supporting troops), approximately as numerous. The two corps together numbered about 300,000 men. To them must be added the garrison of Italy—the praetorian cohorts, or emperor's bodyguard, about 10,000 strong—and the marines of the imperial fleet, which had its main headquarters at Misenum and Ravenna in Italy and subsidiary stations and flotillas on seas and rivers elsewhere (the marines, however, were not reckoned good combat forces). All these troops were long-service professionals—the praetorians serving 16 years; legionaries, 20; auxiliaries, *25;* and marines, 28—with differing pay scales, the praetorians' being highest. In addition to their pay, the men received donatives, shares of booty, and retirement bonuses from a special treasury (*aerarium militare*) established in AD 6 and maintained out of the sales tax and Roman citizens' death duties. Under Augustus the praetorians were normally Italians, but many legionaries and virtually all auxiliaries were provincials, mainly from the imperial provinces in the west, the legionaries coming from municipal towns and the auxiliaries from tribal areas. The tendency to use provincials grew, and by the year 100 the Roman imperial army was overwhelmingly non-Italian. Nevertheless, it helped greatly to Romanize the empire. The legionaires were Roman citizens from the day they enlisted, if not before, and the auxiliaries (after Claudius anyway) from the day they were discharged; and, though serving soldiers could not legally marry, many had mistresses whose children often became Roman citizens. The troops, other than praetorians and marines, passed their years of service in the "armed" imperial provinces—the auxiliaries in forts near the frontier and the legionaires at some distance from it in camps that showed an increasing tendency, especially after AD 69, to become permanent (some of them, indeed, developed into great European cities). There was no central reserve, because, although desirable for emergencies, it might prove dangerous in peacetime.

The officers were naturally Roman citizens. In the legions those of the highest rank (legati and tribuni) were senators or equites; lower officers (*centuriones*) might enter directly from Italian or provincial municipalities or might rise through the ranks, and by the time they retired, if not sooner, many of them were equites. In the auxiliaries the unit commanders (pmefecti) were equites, often of provincial birth. On retirement the soldiers frequently settled in the provinces where they had served, made friends, and perhaps acquired families; imperial policy favoured this practice. Indeed, many emperors founded colonies of veterans in the provinces, establishing the men in Roman communities of regular municipal type with free grants of land. Some of Augustus' *coloniae* were in Italy; but many more were overseas, as those of his successors regularly were. These military coloniae were the most highly regarded type of urban community under the empire, and excellent Romanizing agents.

Territorial expansion under Augustus

Foreign policy. After Actium and on two other occasions, Augustus solemnly closed the gates of the shrine of Janus (a gesture of peace), to show that Rome had peace as well as a princeps. These well-publicized gestures were purely temporary; the gates were swiftly reopened. His proconsular *imperium* made Augustus the arbiter of peace and war, and an ostensible search for defensible frontiers made his a very warlike reign. He sought to repair the omission of the republic that had left the limits of Roman territorial claims rather vague and indefinite; he planned conquests stretching to the boundaries defined by nature (deserts, rivers, and ocean shores), not always, however, with immediate annexation in mind. When annexation did occur, it was followed by the construction of solidly built military roads, paved with thick stone blocks: these also served the official post system (*cursus publicus*) and were provided with rest stages and overnight lodges at regular inververals.

Client kingdoms. Areas where subjugation looked arduous and where Romanization seemed problematic were left to client kings, dependent on the Emperor's support and goodwill and under obligation to render military aid to Rome. Such satellite kingdoms spared Augustus the trouble and expense of maintaining strong defenses everywhere; nevertheless, their ultimate and intended destiny was incorporation as soon as it suited their overlord's convenience. Usually, territory was gained more easily by creating and subsequently incorporating a client kingdom than by launching an expansionist war.

Southern frontier. In the south, Augustus found suitable frontiers quickly. In 25 BC an expedition under Aelius Gallus opened the Red Sea to Roman use and simultaneously revealed the Arabian Desert as an unsurpassed and, indeed, unsurpassable boundary. The same year Gaius Petronius, the prefect of Egypt, tightened Rome's grip as far as the First Cataract and established a broad military zone beyond it. The vast region north of the Sahara and the Atlas Mountains was also secured (c. *25)* after a series of punitive raids against native tribes and the annexation of one client kingdom (Numidia) and the creation of another (Mauretania). Three legions, two in Egypt and one in Africa (a senatorial province), policed the southern shore of the Mediterranean.

Western frontier. In the west, consolidation was extended to the Atlantic. Gaul, Julius Caesar's conquest, was organized as four provinces: senatorial Narbonensis and the imperial three Gauls (Aquitania, Belgica, and Lugdunensis). In Spain, after Agrippa successfully ended in 19 BC the terminal campaign that Augustus had launched in person in 26, three provinces were formed: senatorial Baetica and imperial Lusitania and Tarraconensis. Three legions enforced Roman authority from Gibraltar to the mouth of the Rhine. Augustus ignored the advice of court poets and others to advance still farther and annex Britain.

Difficulties in the east and north

Eastern frontier. In the east, Parthia had demonstrated its power against Crassus and Antony, and Augustus proceeded warily. He retained Antony's ring of buffer client kingdoms, although he incorporated some, including the most celebrated of them, Judaea; he made Judaea a province in AD 6, respecting some of the customs of its Jewish inhabitants in the process. Augustus stationed four legions in Syria and obviously envisaged the Euphrates River and the northern extension of the Arabian Desert as the desirable frontier with Mesopotamia. Farther north, however, no such natural line existed. North of the Black Sea the client kingdom of the Cimmerian Bosporus, under its successive rulers Asander and Polemo, helped to contain southward and westward thrusts by the Scythians, an Iranian people related to the Parthians, and this provided protection in the north for Asia Minor and its provinces (senatorial Asia and Bithynia-Pontus and imperial Cilicia and Galatia, the latter a large new province created in 25 BC out of Amyntas' client kingdom). By a show of force, Augustus' stepson Tiberius in 20 recovered the standards lost at Carrhae and installed Tigranes as client king of Armenia, and Augustan propaganda depicted this as a famous victory; but strategic considerations inevitably obliged the Parthians, once they settled their internal, dynastic dissensions, to dispute Roman control of Armenia, so that Augustus hardly settled the eastern frontier. Missions were sent to the East repeatedly (Agrippa, 17–13 BC; Gaius Caesar, AD 1–4; Germanicus, 18–19), and Armenia remained a problem for Augustus' successors: Tiberius successfully maintained Roman influence there, but Gaius and Claudius failed to do so, leaving Nero with a difficult situation.

Northern frontier. In the north, too, there was difficulty. The Alps and their passes were finally subjugated early in Augustus' reign. This enabled Tiberius and his brother Drusus between 16 and 8 BC to conquer all the way to the great rivers of central Europe. New provinces were created in the Alps and Tyrol (Maritime and Pennine Alps, Raetia, Noricum) and also farther east (Pannonia, Moesia). Stability along the Danube was precariously maintained, under Augustus and later, by means of periodical alliances with Maroboduus and his successors, who ruled Germanic tribes like the Marcomanni and Quadi in Bohemia to the north of the river, and by the existence of a Thracian client kingdom to the south of its lowest course. The push across the Rhine began in 12 BC; it reached the Elbe, but consolidation beyond the Rhine proved elusive. A revolt in Pannonia (AD 6–9) interrupted it, and, in AD 9, German tribes under Arminius annihilated Quinctilius Varus and three legions in the Teutoburg Forest. This disaster reduced the number of legions to 25 (it did not reach 28 again until half a century later), and it disheartened Augustus. Old and weary, he withdrew to the Rhine and decided against all further expansion, a policy he urged upon his successor. For the watch on the Rhine the military districts of Upper and Lower Germany were created, containing eight legions between them. Another seven garrisoned the Danubian provinces. These figures reveal imperial anxiety for the northern frontier.

Effects of peace

Economic life. Although widespread, Augustus' wars chiefly affected the frontier districts. Elsewhere, peace prevailed. Indeed, never before had so large an area been free of war for so long. This state of affairs helped trade. The suppression of piracy and the use of military roads, which the frontier warfare itself brought into being, provided safe arteries of commerce. Stable currency also aided economic growth. Activity directly connected with the soil predominated; but there were also many establishments, usually small, engaged in manufacturing, and such products as textiles, pottery, tiles, and papyrus were turned out in surprising quantities. Advanced techniques were also known: glassblowing, for example, dates from the Augustan age. Most products were consumed locally, but the specialties or monopolies from any region usually exceeded its needs, and the surplus was sold elsewhere, generating a brisk interchange of goods. Some travelled great distances, even beyond the empire: trade with India, for example, reached respectable proportions once the nature of the monsoon was understood and the Red Sea was opened to Roman shipping. Merchants, especially Levantines, travelled everywhere, and fairs were frequent. The Mediterranean world was linked together as never before, and standardization made considerable headway. In Augustus' day Italy was economically the most important part of the empire. It could afford to import on a large scale, thanks partly to provincial tribute but above all to its own large productivity. The eastern provinces, for their part, recovered rapidly from the depredations of the civil wars and were industrially quite advanced. The other provinces were less developed, but they soon ceased being mere suppliers of raw materials; they learned to exploit their natural resources by using new techniques and then began overtaking the more advanced economies of Italy and the Greek-speaking regions. The importance of trade in unifying the empire should not be underestimated.

Augustan art and literature. In 17 BC Rome held Secular Games, a traditional celebration to announce the entry into a new epoch (*saeculum*). New it certainly was, for, although Augustus preserved what he could of republican institutions, he also added much that was his own. His Rome had become very Italian, and this Italian spirit is reflected in the art and literature of his reign. Its greatest writers were native Italians, and, like the ruler whose program they glorified, they used the traditional as the basis for something new. Virgil, Horace, and Livy imitated the writing of classical Greece, but chiefly in outward form, their tone and outlook being un-Hellenic. The glory of Italy and faith in Rome are what inspired Virgil's Georgics and *Aeneid*, Horace's Odes, and the first ten books of Livy's history.

In Augustan art a similar fusion was achieved between the prevailing Attic and Hellenistic models and Italian naturalism. The sculptured portraits on the Ara Pacis Augustae of 9 BC, for all their lifelike quality, are yet in harmony with the classical poise of the figures, and they strike a fresh note: the stately converging processions (Rome's imperial family and magistrates on one side; senators, equites, and citizens on the other) became the prototypes for all later processional reliefs. Augustan painting likewise displays a successful combination of Greek and Roman elements, to judge from the frescoes in the house of Livia on the Palatine. In Augustan architecture Greek features abound, elaborated on basically Italic plans. Unless Agrippa's Pantheon was domed, there is not as yet significant evidence of the soaring vaults and domes that Roman concrete could have already made possible. Building was, however, very active and widespread.

Cultural excellence of the Augustan age

The culture of the age undoubtedly attained a high level of excellence, dominated by the personality of the Emperor and his accomplishments. Imperial art had already reached full development, and this is a matter of no small moment, because Rome's political predominance made the spread of its influence inevitable. The Mediterranean world was soon assuming a Roman aspect, and this is a measure of Augustus' extraordinary achievement. Yet it was an achievement with limitations. His professed aim —to promote stability, peace, security, and prosperity— was irreproachable; but perhaps it was also unexciting. Emphasizing conservatism by precept and his own example, he encouraged the simpler virtues of a less sophisticated age, and his very success made this sedate but rather static outlook fashionable. Men comfortably accepted the routine of his continuing rule, at the cost, however, of some loss of intellectual energy and moral fervour. The great literature, significantly, belongs to the years near Actium, when men's imagination still nursed heady visions of Roman victory and Italian destiny. After the Secular Games the atmosphere became more commonplace and produced the frivolities of Ovid and the pedestrian later books of Livy.

Appraisal of Augustus. Augustus' position as princeps cannot be defined simply. He was neither a Roman king (*rex*) nor a Hellenistic monarch (basileus), nor was he, as the 19th-century German historian Theodor Mommsen thought, a partner with the Senate in a dyarchy. He posed as the first servant of an empire over which the Roman Senate presided, and it would appear that his claim to have accepted no office inconsistent with ancestral custom was literally true. Proconsular *imperium* was a republican institution; and, although tribunician power was not, it contained nothing specifically unrepublican. But, while precedents can be cited for Augustus' various powers, their concentration and tenure were absolutely unparalleled. Under the republic, powers like his would have been distributed among several holders, each serving for a limited period with a colleague. Augustus wielded them all, by himself, simultaneously and without any time limit (in practice, at least). This fact made him an emperor, but it did not necessarily make him a military tyrant.

In discharging both military and civilian functions, Au-.gustus was no different from republican consuls or praetors. Admittedly his military power was overwhelming; but, if he chose not to brandish it, the tone of his reign could remain essentially civilian. Constitutional safeguards were indeed lacking; everything was at the emperor's discretion, and even Augustus passed legislation that made anti-imperial behaviour, real or suspected, treasonable (men were, in fact, executed for conspiracy during his reign). But there had been no constitutional safeguards in the republic, under Sulla, Pompey, the triumvirs, or even Julius Caesar. Augustus' improved police services probably made lower class Romans at least feel safer under him. Augustus had not even deprived the Romans of "liberty," if *libertas* be understood as meaning freedom from arbitrary rule and arbitrary interference.

The principate was in truth something personal. It was what the individual emperor chose to make it, and the relations prevailing between emperor and Senate usually indicated what a reign was like. In Augustus' case they reveal a regime that was outwardly constitutional, generally moderate, and certainly effective. But, as he himself implied at the end of his life, he was a skillful actor in life's comedy. Later emperors might have lacked his sureness of touch.

Augustus' accomplishments

When Augustus died, the Senate unhesitatingly pronounced him divus—the deified one who had restored peace, organized a standing army to defend the frontiers, expanded those frontiers farther than any previous Roman, improved administrative practices everywhere, promoted better standards of public and private behaviour, integrated Rome and Italy, embellished Rome, reconciled the provinces, expedited Romanization, and above all maintained law and order while respecting republican traditions.

Augustus' luck was hardly inferior to his statecraft. Despite indifferent health, he headed the Roman state in one capacity or another for 56 years. This is the longest personal supremacy in European history, and it consolidated the principate so firmly that what might have been an episode became an epoch. At his death there was practically no one left with any personal memory of the republic, and Augustus' wish came true: he had fashioned a lasting as well as a constitutional type of government. The principate endured with only minor changes for something like 200 years.

The other Julio-Claudian emperors

The succession. Augustus' first four successors were all related to him by either birth or marriage; he and they collectively formed the Julio-Claudian dynasty (27 BC–AD 68). There being no constitutional provision for the succession, none of them reached the purple by legal right. The Emperor's position was personal, and, on his death, revival of the republic was in theory possible; in practice, however, it was out of the question-even in Augustus' lifetime, appointment of a senator as city prefect was necessary, to deputize for the Emperor during his absences from Rome. Consequently, Augustus began thinking early about who should follow him. The soldiers' views on legitimacy reinforced his own natural desire to found a dynasty; but he had no son and was therefore obliged to select his successor. Death played havoc with his attempts to do so. His nephew Marcellus, his son-in-law Agrippa, his grandsons Gaius and Lucius (Julia's children by Agrippa), were groomed in turn; but they all predeceased him. Augustus, finally and reluctantly, chose a member of the republican nobility, his stepson Tiberius, a scion of the ultra-aristocratic Claudii. In AD 4 Augustus adopted Tiberius as his son and had tribunician power and probably proconsular imperium as well conferred upon him. This arrangement was confirmed in 13, and when Augustus died the following year, Tiberius automatically became emperor.

Tiberius. Tiberius (ruled 14–37), during whose reign Christ was crucified, was a soldier and administrator of proved capacity but of a reserved and moody temperament that engendered misunderstanding and unpopularity. Slander blamed him for the death in 19 of his nephew and heir apparent, the popular Germanicus; and, when informers (delatores), who functioned at Rome like public prosecutors, charged notables with treason, Tiberius was thought to encourage them. By concentrating the praetorian cohorts in a camp adjoining Rome he increased the soldiers' scope for mischief-making without finding any real security, and in 26 he left Rome permanently for the island of Capreae (Capri), entrusting

Rome to the care of the city prefect. Tiberius heeded the aged Augustus' advice and did not extend the empire. (The annexation of Cappadocia, a client kingdom, represented no departure from Augustan policy.) In general he took his duties seriously; but by administering the empire from Capreae he offended the Senate and was never fully trusted, much less really liked. At his death he was not pronounced *divus.* His great-nephew, Germanicus' son Gaius, succeeded him.

Gaius. Gaius (better known by his nickname, Caligula, meaning Little Boot) ruled from 37 to 41 with the absolutism of an Oriental monarch: his short reign was featured by reckless spending, callous murders, and humiliation of the Senate. Gaius' foreign policy was inept. Projected annexation proved abortive in Britain; it touched off heavy fighting in Mauretania. In Judaea and Alexandria, Gaius' contemptuous disregard of Jewish sentiment provoked near-rebellion. When assassination ended his tyranny, the Senate contemplated restoration of the republic but was obliged by the Praetorian Guard to recognize Claudius, Germanicus' brother and therefore Gaius' uncle, as emperor.

Centralization and expansion under Claudius I

Claudius I. Claudius I (ruled 41–54) went far beyond Augustus and Tiberius in centralizing government administration and, particularly, state finances in the imperial household. His freedmen secretaries consequently acquired great power; they were in effect directors of government bureaus. Claudius himself displayed much interest in the empire overseas; he enlarged it significantly, incorporating client kingdoms (Mauretania in 42; Lycia, 43; Thrace, 46) and, more important, annexing Britain. Conquest of Britain began in 43, Claudius himself participating in the campaign; the southeast was soon overrun, a *colonia* established at Camulodunum (Colchester) and a *municipium* at Verulamium (St. Albans), while Londinium (London) burgeoned into an important entrepôt. Claudius also promoted Romanization, especially in the western provinces, by liberally granting Roman citizenship, by founding *coloniae,* and by inducting provincials directly into the Senate—he became censor in 47 and added the senators he wanted, bestowing appropriate quaestorian or praetorian rank upon them to spare the maturer ones among them the necessity of holding junior magistracies; lest existing senators take offense, he elevated some of them to patrician status (a form of patronage often used by later emperors). Claudius' provincial policies made the primacy of Italy less pronounced, although that was hardly his aim. In fact, he did much for Italy, improving its harbours, roads, and municipal administration and draining its marshy districts. The execution of many senators and equites, the insolence and venality of his freedmen, the excessive influence of his wives, and even his bodily infirmities combined to make him unpopular. Nevertheless, when he died (murdered probably by his fourth wife, Agrippina, Augustus' great-granddaughter, who was impatient for the succession of the 16-year-old Nero, her son by an earlier marriage), he was pronounced *divus.*

Nero. Nero (ruled 54–68) left administration to capable advisers for a few years but then asserted himself as a vicious despot. He murdered successively his stepbrother Britannicus, his mother Agrippina, his wife Octavia, and his tutor Seneca. He also executed many Christians, accusing them of starting the great fire of Rome in 64 (this is the first recorded Christian persecution). In Rome his reliance on Oriental favourites and his general misgovernment led to a conspiracy by Gaius Calpurnius Piso (65), but it was suppressed, leading to yet more executions; the victims included the poet Lucan. The empire was not enlarged under this unwarlike emperor, but it witnessed serious disorders. In Britain in 60–61 the rapacity and brutality of Roman officials provoked a furious uprising under Queen Boudicca; thousands were slaughtered, and Camulodunum, Verulamium, and Londinium were destroyed. In the east, a major military effort under Corbulo, Rome's foremost general, was required (62–65) to re-establish Roman prestige; a compromise settlement was reached, the Romans accepting the Parthian nominee in Armenia and the Parthians recognizing him as Rome's client king. In 66, however, revolt flared in Judaea, fired by Roman cruelty and stupidity, Jewish fanaticism, and communal hatreds; the Prefect of Egypt, Julius Alexander, prevented involvement of the Jews of the Diaspora. A strong army was sent to Judaea under Titus Flavius Vespasianus to restore order; but it had not completed its task when two provincial governors in the west rebelled against Nero—Julius Vindex in Gallia Lugdunensis and Sulpicius Galba in Hispania Tarraconensis. When the praetorians in Rome also renounced their allegiance, Nero lost his nerve and committed suicide. He brought the Julio-Claudian dynasty to an ignominious end by being the first emperor to suffer *damnatio memoriae—his* reign was officially stricken from the record by order of the Senate.

The end of the Julio-Claudian dynasty

GROWTH OF THE EMPIRE UNDER THE FLAVIANS AND ANTONINES

The year of the four emperors. Nero's death ushered in the so-called year of the four emperors. The extinction of the Julio-Claudian imperial house robbed the soldiers of a focus for their allegiance, and civil war between the different armies ensued. The army of Upper Germany, after crushing Vindex, urged its commander, Verginius Rufus, to seize the purple for himself. But he elected to support Galba—scion of a republican patrician family claiming descent from Jupiter and Pasiphae—who was recognized as emperor by the Senate. A treasury emptied by Nero's extravagance entailed economy, and this bred unpopularity for Galba; his age (73) was also against him, and unrest grew. Early in January 69 the Rhineland armies acclaimed Aulus Vitellius, commander in Lower Germany; at Rome the praetorians preferred Marcus Salvius Otho, whom Galba had alienated by choosing a descendant of the old republican aristocracy for his successor. Otho promptly procured Galba's murder and obtained senatorial recognition; this ended the monopoly of the purple that the republican nobility had hitherto enjoyed.

Otho, however, lasted only three months; defeated at Bedriacum, near Cremona in northern Italy, by Vitellius' powerful Rhineland army, he committed suicide (April 69). The Senate thereupon recognized Vitellius; but the soldiers along the Danube and in the east supported Vespasianus, the commander in Judaea. In a second battle near Bedriacum, Vitellius' troops were defeated in their turn, and on his death soon afterward an accommodating Senate pronounced Vespasian emperor.

The Flavian emperors. On December 22, 69, the Senate conferred all the imperial powers upon Vespasian en *bloc* with the famous Lex de Imperio Vespasiani, and the Assembly ratified the Senate's action. This apparently was the first time that such a law was passed; a fragmentary copy of it is preserved on the Capitol in Rome.

Vespasian. Vespasian (ruled 69–79) did not originate from Rome or its aristocracy. His family came from Sabine Reate, and with his elevation the Italian bourgeoisie came into its own. He and his two sons, both of whom in turn succeeded him, constituted the Flavian dynasty (69–96). Vespasian faced the same difficult task as Augustus—the restoration of peace and stability. The disorders of 69 had taken troops away from the Rhine and Danube frontiers. Thereupon, the Danubian lands were raided by Sarmatians, a combination of tribes who had overwhelmed and replaced the Scythians, their distant kinsmen, in eastern Europe. The assailants were repelled without undue difficulty, but, because the region between the Rivers Tisa and Danube was now firmly in the hands of the Sarmatian Iazyges, they were an omen for the future.

Developments in the Rhineland were more immediately serious. There in 69 a certain Civilis incited the Batavians serving as auxiliaries in the Roman army to rebel. Gallic tribes joined the movement, and the insurgents boldly overran all but two of the legionary camps along the Rhine. Vespasian sent his relative Petilius Cerealis to deal with the rebels, who, fortunately for Rome, were not united in their aims; and by 70 Cerealis had restored order. That same year Vespasian's elder son, Titus, brought the bloody war in Judaea to its end by besieging, capturing, and destroying Jerusalem.

Vespasian's policies

To rehabilitate the public finances, Vespasian introduced new imposts, including a poll tax on Jews, and practiced stringent economies. With the Senate he was courteous but firm. He allowed it little initiative but used it as a reservoir from which to obtain capable administrators. To that end he assumed the censorship and added senators on a larger scale than Claudius had done, especially from the municipalities of Italy and the western provinces. Already before 69 an aristocracy of service had been born, and the provincialization of the Roman Senate had begun; thereafter this development made rapid headway. Besides the censorship, Vespasian also often held the consulship, usually with Titus as his colleague. His object presumably was to ensure that his own parvenu Flavian house outrank any other. In this he succeeded; the troops especially were ready to accept the Flavians as the new imperial family. On Vespasian's death in 79, Titus, long groomed for the succession, became emperor and immediately had his father deified.

Titus. Titus (ruled 79–81) had a brief reign, marred by disasters (the volcanic eruption that buried Pompeii and Herculaneum and another great fire in Rome); but his attempts to alleviate the suffering and his general openhandedness won him such popularity that he was unhesitatingly deified after his early death.

Dornitian. Domitian (ruled 81–96), Titus' younger brother, had never been formally indicated for the succession; but the praetorians acclaimed him, and the Senate ratified their choice. Throughout his reign Domitian aimed at administrative efficiency, but his methods were very high-handed. For him the Senate existed merely to supply imperial servants. He also used equites extensively, more than any previous emperor. He held the consulship repeatedly, was *censor perpetuus* from 85 on, and demanded other extravagant honours. On the whole, his efficiency promoted the welfare of the empire. Above all, he retained the allegiance of the troops. Although scornful of the Senate's dignity, he insisted on his own and mercilessly punished any act of disrespect, real or fancied, toward himself. He became even more suspicious and ruthless when Saturninus, commander in Upper Germany, attempted rebellion in 89. He crushed Saturninus; executions and confiscations ensued, and *delatores* flourished. The tyranny was particularly dangerous to senators, and it ended only with Domitian's assassination in 96. The Flavian dynasty, like the Julio-Claudian, ended with an emperor whose memory was officially damned.

Military reforms under the Flavians

Military developments and frontiers. The disorders in 69 were the cause of some military reforms. Under the Flavians, auxiliaries usually served far from their native heaths under officers of different nationality from themselves. At the same time, the tasks assigned to them came increasingly to resemble those performed by the legionaries. The latter grew less mobile, as stone camp buildings came to be the rule; and it became common for detachments from a legion *(vexillationes),* rather than the entire legion, to be used for field operations. This new-type army proved its mettle in Britain, where the advance halted by Boudicca's revolt was now resumed. Between 71 and 84 three able governors—Petilius Cerealis, Julius Frontinus, and Julius Agricola, the latter Tacitus' father-in-law—enlarged the province to include Wales and northern England; Agricola even reached the Scottish highlands before Domitian recalled him.

Along the Rhine, weaknesses revealed by Civilis' revolt were repaired. Vespasian crossed the river in 74 and annexed the Agri Decumates, the triangle of land between the sources of the Rhine and the Danube. To consolidate the position, he and Domitian after him penetrated the Neckar Valley and Taunus Mountains, and fortifications began to take shape to the east of the Rhine, a military boundary complete with strongpoints, watchtowers, and, later, a continuous rampart of earthworks and palisades. Once Saturninus' revolt in 89 had been suppressed, Domitian felt the situation along the Rhine sufficiently stable to warrant conversion of the military districts of Upper and Lower Germany into regular provinces and the transfer of some Rhineland troops to the Danube. To the north of this latter river, the Dacians had been organized into a strong kingdom, ruled by Decebalus and centring on modern Romania; in 85 they raided southward across the Danube, and in the next year they defeated the Roman punitive expedition. Domitian restored the situation in 88, but Saturninus' rebellion prevented him from following up his success. Domitian and Decebalus thereupon came to terms: Decebalus was to protect the lower Danube against Sarmatian attack, and Domitian was to pay him an annual subsidy in recompense. The Danubian frontier, however, remained disturbed, and Domitian wisely strengthened its garrisons; by the end of his reign it contained nine legions, as against the Rhineland's six, and Pannonia was soon to become the military centre of gravity of the empire.

Strengthening of the eastern frontier

The Flavians also took measures to strengthen the eastern frontier. In Asia Minor, Vespasian created a large "armed" province by amalgamating Cappadocia, Lesser Armenia, and Galatia; and the whole area was provided with a network of military roads. South of Asia Minor, Judaea was converted into an "armed" province by getting legionary troops; and two client kingdoms—Commagene and Transjordan—were annexed and added to Syria. Furthermore, the legionary camps seem now to have been established right on the Euphrates at the principal river crossings. This display of military strength kept the empire and Parthia at peace for many years.

The Antonine emperors. Marcus Cocceius Nerva, an elderly senator of some distinction, was the choice of Domitian's assassins for emperor; and the Senate promptly recognized him. The soldiers did so much more reluctantly, however; and, because the year 69 had revealed that emperors no longer needed to be Roman aristocrats and could be chosen elsewhere than at Rome, their attitude imposed caution.

Nerva. Nerva (ruled 96–98) adopted a generally lavish and liberal policy, but it failed to win the soldiers over completely, and he proved unable to save all Domitian's murderers from their vengeance. Unrest subsided only when, overlooking kinsmen of his own, he adopted an outstanding soldier, Marcus Ulpius Trajanus, governor of Upper Germany, as his successo . Nerva himself died a few months later.

Trajan. Trajan (ruled 98–117) was the first and perhaps the only emperor to be adopted by a predecessor totally unrelated to him by either birth or marriage. He was also the first in a series of "good" rulers who succeeded one another by adoption and for most of the 2nd century provided the empire with internal harmony and careful government; they are collectively, if somewhat loosely, called the Antonine emperors. More significantly still, Trajan, a Spaniard, was also the first *princeps* to come from the provinces; with the greater number of provincials now in the Senate, the elevation of one of them, sooner or later, was practically inevitable. Throughout his reign, Trajan generally observed constitutional practices. Mindful of the susceptibilities of the Senate, he regularly consulted and reported to it. Modest in his bearing, he did not claim ostentatious honours such as frequent consulships or numerous imperial salutations, and he mixed easily with senators on terms of cordial friendship. This re-established mutual respect between *princeps* and Senate. Empire and liberty, in Tacitus' words, were reconciled, and the atmosphere of suspicion, intrigue, and terror surrounding the court in Domitian's day disappeared. Trajan endeared himself also to the populace at large with lavish building programs, gladiatorial games, and public distributions of money. Above all, he was popular with the armed forces; he was the soldier-emperor *par excellence.* Understandably, he received the title Optimus (Best), officially from 114 on (and unofficially for many years earlier).

Yet Trajan was a thoroughgoing autocrat who intervened without hesitation or scruple even in the senatorial sphere, whenever it seemed necessary. His aim was efficiency; his desire was to promote public welfare everywhere. He embellished Rome with splendid and substantial structures, and he showed his care for Italy by refurbishing and enlarging the harbours at Ostia, Centumcellae, and Ancona. He sent officials called *curatores* to Italian municipalities in financial difficulties and helped to rehabilitate them. He greatly expanded an ingenious charity scheme, which Nerva had probably begun: money was loaned to farmers on easy terms, and the low interest they paid went into a special fund for supporting indigent children. Nor did Trajan neglect Italy's highway network: he built a new road (Via Traiana) that soon replaced the Via Appia as the main thoroughfare between Beneventum and Brundisium.

Interest in Italy implied no neglect of the provinces. *Curatores* were also sent to them; and to rescue Achaea and Bithynia, senatorial provinces, from threatened bankruptcy, Trajan made them both temporarily imperial, sending special commissioners of his own to them. His correspondence with his appointee in Bithynia, the younger Pliny, has survived and reveals how meticulously the Emperor superintended even the smallest details.

Trajan's sense of responsibility cannot be doubted, and his well-meaning example inspired his immediate successors to imitate his paternalism. Such imperial encroachment, however, helped to create a huge and burdensome bureaucracy and to stifle the local initiative on which municipal autonomy throughout the empire depended.

Expansion of the empire under Trajan

Trajan's efforts were as notable in the military sphere as in the civilian. He decided to strengthen the dangerous Danube frontier by converting Dacia into a salient of Roman territory north of the river in order to dismember the Sarmatian tribes and remove the risk of large, hostile combinations to a safer distance. Accordingly, he mustered a large army, conquered Decebalus in two hard-fought wars (101–102; 105–106), and annexed Dacia and filled it with settlers from neighbouring parts of the empire. On the eastern frontier he planned a similar operation, evidently in the conviction, shared by many eminent Romans both before and after him, that only conquest could solve the Parthian problem. Possibly, too, he wished to contain the menace of the Sarmatian Alani in the Caspian region. In a preliminary move, the Nabataean kingdom of Arabia Petraea was annexed in 105–106. Then in 114 Trajan assembled another large army, incorporated the client kingdom of Armenia, and invaded Parthia.

After spectacular victories in 115 and 116, he created additional provinces (Northern Mesopotamia, Assyria) and reached the Persian Gulf. But he had merely overrun Mesopotamia; he had not consolidated it, and, as his army passed, revolts broke out in its rear. The Jews of the Diaspora and others seized their chance to rebel, and before the end of 116 much of the Middle East besides Parthia was in arms (Cyrene, Egypt, Cyprus, Asia Minor). Trajan proceeded resolutely to restore the situation, but death found him still in the East.

Before his last illness, he had not formally indicated his successor. But high honours and important posts had been accorded his nearest male relative, Publius Aelius Hadrianus, the governor of Syria; and, according to Trajan's widow, Hadrian had actually been adopted by Trajan on his deathbed. Accordingly, both Senate and soldiers recognized him. Trajan's posthumous deification was never in doubt.

Hadrian. Hadrian (ruled 117–138), also a Spaniard, was an emperor of unusual versatility. Unlike Trajan, he was opposed to territorial expansion. Being himself in the East in 117, he renounced Trajan's conquests there immediately and contemplated evacuating Dacia as well. Furthermore, four of the consular generals particularly identified with Trajan's military ventures were arrested and executed "for conspiracy"; Hadrian claimed later that the Senate ordered their deaths against his wishes. His policies made his a peaceful reign. The only heavy fighting occurred in Judaea—or Syria Palaestina, as it was thenceforth called—where Bar Kokhba led a furious, if futile, Jewish revolt (132–135) against Hadrian's conversion of Jerusalem into a Roman colony named Aelia Capitolina.

Hadrian's policies

Instead of expansion by war, Hadrian sought carefully delimited but well-defended frontiers, with client states beyond them where possible. The frontiers themselves, when not natural barriers, were strongly fortified: in Britain, Hadrian's Wall, a complex of ditches, mounds, forts, and stone wall, stretched across the island from the Tyne to the Solway; Germany and Raetia had a *limes* (fortified boundary) running between Mainz on the Rhine and Regensburg on the Danube. Within the frontiers the army was kept at full strength, mostly by local recruiting of legionaries and apparently of auxiliaries, too (so that Vespasian's system of having the latter serve far from their homelands gradually ceased). Moreover, the tendency for auxiliaries to be assimilated to legionaries continued; even the officers became less distinguishable, because equites now sometimes replaced senators in high posts in the legions. To keep his essentially sedentary army in constant readiness and at peak efficiency (no easy task), Hadrian carried out frequent personal inspections, spending about half his reign in the provinces (121–125; 128–134). The transformation of field troops into static garrisons frequently meant entrusting mobile operations to the so-called numeri—light-armed tribal units, often obtained outside the empire—and this contributed to the barbarization of the Roman imperial army.

Hadrian was also responsible for significant developments on the civilian side. Under him, equites were no longer required to do military service as an essential step in their career, and many of them were employed in the imperial civil service, more even than under Domitian. By now the formative days of the civil service were over; its bureaucratic phase was beginning, and it offered those equites who had no military aspirations an attractive, purely civilian career. Formal titles now marked the different equestrian grades of dignity: a procurator was *vir egregius;* an ordinary prefect, *vir perfectissimus;* a praetorian prefect, *vir eminentissimus,* the latter title being obviously parallel to the designation *vir clarissimus* for a senator. Thenceforth, equites replaced freedmen in the imperial household and bureaus, and they even appeared in Hadrian's imperial council.

Hadrian also improved legal administration. He had his expert jurists codify the *edictum perpetuum* (the set of rules gradually elaborated by the praetors for the interpretation of the law). He also appointed four former consuls to serve as circuit judges in Italy. This brought Italy close to a province; not that Hadrian was anxious to reduce the status of Italy—his policy was to make all parts of the empire important. For one part he was exceptionally solicitous: he spent much time in Greece and lavishly embellished Athens.

Hadrian maintained good relations with but was never fully trusted by the Senate. His foreign policy seemed to be unheroic, his cosmopolitanism to be un-Roman, and his reforms to encroach on activities traditionally reserved to senators. Moreover, in his last two years he was sometimes capricious and tyrannous. Like Augustus, he had no son of his own and conducted a frustrating search for a successor. After executing his only male blood relative, his grandnephew, in 136, he adopted Lucius Ceionius Commodus, renaming him Lucius Aelius Caesar. The latter, however, died shortly afterward, whereupon Hadrian in 138 chose a wealthy but sonless senator, the 51-year-old Titus Aelius Antoninus; but, evidently intent on founding a dynasty, he made Antoninus in his turn adopt two youths, 16 and seven years old, respectively—they are known to history as Marcus Aurelius (the nephew of Antoninus' wife) and Lucius Verus (the son of Aelius Caesar). When Hadrian died soon thereafter, Antoninus succeeded and induced a reluctant Senate to deify the deceased emperor. According to some, it was this act of filial piety that won for Antoninus his cognomen, Pius.

Antoninus' reign of quiet prosperity

Antoninus Pius. Antoninus Pius (ruled 138–161) epitomizes the Roman Empire at its cosmopolitan best. He himself was of Gallic origin; his wife was of Spanish origin. For most men his was a reign of quiet prosperity, and the empire under him deserves the praises lavished upon it by the contemporary writer Aelius Aristides. Unlike Hadrian, Antoninus travelled little; he remained in Italy, where in 148 he celebrated the 900th anniversary of

Rome. *Princeps* and Senate were on excellent terms, and coins with the words *tranquillitas* and *concordia* on them in Antoninus' case mean what they say. Other of his coins not unreasonably proclaim *felicitas ternporum* ("the happiness of the times"). Yet raids and rebellions in many of the borderlands (in Britain, Dacia, Mauretania, Egypt, Palaestina, and elsewhere) were danger symptoms, even though to the empire at large they seemed only faraway bad dreams, to use the expression of Aelius Aristides. Antoninus prudently pushed the Hadrianic frontiers forward in Dacia, the Rhineland, and Britain (where the Antonine Wall from the Forth to the Clyde became the new boundary) and carefully groomed his heir apparent for his imperial responsibilities.

Marcus Aurelius. Marcus Aurelius (ruled 161–180) succeeded the deified Antoninus and more than honoured Hadrian's intentions by immediately co-opting Lucius Verus as his full co-emperor. Because Verus' competence was unproved, this excess of zeal was imprudent. Fortunately, Verus left decision making to Marcus. Marcus' action was also dangerous for another reason; it represented a long step away from imperial unity and portended the ultimate division of the empire into Greek- and Latin-speaking halves. Nor was this the only foreboding development in Marcus' reign—formidable barbarian assaults were launched against the frontiers, anticipating those that were later to bring about the disintegration of the empire. Marcus himself was a Stoic philosopher; and his humanistic if somewhat pessimistic *Meditations* reveal how conscientiously he took his duties. Duty called him to war; he responded to the call and spent far more of his reign in the field than had any previous emperor.

At Marcus' very accession the Parthians turned aggressive, and he sent Verus to defend Roman interests (162). Verus greedily took credit for any victories but left serious fighting to Avidius Cassius and the army of Syria. Cassius succeeded in overrunning Mesopotamia and even took Ctesiphon, the Parthian capital; he was therefore able to conclude a peace that safeguarded Rome's eastern provinces and client kingdoms (166). In the process, however, his troops became infected with plague, and they carried it back with them to the west with calamitous results. The Danube frontier, already weakened by the dispatch of large detachments to the East, collapsed under barbarian assault. Pressed on from behind by Goths, Vandals, Lombards, and others, the Germanic Marcomanni and Quadi and the Sarmatian Iazyges poured over the river; the Germans actually crossed Raetia, Noricum, and Pannonia to raid northern Italy and besiege Aquileia. Marcus and Verus relieved the city shortly before Verus' death (169). Then, making Pannonia his pivot of manoeuvre, Marcus pushed the invaders back; by 175 they were again beyond the Danube. At that moment, however, a false report of Marcus' death prompted Avidius Cassius, by now in charge of all eastern provinces, to proclaim himself emperor. Cassius was himself a native Syrian, and his action may represent incipient separatism by the Greek-speaking East. In any case, it undid Marcus' achievements along the Danube because it took him to the East and reopened the door to barbarian attacks. Fortunately, Cassius was soon murdered, and Marcus could return to central Europe (177). But he had barely restored the frontier again when he died at Vindobona (Vienna) in 180, bequeathing the empire to his son, the 19-year-old Commodus, who had actually been named co-emperor three years earlier.

Barbarian assaults

Cornmodus. Commodus (ruled 180–192), like Gaius and Nero, the youthful emperors before him, proved incompetent, conceited, and capricious. Fortunately, the frontiers remained intact, thanks to able provincial governors and to barbarian allies, who had been settled along the Danube with land grants and who gave military service in return. But Commodus abandoned Marcus' scheme for new trans-Danubian provinces, preferring to devote himself to sensual pleasures and especially to the excitements of the arena in Rome, where he posed as Hercules Romanus and forced the Senate to recognize his godhead officially. He left serious business to his favourites; and their ambitions and intrigues led to plots, treason trials, confiscations, and insensate murders. Commodus' assassination on the last day of 192 terminated a disastrous reign; and the Antonines, like the Julio-Claudians, came to an ignominious end. And there was a similar sequel. Commodus' *dnmnatio memorine,* like Nero's, was followed by a year of four emperors.

THE EMPIRE IN THE 2ND CENTURY

The century and three-quarters after Augustus' death brought no fundamental changes to the principate, although so long a lapse of time naturally introduced modifications and shifts of emphasis. By Flavian and Antonine times the principate was accepted universally. For the provinces, a return to the republic was utterly unthinkable; for Rome and Italy, the year 69 served as a grim warning of the chaos to be expected if, in the absence of a *princeps,* the ambitions of a few powerful individuals obtained unfettered scope. *A princeps* was clearly a necessity, and men were even prepared to tolerate a bad one, although naturally they always hoped for a good one.

Nor did the *princeps* have to be chosen any longer from the Julio-Claudians. The great achievement of the Flavians was to reconcile the soldiers and the upper classes everywhere to the idea that others were eligible. The Flavians' frequent tenure of consulship and censorship invested their middle class Italian family with the outward trappings of prestige and the aristocratic appearance of an authentic imperial household. The deification of the first two Flavians contributed to the same end, and so did the disappearance of old republican families that might have outranked the reigning house (by 69 most descendants of the republican nobility had either died or been exterminated by imperial persecution). After the Flavians, the newness of a man's senatorial dignity and the obscurity of his ultimate origin, Italian or otherwise, no longer forbade his possible elevation. Indeed, Domitian's successors and even Domitian himself in his last years did not need to enhance their own importance by repeated consulships. The Antonine emperors, like the Julio-Claudians, held the office infrequently. They did, however, continue the Flavian practice of emphasizing the loftiness of their families by deifying deceased relatives (Trajan deified his sister, niece, and father; Antoninus, his wife; etc.).

Eligibility for the principate

Trend to absolute monarchy. Glorification of the reigning house, together with a document such as Vespasian's Lex de Imperio, helped to advertise the emperor's position; and under the Flavians and Antonines the principate became much more like an avowed monarchy. Proconsular *imperium* began to be reflected in the imperial titulary, and official documents started calling the emperor *dominus noster* (" our master").

The development of imperial lawmaking clearly illustrates the change. From the beginnings of the principate, the emperor had had the power to legislate, although no law is known that formally recognized his right to do so; by Antonine times, legal textbooks stated unequivocally that whatever the emperor ordered was legally binding. The early emperors usually made the Senate their mouthpiece and issued their laws in the form of senatorial decrees; by the 2nd century the emperor was openly replacing whatever other sources of written law had hitherto been permitted to function. After 100 the Assembly never met formally to pass a law, and the Senate often no longer bothered to couch its decrees in legal language, being content to repeat verbatim the speech with which the ruler had advocated the measure in question. After Hadrian, magistrates ceased modifying existing law by their legal interpretations because the praetors' *edictum perpetuum* had become a permanent code, which the emperor alone could alter. By 200, learned jurists had lost the right they had enjoyed since the time of Augustus of giving authoritative rulings on disputed points *(responsa prudentium).* Meanwhile, the emperor more and more was legislating directly by means of edicts, judgments, mandates, and rescripts—collectively known as *constitutiones principum. He* usually issued such *constitutiones* only after consulting the "friends" *(amici Caesaris)* who composed his imperial council. But a ***constitutio*** was nevertheless a fiat. The road to the latter dominate (after 284) lay open.

Political life. Nevertheless, the autocratic aspect of the Flavian and Antonine regimes should not be overstressed. Augustus himself had been well aware that it was impossible to disguise permanently the supremacy that accumulation of powers gained piecemeal conferred; his deportment in his last years differed little from that of Vespasian, Titus, and the so-called five good emperors who followed them. Nor had other Julio-Claudians hesitated to parade their predominance—Claudius, by centralizing the imperial powers, reduced their apparent diversity to one all-embracing *imperium;* Gaius and Nero revealed the autocracy implicit in the principate with frank brutality.

Civilian character of the Flavian and Antonine reigns

What impresses perhaps as much as the undoubtedly autocratic behaviour of the Flavians and Antonines is the markedly civilian character of their reigns. They held supreme power, and some of them were very distinguished soldiers; yet they were not military despots. For this the old republican tradition—that a state official might serve it in both a civilian and a military capacity—was largely responsible. Matters could change after Hadrian separated the two roles. Actually, the 3rd century soon showed what it meant to have a ***princeps*** whose whole experience had been confined to camps and barracks.

As imperial powers became more concentrated, republican institutions decayed; the importance of imperial officials grew, while the authority of urban magistrates declined. Quaestorship, praetorship, and consulship (the last named now reduced to a two-month sinecure) became mere stepping stones to the great imperial posts that counted most in the life of the empire. Governors of imperial provinces and commanders of legions were Roman senators; but they were equally imperial appointees. Clearly, the emperor was the master of the Senate; and it was disingenuous for him to get impatient, as some emperors did, with the Senate's lack of initiative and reluctance to take firm decisions of its own. The emperor might not even consult the Senate much, preferring to rely on his imperial council, which by the 2nd century was no longer composed exclusively of senators.

The Senate, however, while exercising little power, was treated courteously by most Flavians and Antonines. They recognized its importance as a lawcourt, as the body that formally appointed a new emperor, and as a sounding board of informed opinion. Senators came increasingly from the provinces, and, although this meant pre-eminently the western provinces (the Greek-speaking East being underrepresented), the Senate did reflect to some extent the views of the empire at large.

The equites, meanwhile, steadily acquired greater importance as imperial officials. In newly created posts they were invariably the incumbents, and in posts of long standing they replaced freedmen and ***publicani.*** The 2nd century saw equestrian procurators become the solid basis of the civil service—four grades, distinguished by salary, were established. The great .bureaucracy of later times was fairly launched. This development inevitably bred bitter rivalry with the Senate, mitigated somewhat by the similar social backgrounds of the two orders.

Rome as a melting pot

Rome and Italy. By the 2nd century the city of Rome had attracted freeborn migrants from all over the empire; it housed, additionally, large numbers of manumitted slaves. These newcomers were all assimilated, making Rome a veritable melting pot, and they hoped—especially the many who came from the East—to dilute the city's Italian flavour. The vast majority of them were poor, a handful of opulent imperial freedmen being entirely exceptional. But many were energetic, enterprising, and lucky and made their way in the world. Freedmen laboured under a social stigma, although some of them managed to become equites. Their sons, however, might overcome discrimination, and their grandsons were even eligible for membership in the Senate.

Inevitably, there was extensive trade and commerce (much of it in freedman hands) in so large a city, which was also the centre of imperial administration. There was little industry, however, and the urban poor had difficulty finding steady employment. Theirs was a precarious existence, dependent on the public grain dole and on the private charity of the wealthy. Large building programs gave Flavian and Antonine emperors the opportunity not only to repair the damage caused by fire and falling buildings (a frequent hazard among the densely packed and flimsily built accommodations for the urban plebs) but also to relieve widespread urban unemployment. This also made imperial Rome a city of grandeur. Augustus' building program had been vast but mostly concerned with repairing or rebuilding structures already existing, and his Julio-Claudian successors had built relatively little until the great fire made room for the megalomaniac marvels of Nero's last years. It was under the Flavians and Antonines that Rome obtained many of its most celebrated structures: the Colosseum, Palatine palaces, Trajan's Forum, the Pantheon, the Castel Sant' Angelo (Hadrian's mausoleum), the Temple of Antoninus and Faustina, Aurelius' Column, as well as the aqueducts whose arches marched across the Campagna to keep the city and its innumerable fountains supplied with water.

Italy was much less cosmopolitan and sophisticated and, according to literary tradition, much more sober and straitlaced than was Rome. It was the mistress of the empire, although the gap between it and the provinces was narrowing. Hadrian's policies especially helped to reduce its privileged position. His use of circuit judges was resented precisely because with them Italy resembled a province; actually, Italy badly needed them, and their abolition by Antoninus Pius was soon reversed by Marcus Aurelius. Also in Aurelius' reign a provincial fate overtook Italy in the form of barbarian invasion; and a few years later the country got its first legionary garrison under Septimius Severus.

The economic importance of Italy also declined. By Antonine times the eastern provinces had overtaken it, and it was hardly ahead even of those of the west. Gaul, for instance, could rival Italy industrially. In agricultural products, northern Africa and Spain are revealed by the potsherds of Monte Testaccio to have become serious competitors to Italian growers. The emperors, however, were not indifferent to Italy—Domitian, for example, protected Italian viticulture by restricting vine growing in the provinces; Trajan and his successors forced Roman senators to take an interest in the country, which was no longer the homeland of many of them, by investing a high proportion of their capital in Italian land (one-third under Trajan, one-quarter under Aurelius).

Attempts to achieve stability in the provinces

Developments in the provinces. In the empire at large, Flavians and Antonines, like the better Julio-Claudians, aimed at stability in order that its inhabitants might live in security and self-respect. In this they largely succeeded. Gibbon's famous description of the 2nd century as the period when men were happiest and most prosperous is not entirely false. Certainly, by then men had come to take for granted the unique greatness and invincible eternity of the empire; even the ominous events of Aurelius' reign failed to shatter their conviction that the empire was impregnable. The empire was a vast congeries of peoples and races with differing religions, customs, and languages, and the emperors were content to let them live their own lives. Imperial policy favoured a veneer of common culture transcending ethnic differences, but there was no deliberate denationalization. Ambitious men striving for a career naturally found it helpful, if not necessary, to become Roman in bearing and conduct and perhaps even in language as well (although speakers of Greek often rose to exalted positions). But local self-government was the general rule, and neither Latin nor Roman ways were imposed on the communities composing the empire. The official attitude to religion illustrates this—in line with the absolutist trend, emperor worship was becoming slowly but progressively more theocratic (Domitian relished the title of god, Commodus demanded it); yet this did not lead to the suppression of non-Roman or even outlandish cults, unless they were thought immoral (like Druidism, with its human sacrifice) or conducive to public disorder (like Christianity, with its uncompromising rejection of emperor worship and its liability to become a target for riots).

Local autonomy. Where possible, the emperors kept direct administration from Rome to a minimum (except perhaps in Egypt), and the 2nd century was the most flourishing period of urban civilization that the empire ever knew. Administration everywhere was in the hands of the local well-to-do, who alone could afford the costs attaching to it. Local magistrates paid substantial fees on election; nor did their expenses end there, for they remained members of their local council after their term of office, and local councillors (decuriones, later curiales) were expected to contribute handsomely to local amenities. The cost involved inevitably reduced local elections to formalities, and the local councils became self-perpetuating bodies whose members were persons of some consequence. It was from these local worthies that the emperor often found his candidates for the Senate at Rome, an honour that was eagerly sought by individuals but that was a mixed blessing for their local communities, which stood thereby to lose prospective benefactors.

It is impossible not to be impressed by the spectacle of the Roman Empire in its 2nd-century heyday, with its panorama of splendid and autonomous communities. Nevertheless, before the century was over, there was growing difficulty in maintaining flourishing municipal life in a world where the ordinary man was encouraged to regard the emperor as a sort of terrestrial Providence and where the emperor himself with responsible earnestness accepted the role of universal dispenser of justice. The letters of the younger Pliny and of Marcus Cornelius Fronto reveal how seriously the 2nd-century emperors took their duty and strove for orderly government everywhere. But the emperors' very conscientiousness led inevitably to interference with local autonomy. Perhaps the civil service that Augustus founded would have burgeoned in any event and encroached on the self-governing communities that made up the provinces; but the well-meaning efforts of the Antonines hastened such a development. The ultimate effect was to dampen civic ardour and to foster listlessness. Faced with the prospect of increasing direction from above, municipal notables began avoiding local office. Inability to pay the cost involved may also in part explain the growing reluctance of men to undertake municipal responsibilities; although the provincial bourgeoisie remained generally prosperous, economic recession had set in before the 2nd century ended. For whatever reason, local officeholders became less easy to find; and, well before 200, men were being compelled to accept local office. This boded ill for the future.

The emperors' interference withlocal autonomy

Social structure. Information about the lower orders throughout the empire is not very plentiful. So long as economic activity remained at a high level, there was apparently much social mobility. This was just as well, because there were great disparities of wealth and of class. Society was stratified, and the gap between poor and rich in provincial towns somewhat resembled that between plebs and aristocracy, senatorial or equestrian, in Rome. Indeed, it was precisely in Antoninus' reign that Roman law began to distinguish the lower classes (*humiliores*) from the upper (*honestiores*), and, although local customs and traditions were respected, the civil law of Rome was gradually replacing local codes. The 2nd century and the opening years of the 3rd comprised the classical period of Roman law, producing such illustrious jurists as Salvius Julianus, Gaius, Papinian, Ulpian, and Julius Paulus; and the temptation for the communities of the empire to adopt it must have been strong.

Spread of citizenship. Clearly, Romanization was spreading and with it the Roman citizenship, especially from Hadrian's reign on. The number of provincials with citizenship increased so rapidly that 20 years after Commodus' death the emperor Caracalla could enfranchise all the communities of the empire and excite very little comment in doing so. The subjects of the empire found this levelling of status comfortable; they were also grateful for the peace that prevailed over much of the empire, but they were not particularly eager to defend it. Roman citizens in general avoided military service, and the empire obtained its soldiers by offering Roman citizenship as a reward for enlistment. The ever-wider extension of the citizenship ended its usefulness as a bribe, and the recruiting areas steadily contracted to the least Rornanized districts. Finally, recruits had to be obtained outside the empire altogether.

Evidently, the empire did not inspire a sense of duty or high resolve. There were both political and economic reasons. The average nonwealthy man hardly participated in political life, even in his own community, and his share of the prevailing prosperity was small. Slavery was ubiquitous; and, although humanitarian legislation gradually softened its harsher features, its unhappy victims depressed the living standards of the free workers so that poverty was a serious social problem. Drab materialism was the lot of many—in Asia Minor, for instance, despite unparalleled prosperity brought about by its having the whole Mediterranean open to it as at no other time, great unrest developed as a result of extremes of affluence and poverty. Many men sought solace from worldly cares in the mystery religions, and the prevailing polytheism enabled them to switch or multiply their gods without much problem of conscience (except for the increasingly numerous Christian converts).

Growing use of Latin in Iberia and Gaul

Romanization. By the 2nd century considerable Latinization had occurred in the West. Modem Spanish, Portuguese, and French show this was particularly true of the Iberian Peninsula, which had been provincial soil ever since the Second Punic War, and of Gaul, where Latin enjoyed the advantage of some relationship to Celtic. In these regions even the lower orders adopted Latin, although the survival of Basque proves that the older native languages did not perish utterly. Moreover, enemy threats were far enough away and integration far enough advanced to render large garrisons unnecessary. Elsewhere, and especially in central Europe, large forces remained and were themselves potent instruments of Latinization, because in the army Latin was both the official language and also the lingua franca for heteroglot recruits. Furthermore, thanks to the coloniae, urbanization increased; and urbanization promoted Latinization. In the Danubian region, for instance, where the barbarian threat was acute, ceaseless military activity led to the establishment of numerous coloniae, to the spread of Latin (and its survival in modem Romanian), to the disappearance of ancient Illyrian and Thracian (unless modern Albanian preserves them), and to the birth of a strong pro-Roman patriotism, which increased men's eagerness to learn Latin. In areas away from the frontiers and especially in areas lacking the economic resources to generate cities, often only the bourgeoisie acquired Latin; among the lower orders the pre-Roman languages, even if seldom written, continued in use: Celtic in Britain and Switzerland and, to a lesser extent, in Gaul; Berber and some Punic in Africa and Mauretania. Moreover, from Hadrian on—when local recruiting and local service became the rule, and coloniae were no loager new foundations but simply old towns upgraded in status—the army lost much of its Latinizing role. Complete linguistic unity was therefore never achieved in the West.

This was true also of the East, where Greek had the status of an official language. Rome readily accepted and even encouraged Hellenization of the whole region; and, although the Romans were gratified if Orientals adopted Latin, they realized the futility of trying to impose it. In many districts, however, Greek was spoken only by the administrative and propertied classes, just like Latin in the West. Lower orders, especially in inland districts, continued to use their vernaculars: Aramaic in Syria, Coptic in Egypt, and a variety of tongues in Asia Minor.

The official emperor worship became universal, but Rome was no fountainhead of religious ideas; local cults persisted in both East and West, and the East even gave birth to new religions, such as Christianity. The native divinities were worshipped either under their own names or under those of Greco-Roman deities with whom they could be conveniently identified, for there was much syncretism. In religion as in language the army played an important role. The soldiers, with some help from eastern merchants, spread the worship of such deities as Mithra and Dolichenus, with whom they had become acquainted in the East, to the farthest corners of the empire; Mithra became known even in distant Britain, and Dolichenus was assimilated to Jupiter everywhere.

Differences between the Greek East and the Latin West

In the material aspects of life there was much Romanization. Towns in the Latin-speaking provinces looked very Roman in layout and in buildings: their streets were arranged on a checkerboard plan, and they usually boasted forums, triumphal arches, basilicas, baths, and Roman-type temples, theatres, and amphitheatres, graced with Roman sculpture and decoration. The eastern cities were much less receptive to Roman influences; in fact, the Latin West and the Greek East tended to go their separate ways. Differing outlooks and values, divergent interests and degrees of development, mutual suspicions and antipathies combined to keep them apart; and by Marcus Aurelius' day the blueprint for a divided empire had already been sketched in outline, no matter how faintly. Yet, even in the East, Roman influence was not negligible: some Greek cities became partly Roman in aspect, went in for gladiatorial games and similar Roman entertainments, and adopted the Roman civil law, the most enduring of the empire's legacies.

Economic life. Romanization was due chiefly to Rome's political authority, but the conditions fostered by the Pax Romana also helped. Harbour dues and customs duties prevented the Roman Empire from being one large free-trade area. Nor was there complete laissez faire in other respects, despite a general lack of policies for improving the increasingly complex economy by promoting trade and regulating commerce: there was much routine administration directly concerned with production and distribution of goods. The supervision of the enormous and ubiquitous imperial properties, the provision of Rome's food supply, and procurement for the armed services and other branches of government constituted a significant proportion of the empire's economic activity. Nevertheless, in the prevailing peace, the private sector offered great scope for initiative, enterprise, the free play of competition, and general economic development. Industrial and commercial methods, banking, and credit facilities were all improved, and goods moved easily along the highways crisscrossing the empire from frontier to frontier. There was some trade with countries beyond the frontiers, such as Scandinavia and even China (indirectly), although naturally there was far more within the empire itself. In the West, large Roman (or at least Romanized) towns came into being to serve the needs of trade. In the East, Asia Minor and Syria interchanged goods with the whole Mediterranean world, and the Greek cities flourished as never before, Alexandria and Antioch being the greatest of them but with those of Asia Minor not far behind.

Waning of prosperity during the 2nd century

But in economic as in other matters there was premonition of trouble to come. Before the 2nd century ended, prosperity had begun to wane. A clear symptom of this was the reduction in the weight and fineness of the standard silver coin, the denarius. The economy faltered for various reasons: the consistent drain of metal to the East to pay for spices and Oriental luxuries, the maldistribution of wealth, the lack of industrial inventiveness in a slave-based society, the depopulation caused by war (by Aurelius' reign no longer confined to frontier areas). The economic chaos of the 3rd century, however, was still some distance away.

The army. The army that enforced the Pax Romana had expanded little beyond the size envisaged for it by Augustus, despite the enlargement of the empire by Claudius, the Flavians, and Trajan. It reached 31 legions momentarily under Trajan, but it usually totalled 28 under the Flavians and Antonines until the onset of the frontier crisis in Aurelius' reign brought it to 30. This, seemingly, was the limit for which either recruits or money could be found; and it probably explains why Hadrian, and later Commodus, halted further expansion.

The army was used not to prop up a militarist government but to defend the frontiers. Shifts in enemy pressures, however, caused the legions to be distributed differently than in Julio-Claudian times. Under Antoninus Pius, the Danubian provinces (Pannonia, Moesia, Dacia) had ten, and the East (Asia Minor, Syria, Palestine, Egypt) had nine, and both regions also had supporting naval flotillas; of the remaining nine legions, Britain contained three and the Rhineland four. Clearly, the army was thinly stretched elsewhere, with auxiliaries bearing the burden of defense. In much of the empire the sight of a soldier, especially a legionary, was a comparatively rare event, the more so because normally troops not only remained in frontier areas while serving but were also largely recruited there.

The provinces underwent changes in size as well as in garrison. For prudential reasons the emperors refused to entrust many legions to a single commander; they split a province in two as soon as it needed a large military establishment. Thus, the number of provinces was constantly growing; by Hadrian's day subdivision began to anticipate the fragmentation later carried out by Diocletian.

Imperial literature

Cultural life. The literature of the empire is both abundant and competent, for which the emperors' encouragement and financing of libraries and higher education were perhaps in part responsible. The writers, however, with the possible exception of Christian apologists, were seldom excitingly original and creative. As Tacitus said, the great masters of literature have ceased to be. Perhaps Augustus' emphasis on tradition affected more than political ideals and practice. At any rate, men of letters, too, looked often backward. At the same time they clearly reveal the success of the empire in spreading Greco-Roman culture, for the majority of them were natives of neither Italy nor Greece. Of the writers in Latin, the two Senecas, Lucan, Martial, Columella, Hyginus, and Pomponius Mela came from Sapin; Fronto, Apuleius, and probably Florus and Aulus Gellius, from Africa. Tacitus was perhaps from Gallia Narbonensis. The Latin writers in general sought their models less in Greece than in Augustus' Golden Age, when Latin literature had reached maturity. Thus, the poets admired Virgil and imitated Ovid; lacking genuine inspiration, they substituted for it an erudite cleverness, the fruit of an education that stressed oratory of a striking but sterile kind. Authentic eloquence in Latin came to an end when, as Tacitus put it, the principate "pacified" oratory. Under the Flavians and Antonines, an artificial rhetoric, constantly straining after meretricious effects, replaced it. The epigrammatic aphorism (sententia) was especially cultivated; the epics of Lucan, Valerius Flaccus, Silius Italicus, and Statius are full of it, and it found a natural outlet in satirical writing, of which the Latin instinct for the mordant always ensured an abundance. In fact, Latin satire excelled: witness Martial's epigrams, Petronius' and Juvenal's pictures of the period, and Persius' more academic talent. For that matter, Tacitus' irony and pessimism were not far removed from satire.

In the East the official status of Greek and the favour it enjoyed from such emperors as Hadrian gave new life to Greek literature. It had something in common with its Latin counterpart in that it looked to the past but was chiefly written by authors who were not native to the birthplace of the language. The so-called Second Sophistic reverted to the atticism of an earlier day but often in a Roman spirit; its products from the Asian pens of Dio Chrysostom and Aelius Aristides are sometimes limpid and talented tours de force but rarely great literature. In Greek, too, the best work was in satire, the comic prose dialogues of the Syrian Lucian being the most noteworthy and original literary creations of the period. Among minor writers the charm of Arrian and Pausanias, Asians both, and above all of Plutarch abides (although Plutarch's talents were mediocre, and his moralizing was shallow, his biographies, like those of his Latin contemporary Suetonius, are full of information and interest).

Imperial encouragement of Greek culture and a conviction, no longer justified, of its artistic and intellectual superiority caused the East to resist Latinization. This attitude was bound to lead to a divided empire, and thoughtful observers must have noted it with misgivings. The split, however, was still far in the future. Meanwhile, there was a more immediate cause for disquiet. The plethora of summaries and anthologies that appeared implies a public progressively indifferent to reading. In other words, the outlook for letters was poor, and this had an unfortunate effect on the scientific literature of the age, which was in itself of first-class quality. Dioscorides on botany, Galen on medicine, and Ptolemy on mathematics, astronomy, and geography represent expert scholars, expounding carefully, systematically, and lucidly the existing knowledge in their respective fields. But their very excellence proved fatal because, as the reading public dwindled, theirs remained standard works for far too long; their inevitable errors became enshrined, and their works acted as brakes on further progress.

Stoicism was the most flourishing philosophy of the age. In the East a sterile scholasticism diligently studied Plato and Aristotle, but Epictetus, the Stoic from Asia Minor, was the pre-eminent philosopher. In the West, Stoicism permeates Seneca's work and much of Pliny's *Natural History*. Evidently, its advocacy of common morality appealed to the traditional Roman sense of decorum and duty, and its doctrine of a world directed by an all-embracing providence struck a responsive chord in the 2nd-century emperors, though they deeply disapproved of its extremist offshoots, the Cynics: Marcus Aurelius was himself a Stoic.

Imperial art

Imperial art, dealing above all with man and his achievements, excelled in portraits and commemoration of events; Roman sculpture and presumably Roman painting, also, owed much to Greek styles and techniques. It emerged, however, as its own distinctive type. The Augustan age had pointed the way that Roman art would go: Italian taste would be imposed on Hellenic models to produce something different and original. The reliefs of the Augustan Ara Pacis belong to Rome and Italy, no matter who actually carved them. By Flavian times this Roman instinct had asserted itself and with it the old Roman tendency toward lively and accurate pictorial representation. This can be seen from the reliefs illustrating the triumph over Judaea in the passageway of the Arch of Titus. The narrative description dear to Roman art found its best expression in the great spiral frieze on Trajan's Column, where the Emperor can be seen among his soldiers at various times in the Dacian campaigns; the story of the war plays a most important part, although, like most imperial monuments, the column is meant to exalt the leader. Under Hadrian a reaction made sculpture less markedly Italian, as if to conform with the slow decline of Italy toward quasi-provincial status. Also under Hadrian, the emperor was more singled out, being made bigger and more frontal than the other figures, as if to illustrate the growing monarchical tone of the principate. This tendency continued under the Antonines, when there was a magnificent flowering of sculpture on panels, columns, and sarcophagi; but its exuberance and splendour foreshadow the end of classical art.

The artistic currents that flowed in Rome were felt throughout the empire, the less developed areas being influenced most. In the West, provincial sculpture closely resembled Roman, although it sometimes showed variations, in Gaul especially, due to local influences (the native element, however, is not always easy to identify). The Roman quality of portraits painted on Egyptian mummy cases shows that the Greek-speaking regions were also affected, although generally they maintained their own traditions. But by now the Greek East had become rather barren; much of its production was imitative rather than vitally creative. Greece proper contributed little, the centre of Hellenism having shifted to Asia Minor, to places like Aphrodisias, where there was a flourishing school of sculpture.

In at least one respect the East was heavily influenced by Rome. The use of concrete and cross vault enabled Roman architects and engineers to span wide areas; their technological achievements included the covered vastness of the huge thermal establishments, the massive solidity of the amphitheatres, and the audacity of the soaring bridges and aqueducts. The East was greatly impressed. Admittedly, the agoras and gymnasiums in Greek towns are hardly Roman in aspect, but, for most structures of a practical utilitarian kind, the Greek debt to Rome was heavy. Sometimes Roman influence can be seen not only in the fundamental engineering of such buildings as market gateways, theatres, and amphitheatres, but even in such decorative details as composite capitals as well. Roman features abound in exotic Petra, Palmyra, Gerasa, and Baalbek, and even in Athens itself. (E.T.S.)

V. The later Roman Empire

THE DYNASTY OF THE SEVERI (AD 193–235)

Septimius Severus. *Establishment of the dynasty.* After the assassination of Commodus on December 31, AD 192, Helvius Pertinax, the prefect of the city, became emperor. In spite of his modest birth, he was well respected by the Senate but was without his own army. He was killed by the praetorians at the end of March 193, after a three-month reign. The praetorians, after much corrupt bargaining, designated as emperor an old general, Didius Julianus, who had promised them the largest *donativum* (a donation given to each soldier on the emperor's accession). The action of the praetorians roused the ire of the provincial armies. The army of the Danube, which was the most powerful as well as the closest to Rome, appointed Septimius Severus in May 193. Severus soon had to face two competitors, supported, like himself, by their own troops: Pescennius Niger, the legate of Syria; and Clodius Albinus, legate of Britain. After having temporarily neutralized Albinus by accepting him as Caesar, Septimius marched against Niger, whose troops, having come from Egypt and Syria, were already occupying Byzantium. The Danubian legions were victorious, and Niger was killed at the end of 194; Antioch and Byzantium were pillaged after a long siege. Septimius even invaded Mesopotamia, for the Parthians had supported Niger. But this campaign was quickly interrupted: in the West, Albinus, disappointed at not being associated with the empire, proclaimed himself Augustus in 196 and invaded Gaul. He was supported by the troops, by the population, and even by the senators in Rome. In February 197, he was defeated and killed in a difficult battle near his capital of Lyons which, in turn, was almost devastated. Septimius Severus remained the sole master of the empire, but the pillagings, executions, and confiscations left a painful memory. A few months later, in the summer of 197, he launched a second Mesopotamian campaign, this time against the Parthian king Vologases IV, who had attacked Nisibis, which the Romans had conquered two years previously. Septimius Severus was again victorious. Having arrived at the Parthian capitals (Seleucia and Ctesiphon), he was defeated near Hatra but in 198 obtained an advantageous peace: Rome retained a part of Mesopotamia, together with Nisibis, the new province being governed by an eques. After having inspected the East, the Emperor returned to Rome in 202. He spent most of his time there until 208, when the incursions of Caledonian rebels called him to Britain, where he carried out a three-year campaign along Hadrian's Wall. He died at Eboracum (York) in February 211.

Severus' campaigns

The Severan revolution. Septimius Severus belonged to a Romanized Tripolitan family that had only recently attained honours. He was born in Leptis Magna in North Africa and favoured his native land throughout his reign. Married to Julia Domna of Emesa, a Syrian woman from an important priestly family, he was surrounded by Easterners. He had pursued a senatorial career and had proved himself a competent general, but he was above all a good administrator and a jurist. Disliking Romans, Italians, and senators, he deliberately relied on the faithful Danubian army that had brought him to power and always showed great concern for the provincials and the lower classes. Although he had sought to appropriate the popularity of the Antonines to his own advantage by proclaiming himself the son of Marcus Aurelius and by naming his own son Marcus Aurelius Antoninus, he in fact carried out a totally different policy—a brutal, yet realistic policy that opened careers to new social classes. Indifferent to the prestige of the Senate, where he had a great many enemies, he favoured the equites. The army thus became the seedbed of the equestrian order and was the object of all of his attentions. The ready forces were increased by the creation of three new legions commanded by equites; and one of these, the Second Parthica, was installed near Rome. Unlike Vespasian, who also owed his power to the army but who knew how to keep it in its proper place, Septimius Severus, aware of the urgency of external problems, established a sort of military monarchy. The praetorian cohorts doubled their ranks, and the dismissal of the old staff of Italian origin transformed the Praetorian Guard into an imperial guard in which the elite of the Danube army were the most important element. The auxiliary troops were increased by the creation of 1,000-man units (infantry cohorts) and cavalry troops, sometimes outfitted with mail armour in the Parthian manner. The careers of noncommissioned officers emerging from the ranks now opened onto new horizons: centurions and noncommissioned grades could attain the tribunate and enter into the equestrian order. Thus, a simple Illyrian peasant might attain high posts: this was undoubtedly the most essential aspect of the "Severan revolution." This "democratization" was not necessarily a barbarization, for the provincial legions had long been Romanized. Their salaries were increased, and donativa were distributed more frequently; henceforth, soldiers were fed at the expense of the provincials. Veterans received lands, mostly in Syria and Africa. The right of legitimate marriage, previously refused by Augustus, was granted to almost all of the soldiers, and the right to form collegia (private associations) was given to noncommissioned officers. The army was tending to become a privileged caste, issuing from the peasantry. In the 1920s, the Russian scholar Michael Rostovtzeff suggested that this policy marked the beginning of a socially dangerous collusion between the peasants and the army—which was recruited from among them—and indicated a class struggle against the bourgeoisie of the city. But many scholars now believe that Rostovtzeff, perhaps influenced by parallels with the Red Army during the Russian Revolution, overstated his case.

The growth of the equestrian order

The administrative accomplishments of Septimius Severus were of great importance: he clearly outlined the powers of the city prefect; he entrusted the praetorian prefecture to first class jurists, such as Papinian; and he increased the number of procurators, who were recruited for financial posts from among Africans and Easterners and for government posts (praesides) from among Danubian officers. Italy lost its privileges and found itself subjected, like all the other provinces, to the new annona, a tax paid in kind, which assured the maintenance of the army and of the officials. The consequent increase in expenditures—for administration, for the salaries and the donativa of the soldiers, for the maintenance of the Roman plebs, and for construction-obliged the Emperor to devalue the denarius in 194. But the confiscations increased his personal fortune, the res privata, which had been previously created by Antoninus.

Severus' social policy favoured both the provincial recruitment of senators (Easterners, Africans, and even Egyptians), causing a sharp decrease in the percentage of Italian senators, and the elevation of the equestrian order, which began to fill the prince's council with its jurists. The cities, which had been favoured by the Antonines, were more and more considered as administrative wheels in the service of the state: the richest decuriones (municipal councillors) were financially responsible for levying the taxes, and it was for this purpose that the towns of Egypt finally received a *boulē* (municipal senate).

The burden of taxes and forced government service was made weightier by numerous transport duties for the army and for the annona service and was regulated by the jurists through financial, personal, or mixed charges. The state was watchful to keep the decuriones in the service of their cities and to provide a control on their administration through the appointment of curatores rei publicae, or officials of the central goverriment. The lower classes were, in principle, protected against the abuses of the rich, but in fact they were placed at the service of the state through the restrictions imposed on shipping and commercial corporations. The state became more and more a policeman, and the excesses of power of numerous grain merchants (frumentarii) weighed heavily on the little man.

Imperial power, without repudiating the ideological themes of the principate, rested in fact on the army and sought its legitimacy in heredity: the two sons of Septimius Severus, Caracalla and Geta, were first proclaimed Caesars (heirs apparent), the former in 196, the latter in 198; later, they were directly associated in imperial power through bestowal of the title of Augustus, in 198 and 209, respectively. Thus, during the last three years of Septimius Severus' reign, the empire had three Augusti at its head.

Caracalla. Caracalla, the eldest son of Septimius Severus, reigned from 211 to 217, after having assassinated his younger brother, Geta. He was a caricature of his father: violent, megalomaniacal, full of complexes, and, in addition, cruel and debauched. He retained, however, the entourage of the equites and jurists who had governed with his father and enforced to an even greater degree his father's militaristic and egalitarian policy. He increased the wages of the army even further and, at the same time, began a costly building program that quickly depleted the fortune left him by his father. He forced the senators to pay heavy contributions, doubled the inheritance and emancipation taxes, and often required the aurum *coron*ariurn (a contribution in gold), thereby ruining the urban middle classes. To stop the inflation that was depreciating the silver denarius, he created a new coin, the *antoninianus,* whose real value was that of the Marcus Aurelius coin but whose official rate was fixed at twice the value, thus insuring that monetary instability would continue for some time to come. In 212, the famous Constitutio Antoniniana de Civitate gave the right of citizenship to nearly all inhabitants of the empire. This law, known only through a mutilated papyrus, has raised many controversies; but even taking into account the exclusion of numerous barbarians and of a part of the indigenous population of Egypt, the effect of this measure was considerable. It sanctioned the political evolution of several centuries and represented true social advancement for many provincials. It also reinforced imperial unity and complied with the wishes of the great jurists of the time. Each inhabitant of the empire would henceforth have a double citizenship—that of Rome and that of his city of origin. The very diffusion, however, of citizenship status diminished its value, and, in effect, all citizens were reduced to subjects of an absolute prince. On the other hand, the recruitment of legionaries and of municipal magistrates became more difficult, for up to that time acquisition of Roman citizenship had been the principal inducement for undertaking such offices.

Extension of Roman citizenship

Although little endowed with military qualities, Caracalla adopted as his patron Alexander the Great, whom he admired greatly, and embarked on an active external policy. He fought successfully against the Teutonic tribes of the Upper Danube, among whom the Alamanni, as well as the Capri of the Middle Danube, appeared for the first time; he often prudently mixed military operations with negotiation and gave important subsidies and money (in sound currency) to the barbarians, thus arousing much discontent. His ambition was to triumph in the East like his hero of old and, more recently, Trajan and his own father. He invaded Armenia and Adiabene and annexed Osroene in northwest Mesopotamia, joining it to the part of Mesopotamia taken by Septimius Severus. In April 217, while pursuing his march on the Tigris, he was assassinated on the order of one of his praetorian prefects, Marcus Opellius Macrinus.

Macrinus. Macrinus was accepted as emperor by the soldiers, who were unaware of the role he had played in the death of his predecessor. For the first time an eques had acceded to the empire after having been no more than a manager of financial affairs. The senators reluctantly accepted this member of the equestrian order, who, nevertheless, proved to be moderate and conciliatory; but the armies despised him as a mere civilian, and the ancient authors were hostile to him. His reign was brief, and little is known of him. He concluded an inglorious peace with the Parthians, which assured Mesopotamia to Rome through the payment of large sums of money. And to make himself popular, he cancelled Caracalla's tax increases and reduced military expenditures. A plot against him was soon organized: two young grandnephews of Septimius Severus were persuaded by their mothers and especially by their grandmother, Julia Maesa, the sister of Julia Domna (who had recently died), to reach for imperial power. The eldest, Bassianus, was presented to the troops of Syria, who had been bought with gold, and was proclaimed in April 218. Shortly afterward, Macrinus was defeated and killed, as was his son (whom he had associated with him on the throne).

Elagabalus and Severus Alexander. The new emperor was presented as the son of Caracalla, whose name he took (Marcus Aurelius Antoninus). He is better known, however, under the name Elagabalus, the god whose high priest he was and whom he quickly and imprudently attempted to impose on the Romans, in spite of his grandmother's counsel of moderation. Fourteen years old, he caused himself to be detested by his heavy expenditures, his orgies, and the dissolute behaviour of his circle. His fanatic devotion to the Baal of Emesa (Elagabal) was sincere, but he shocked the Romans by his extravagances. The praetorians killed him in 222 and proclaimed as emperor his first cousin, Alexianus, who took the name of Severus Alexander.

Although well educated and full of good intentions, Severus Alexander showed some weakness of character by submitting to the counsel of his mother, Mamaea, and of his grandmother, Maesa. The Scriptores *historiae Augustae*, a collection of biographies of the emperors, attributes to him a complete program of reforms favourable to the Senate, but these reforms are not mentioned elsewhere. As in the time of Septimius Severus, his counsellors were equites. Ulpian, the praetorian prefect, was the greatest jurist of this period, and the basic policies of the founder of the dynasty were carried on but with less energy. This weakening of energy had disastrous results: in Persia, the Arsacids were replaced in 224 by the more ambitious Sāsānid dynasty, who hoped to recover the former possessions of the Achaemenids in the East. Their initial attacks were stopped in 232 by a campaign that was, however, poorly conducted by the Emperor and that alienated the army as a result of its ineptitude. In Rome, there were frequent disorders and, as early as 223, Ulpian had been killed by the praetorians. While gathered on the Rhine to fight the Teutons, the soldiers once again revolted and killed Severus Alexander and his mother. A coarse and uneducated but energetic soldier, Maximinus the Thracian, succeeded him without difficulty in March 235. The Severan dynasty had come to an end.

Rise of the Sāsānids in Persia

RELIGIOUS AND CULTURAL LIFE IN THE 3RD CENTURY

Official religions and Eastern cults. Since the reforms of Augustus and his efforts to restore the traditional religion, few religious changes had taken place. The official ceremonies in honour of the Capitoline divinities Jupiter, Juno, and Minerva, the protectors of the state, were still celebrated by the priests of the sacerdotal collegia of Rome, of the provinces, and of the municipalities, but their religious vigour was waning. In the countryside, local agrarian and fertility cults survived, and in the course of the 2nd century AD there was a renaissance of the ancient indigenous cults in the Celtic countries of Gaul and Britain, in Africa, and in Central Europe. Rather than the great gods, who were too "political," the lower classes preferred the more helpful divinities, such as Hercules and Silvanus. Simplistic superstitions still flourished, and even the educated classes did not hesitate to indulge in the magical practices that the law sometimes condemned. The 2nd-century philosopher and rhetorician Apuleius, for example, was brought to trial for having wooed a rich widow by magic arts.

The cult of the emperor had deeply penetrated the traditional religion; and there was a tendency, especially in the East, to make divinities of the emperors while they were still alive. The Severan period was characterized by the triumph of the Oriental religions and by the first efforts of syncretism. For some time now, the Eastern religions had spoken successfully to the uneasy soul of the masses. To be sure, the Greek gods Demeter and Dionysus still had their mysteries and retained their followers. But the cults of the East were aided by their very strangeness, by the fervour of their ceremonies, and by the hope of a future life they reserved to their initiates, who were fraternally united around a well-organized clergy. The oldest established religions were those of Cybele (the Phrygian Great Mother) and of her ancillary god Attis. Proper worship of the Great Mother, however, was open only to the wealthy, for salvation could be obtained only through the sacrifice of the bull. Other long-established deities were the Egyptian gods Isis and Sarapis, whose clergy were particularly honest and fraternal. The cult of Mithra and that of the Syrian gods had expanded more recently, thanks to the perpetual movement of the troops and of the merchants. Mithra, the Sun, representing the supreme god, appealed to the military classes by promising courageous initiates a glorious life in the hereafter. The religion had numerous sanctuaries (*mithraea*) throughout the empire—in the Danubian and Germanic areas, in Britain, and even in Rome. There were large numbers of followers also of the great Syrian Baals: Zeus Dolichenus (from Doliche in northern Syria); Zeus of Heliopolis—Baalbek; and the highland Baal, the El-Gabal (Elagabal) of Emesa. During the Severan era and at their very court, there was developed that syncretism which tended to fuse into a single supreme god all divinities, and especially the solar gods (Sol, Helios, Sarapis, and Mithra). This tendency toward monotheism was accompanied by a lively atmosphere of religiosity and by a taste for magic and miracles.

The Cult of the Great Mother

The rise of Christianity. During the 1st and 2nd centuries, Christianity spread with relative slowness. The doctrines of Jesus, who was crucified about AD 30, first took root among the Jews of Palestine, where a large number of sects were proliferating--orthodox sects, such as the Sadducees and the Pharisees, as well as dissident and sometimes persecuted sects such as the Essenes, whose ascetic practices have been illuminated by the discovery of the Dead Sea Scrolls. At the end of Tiberius' reign, Christianity had spread to the Gentiles as a result of the preaching of St. Paul in Asia Minor and in Greece. At the same time, Christianity continued to make progress among the Jews of Jerusalem, Alexandria, and Syria and quickly reached even Osroene and the Parthian towns of the Euphrates, where Jewish colonies were numerous. The Roman authorities at first had difficulty in distinguishing the "Christos" believers from the orthodox Jews, but the religion of the former, on leaving its original milieu, quickly became differentiated: to the Romans it appeared illicit because it was mysterious; vile, because its first followers belonged to the lower classes; and dangerous, because, unlike Judaism, which was tolerated as an "ethnic" religion, it was without national ties. The 1st-century historian Tacitus declared that there dwelt in Christians a hatred of humankind. Christians did not participate in the Jewish revolt of 66–73, and under the Flavians, Christianity completely severed itself from its origins. At that time, the East was the centre of the new religion, whose followers were multiplying from Egypt to the Black Sea and were beginning to be noticed in Bithynia and in Greece. Christians seemed fairly numerous in Rome as early as the end of the 1st century. When the age of the Apostles ended, the age of the church began, with its bishops, presbyters, and deacons, with its catechism, preaching, and celebration of the Eucharist. In the 2nd century, Christianity began to reach the intellectuals. Hellenistic culture offered educated Christians the resources of philosophical dialectic and of sophist rhetoric. The example of Philo of Alexandria had shown in the 1st century that it was possible to reconcile the Bible with the great Platonic ideas. By the 2nd century the Christian "apologists" tried to show that Christianity

Beginnings of Christian church organization was in harmony with Greco-Roman humanism and that it was intellectually, and above all morally, superior to paganism.

But the Christians did not succeed in convincing the authorities. The first persecution, that of Nero, was related to a particular incident and had no juridical foundation (the fire at Rome in 64). The concept of Christianity as a crime was undoubtedly established under the Flavians, for about 111, Trajan ordered Pliny, the governor of Bithynia, to put to death all those who declared themselves Christians, without the necessity of their having committed any other crime. At the same time, however, he prescribed that no searches be made for them, and this line of conduct was followed by his successors. Thus, the persecutions remained localized and sporadic and were the result of private denunciations or of spontaneous popular protests. Under Marcus Aurelius, the difficulties of the times often caused the Christians, who refused to sacrifice to the state gods and to participate in the imperial cult, to be accused of provoking the wrath of the gods: martyrs appeared in the East, in Rome, in Gaul, and in Africa. Commodus' reign was more favourable to them, perhaps because certain members of his circle, not a very edifying one in other respects, were Christians or Christian sympathizers. This reprieve, however, was short-lived: Septimius Severus inaugurated the first systematic persecution. In 202 an edict forbade Christian (and Jewish) proselytism. Members of extremist sects were persecuted for preaching continence (which violated Augustus' laws against celibacy), for holding the state in contempt, and especially for refusing military service. Under Caracalla, the situation quieted and the church continued to progress, favoured perhaps by the relative freedom which the law granted to funerary collegia (whence the first catacombs).

Cultural life from the Antonines to Constantine. Latin literature enjoyed its "Silver Age" under the Antonines, with the majority of great authors, such as Tacitus, Juvenal, and Pliny the Younger, having begun their careers under Domitian. They had no heirs: after Tacitus, Roman history was reduced to biography. It was only in the 4th century that history began to flourish again, with Ammianus Marcellinus, a Greek writing in Latin. Satire, the Roman genre par excellence, came to an end with Juvenal; and Pliny the Younger, a diligent rhetorician but with a lesser degree of talent, had only the mediocre Fronto as a successor. Another African, Apuleius of Madauros, was more original, being a rhetorician, scholar, and picaresque novelist (Metamorphoses).

Decline of Latin literature

A Greek renaissance, however, took place during the 2nd century. The second sophistic school reigned in every area, in rhetoric, history, philosophy, and even in the sciences. Schools of rhetoric and philosophy prospered in the East, in Smyrna, Ephesus, Pergamum, Rhodes, Alexandria, and even in Athens, protected and subsidized by the emperors, from Vespasian to Marcus Aurelius. The great sophists were Herodes Atticus, a multimillionaire from Athens; Polemon; and Aelius Aristides, a valetudinarian devotee of Asclepius. Dio Cassius and Herodian were conscientious and useful historians (first half of the 3rd century), as was later Dexippus the Athenian, whose work survives only in fragments. Science was represented by the mathematician Nicomachus of Gerasa; medicine, by Galen of Pergamum; astronomy, by the Alexandrian Ptolemy. Law remained the only Roman science, exemplified under the Antonines by Salvius Julianus and Gaius (the *Institutiones*) and later rising to its zenith in the 3rd century as a result of the works of three Greeks: Papinian, Ulpian, and Modestinus. Philosophy, heavily influenced by rhetoric and ethics, was represented under Domitian and Trajan by Dio (or Chrysostom) of Pmsa, who outlined the stoical doctrine of the ideal sovereign. The biographer Plutarch and Lucian of Samosata were more eclectic, especially Lucian, who resembled Voltaire in his caustic skepticism. Under Marcus Aurelius one of Lucian's friends, Celsus, wrote the first serious criticism of Christianity, "The True Word," known through Origen's refutation of it in the 3rd century. At this time philosophy leaned toward religious mysticism: under the Severans, Ammonius Saccas created the school of Alexandria, and his disciple Plotinus founded the Neoplatonist school, which was to fight bitterly against Christianity. After the apologists and, above all, Tertullian (c. 160–after 222), Christian thought deepened, and theology made its appearance. Clement and Origen (*c.* 185–c. 254), the greatest theologian of the time, were the luminaries of the church of Alexandria; the Roman church still wrote in Greek and was represented by the slightly old-fashioned Hippolytus; and the church of Africa had a powerful personality, Saint Cyprian, bishop of Carthage.

The Christian theologians

The disappearance of the great lyric and poetic styles, the fossilizing of education as it came to be completely based on rhetoric (paideia), and the growing importance of philosophical and religious polemical literature among both pagans and Christians were the basic traits that, as early as the 3rd century, foreshadowed the intellectual life of the late empire.

MILITARY ANARCHY AND THE DISINTEGRATION OF THE EMPIRE (235-270)

Succession of emperors and usurpers. The period from the death of Severus Alexander to the time of Claudius II Gothicus was marked by usurpations and barbarian invasions. After Maximinus the Thracian, who bravely fought the Germans but showed great hostility toward the Senate and the educated elite, the Gordians rose to power as a result of a revolt by wealthy African landowners. A senatorial reaction first imposed civilian emperors, Pupienus and Balbinus together, and then named Gordian III, a youth backed by his father-in-law, the praetorian prefect Timesitheus. Gordian III was murdered by the soldiers during a campaign against the Persians and was replaced, first by Philip the Arabian and then by Decius, both soldiers. Decius tried to restore Roman traditions and also persecuted the Christians, but he was killed by the Goths in 251 in a battle near the Black Sea. From 253 to 268 two Roman senators, Valerian and his son Gallienus, reigned. Valerian revived the persecution of the Christians, but he was captured by the Persians during a disastrous campaign and died in captivity (260). His son then reigned alone, facing multiple invasions and several usurpations. He moved constantly between the Rhine and the Danube, achieving brilliant victories (Milan in 262, the Nestus in 267), but the Pannonian army raised several competitors against him (Ingenuus, Regalianus, Aureolus). Too busy to protect the Gauls against the Franks and the Alamanni, and the East against the Persians, he had to tolerate the formation of the Gallic empire under the praetorian prefect Marcus Cassianius Postumus (259–268) and the Palmyrene kingdom of Odaenathus (Odenathus; 260–267). Some of his reforms were a foreshadowing of the future: the senators were practically excluded from the army; the equites received the majority of commands and of provincial governorships; and the composition of the army was modified by the creation of new army corps and especially of a strong cavalry, which was placed under the command of a single leader and charged with closing the breaches that the barbarians were opening along the frontiers. Upon his father's death Gallienus had put an end to the persecution of the Christians, preferring to fight the new religion through intellectual means; to that end, he favoured the ancient Greek cults (Demeter of Eleusis) and protected the Neoplatonist philosopher Plotinus. These initiatives increased the number of his enemies, particularly among the patriotic senators and the Pannonian generals. While Gallienus was in Milan besieging the usurper Aureolus, he was killed by his chiefs of staff, who proclaimed Claudius II (268), the first of the Illyrian emperors. The new emperor won a great victory against the Alamanni on the Garda lake and overwhelmed the Goths in Naissus (269) but died of the plague in 270. This fatal period brought to light one of the major defects of the empire: the lack of a legitimate principle of succession and the preponderant role of the army in politics. The structures that had created the strength of the principate were weakened, and the empire required deep reforms. Gallienus had felt their necessity but had been too weak to impose them.

Victories over the Germans

The barbarian invasions. The Goths were East Germans who came from what is now Sweden and were followed by the Vandals, the Burgundians, and the Gepidae. The after-effect of their march to the southeast, toward the Black Sea, was to push the Marcomanni, the Quadi, and the Sarmatians onto the Roman *limes* in Marcus Aurelius' time. Their presence was brusquely revealed when they attacked the Greek towns on the Black Sea c. 238. Timesitheus fought against them under Gordian III, and under Philip and Decius they besieged the towns of Moesia and Thrace, led by their kings, Ostrogotha and Kniva. Beginning in 253, the Crimean Goths and the Heruli appeared and dared to venture on the seas, ravaging the shores of the Black Sea and the Aegean, as well as several Greek towns. In 267 Athens was taken and plundered despite a strong defense by the historian Dexippus. After the victories of Gallienus on the Nestus and Claudius at Naissus (Nish), there was, for a time, less danger. But the countries of the middle Danube were still under pressure by the Marcomanni, Quadi, Iazyges, Sarmatians, and the Carpi of free Dacia, who were later joined by the Roxolani and the Vandals. In spite of stubborn resistance, Dacia was gradually overwhelmed, and it was abandoned by the Roman troops, though not evacuated officially. When Valerian was captured in AD 259/260, the Pannonians were gravely threatened, and Regalianus, one of the usurpers proclaimed by the Pannonian legions, died fighting the invaders. The defense was concentrated around Sirmium and Siscia-Poetovio, the ancient fortresses that had been restored by Gallienus, and many cities were burned.

Alamanni invasions

In the West the invasions were particularly violent. The Germans and the Gauls were driven back several times by the confederated Frankish tribes of the North Sea coast and by the Alamanni from the Middle and Upper Rhine. Gallienus fought bitterly, concentrating his defense around Mainz and Cologne, but the usurpations in Pannonia prevented him from obtaining any lasting results. In 259–260 the Alamanni came through the Agri Decumates (the territory around the Black Forest), which was now lost for the Romans. Some of the Alamanni headed for Italy across the Alpine passes; others attacked Gaul, devastating the entire eastern part of the country. Passing through the Rhône Valley, they eventually reached the Mediterranean: and some bands even continued into Spain. There they joined the Franks, many of whom had come by ship from the North Sea, after having plundered the western part of Gaul. Sailing up the estuaries of the great rivers, they had reached Spain and then, crossing the Strait of Gibraltar, had proceeded to Mauretania Tingitana. Gallienus, outflanked, entrusted Gaul and his young son Saloninus to Postumus. who then killed Saloninus and proclaimed himself emperor. The several invasions had so frightened the people that the new emperor was readily accepted, even in Spain and Britain. He devoted himself first to the defense of the country and was finally considered a legitimate emperor, having established himself as a rival to Gallienus, who had tried in vain to eliminate him but finally had to tolerate him. Postumus governed with moderation, and, in good Roman fashion, minted excellent coins. He, too, was killed by his soldiers, but he had successors who lasted until 274.

Difficulties in the East. In the East the frontiers had been fixed by Hadrian at the Euphrates. But under Nero, the Romans had claimed control over the kings of Armenia, and under Caracalla they had annexed Osroene and Upper Mesopotamia. The Parthian empire had been weak and often troubled, but the Sāsānids were more dangerous. In 241, Shiipiir I (Sapor), an ambitious organizer and statesman, mounted the throne: he united his empire by bringing the Iranian lords into line and by protecting the Zoroastrian religion. He also tolerated the Manichaeans, put an end to the persecutions of the Christians and Jews, and thereby gained the sympathy of these communities. In 252, with a large army at his command, Shiipiir imposed Artavasdes on Armenia, attacked Mesopotamia, and took Nisibis. In 256 his advance troops entered Cappadocia and Syria and plundered Antioch, while Doura-Europus; on the Middle Euphrates, was likewise falling to him. Valerian had rushed to its aid, but he could not remedy the situation; and in 259 or 260 he was imprisoned by Shiipiir during operations about which little is known. Mesopotamia was lost and Rome was pushed back to the Euphrates. Cappadocia, Cilicia, and Syria were again plundered, and a puppet emperor was appointed in Antioch. But these victories were transitory: in Osroene, Eddessa had shown resistance, a defense was organized in Cappadocia and Cilicia, and Odaenathus, the prince of Palmyra, took Shiipiir by surprise and forced him back to Iran. Having thus aided the Roman cause, Odaenathus then began to act in his own interest: he continued the fight against the Persians and took the title "King of Kings." The Romans officially entrusted him with the defense of the East and conferred on him the governorship of several provinces, so that the "kingdom" of Palmyra thus extended from Cilicia to Arabia. He was murdered in 267 without ever having severed his ties with Gallienus. His widow Zenobia had her husband's titles granted to their son Vaballathus. Then in 270, taking advantage of the deaths of Gallienus and Claudius II, she invaded Egypt and a part of Asia Minor with her troops. This invasion was followed by a rupture with Rome, and in 271 Vaballathus was proclaimed Imperator Caesar Augustus. The latent separatism of the Eastern provinces and, undoubtedly, some commercial advantages caused them to accept Palmyrene domination without difficulty, as they had, in the past, supported Avidius Cassius and Pescennius Niger against the legitimate emperors. In 272 unity was restored by Aurelian, but Mesopotamia was lost and the Euphrates became the new frontier of the empire.

Disruption of commerce by the barbarians

Economic and social crisis. The invasions and the civil wars explain the gravity of the economic and social crisis at the end of Gallienus' reign. Many regions had been laid waste (Gaul, Dacia, Moesia, Thrace, and numerous towns on the Aegean Sea); many important cities had been pillaged or destroyed (Byzantium, Antioch, Olbia, Lyons); and in northern Italy, Gallia Cisalpina (Cisalpine Gaul) itself had been crushed by the Alamanni. The upheavals led to brigandage on land and piracy at sea. Plague, brought back from the East by the armies, ravaged the land for 20 years between 250 and 270. There ensued a population decline, the seriousness of which cannot be estimated; and the empire as a whole was impoverished, although certain provinces, such as Africa, Britain, Pannonia, Asia, and Syria, remained prosperous throughout the century. Municipal government atrophied, inscriptions and dedications became rare, and the cities withdrew into themselves behind narrowed walls. Commerce was often paralyzed by the barbarians; and the main trade route from London to Byzantium, through Boulogne, Cologne, and the Danube, was unsafe and often impassable. The loss of Doura, the defeats inflicted by Shiipiir, and the Palmyrene claims to power slowed down the caravan trade with the East and the Far East. Commerce, however, remained active in the West between Spain, Narbonensis, and Italy, while Africa retained its relations with Ostia, Rome, Egypt, and Syria. Agricultural and handicraft production suffered from insecurity, from military requisitions, and from the abuses of a hard-pressed administration that was attempting by every means to assure the *annona* yield and to provision the large towns so as to prevent outbreaks of social conflict. The empire became a besieged fortress.

The emperors after Commodus—above all Septimius Severus—had been aware of the danger and had adopted such harsh measures as the requisition of material resources and the mobilization of personnel into the service of the state. The municipalities and even the corporations were crushed by the taxes: the collection of the *annona* was a heavy burden for the financially responsible decurions, landholders, and colonists, who were watched over by brutal and sometimes thieving noncommissioned officers. Only a military and bureaucratic monarchy could assure—but at great price certainly—the essential services and obtain an adherence to vital priorities. The early empire's policy of liberalism was at an end; and some modern scholars, such as Rostovtzeff, have spoken of the instituting of "state Socialism"; yet the latter term is fallacious, for there was neither appropriation of the means of production nor redistribution of income to the benefit of the lower classes. Rather, the late empire was a military dictatorship.

Deep social transformations foreshadowed the society of the late empire. The urban middle classes were overburdened, impoverished, and enslaved; and it soon became necessary to halt the flight of the municipal officials by force. The senators lost their political power; but they did preserve their social prestige, and their latifundial fortunes increased at the expense of the decurions and the small landholders. To avoid the obligations of urban life, they tended more and more to retire to their magnificent country villas. The equites thus became the dominant class. They took over the state administration, directed the important offices, controlled finances and taxation, and wielded military and governmental authority in the majority of provinces. Some were civilians, jurists by education; the most powerful came out of the army, where the centurionship had provided open access to high positions since the time of the Severi. During the 3rd century, the army was the empire's rampart against the barbarians, the safest base of imperial power, and the mechanism for social advancement. The lower classes, the humble peasants, the settlers, and the town plebs, officially protected by a state that considered itself egalitarian, suffered from the general impoverishment and the decrease in production: fiscal necessities tended to bind the peasant to his land, the craftsman to his workshop, and the merchant to his trade. Even if the law did not yet prescribe these restrictions, social evolution was moving in this direction, as shown, for example, by the plight of the increasing number of settlers who were forced to cultivate the lands of the rich and of the state.

THE RECOVERY OF THE EMPIRE AND THE ESTABLISHMENT OF THE DOMINATE (270–337)

Aurelian's conquests

The Illyrian emperors. After Claudius II's unexpected death, the empire was ruled from 270 to 284 by several "Illyrian" emperors, who were good generals and who tried in an energetic way to restore equilibrium. The most remarkable was Aurelian. He first gained hard-won victories over the Alamanni and the Juthungi, who had invaded the Alpine provinces and northern Italy. To cheer the inhabitants of Rome, who had succumbed to panic, he began construction of the famous rampart known as Aurelian's Wall. And while crossing the Danubian provinces, before marching against Palmyra, he decided on an orderly evacuation of Dacia, an undefendable region that had been occupied by the barbarians since the time of Gallienus. In the East, he defeated Zenobia's troops easily and occupied Palmyra in 272. Shortly afterward, an uprising broke out in Egypt under the instigation of a rich merchant, who, like a great part of the population, was a partisan of the Palmyrene queen. In response, Aurelian undertook a second campaign, plundering Palmyra and subjugating Alexandria. These troubles, however, along with the devastation of the great caravan city, were to set back Roman trade seriously in the East. Later, rounding back on the Gallic empire of Postumus' successors, he easily defeated Tetricus, a peaceful man and not very willing to fight, near Châlons-sur-Marne. The unity of the empire was restored, and Aurelian celebrated a splendid triumph in Rome. He also re-established discipline in the state, sternly quelled a riot of artisans in the mints of Rome, organized the provisioning of the city by militarizing several corporations (the bakers, the pork merchants), and tried to stop the inflation by minting an *antoninianus* of sounder value. His religious policy was original: in order to strengthen the moral unity of the empire and his own power, he declared himself to be the protégé of the Sol Invictus (the Invincible Sun) and built a magnificent temple for this god with the Palmyrene spoils. Aurelian was also sometimes officially called *dominus et deus:* the principate had definitely been succeeded by the "dominate." In 275 Aurelian was murdered by certain officers who mistakenly believed that their lives were in danger.

His successor, for once, was chosen by the Senate—at the army's request and on short notice. But Tacitus, an aged senator, reigned only for a few months. After him, Probus, another Illyrian general, inherited a fortified empire but had to fight hard in Gaul, where serious invasions occurred in 275–277. Afterward, Probus devoted himself to economic restoration; he attempted to return abandoned farmland to cultivation and, with the aid of military labour, undertook works of improvement. To remedy the depopulation, he admitted to the empire, as had Aurelian, a great number of defeated barbarian Goths, Alamanni, and Franks and permitted them to settle on plots of land in Gaul and in the Danubian provinces. After the assassination of Probus in 282 by soldiers, Cams became emperor and immediately associated with himself his two sons, Carinus and Numerian. Carus and Numerian fought a victorious campaign against the Persians but died under unknown circumstances. Carinus, left behind in the West, was later conquered and killed by Diocletian, who was proclaimed emperor in November 284 by the army of the East.

The tetrarchy

Diocletian. *The tetrarchy.* Diocletian may be considered the real founder of the late empire, though the form of government he established—the tetrarchy, or four persons sharing power simultaneously—was transitory. His reforms, however, lasted longer. Military exigencies, and not the desire to apply a preconceived system, explain the successive nomination of Maximian as Caesar and later as Augustus in 286 and of Constantius and Galerius as Caesars in 293. The tetrarchy was a collegium of emperors comprising two groups: at its head, two Augusti, older men who made the decisions; and in a secondary position, two Caesars, younger, with a more executive role. All four were related either by adoption or by marriage, and all were Illyrians who had attained high commands after a long military career. Of the four, only Diocletian was a statesman. The unity of the empire was safeguarded, despite appearances, for there was no territorial partitioning. Each emperor received troops and a sector of operation: Maximian, Italy and Africa; Constantius, Gaul and Britain; Galerius, the Danubian countries; and Diocletian, the East. Practically all governmental decisions were made by Diocletian, from whom the others had received their power. He legislated, designated consuls, and retained precedence. After 287, he declared his kinship with the god Jupiter (Jove), whom Diocletian claimed was his special protector. Diocletian, together with his Caesar Galerius, formed the "Jovii" dynasty, whereas Maximian and Constantius, claiming descent from the mythical hero Hercules, formed the "Herculii." This "Epiphany of the Tetrarchs" served as the divine foundation of the regime. The ideological recourse to two traditional Roman divinities represented a break with the orientalizing attempts of Elagabalus and Aurelian. Even though he honoured Mithra equally, Diocletian wanted to be seen as continuing the work of Augustus. And in dividing power, Diocletian's aim was to avoid usurpations, or at least to stifle them quickly—as in the attempt of Carausius, chief of the army of Britain, who was killed (293), as was his successor, Allectus (296), after a landing by Constantius.

The deification of the imperial function, marked by elaborate rituals, tended to set the emperors above the rest of mankind. But it was still necessary to avoid future rivalries and to assure the tetrarchy a legitimate and regular succession. Apparently, some time between 300 and 303, Diocletian found an original solution. After the anniversary of their 20-year reign the two Augusti abdicated (Maximian quite unwillingly), and on the same day (May 1, 305) the two Caesars became Augusti. Two new Caesars were chosen, Severus and Maximinus Daia, both friends of Galerius, whose strong personality dominated Constantius. In repudiating the principle of natural heredity (Maximian and Constantius each had an adult son), Diocletian took a great risk: absolute divine monarchy, which Diocletian largely established, implies the hereditary transmission of power, and the future would soon demonstrate the attachment of the troops and even of the population to the hereditary principle.

Changes in provincial organization

Administrative and financial reforms. In order to create a more efficient unity between subjects and administrators, Diocletian multiplied the number of provinces; even Italy was divided into a dozen small units of the provincial type. This put an end to Italy's privileges, particularly since Rome was no longer the effective capital of the empire, each emperor having his own residence in the part of the empire over which he ruled (Trier, Milan, Sirmium, Nicomedia). Although a few provinces were still governed by senators (proconsuls or consuls), the majority were given to equestrian *praesides,* usually without any military power but responsible for the entirety of civil administration (justice, police, finances, and taxes). The cities lost their autonomy, and the curiales administered and collected the taxes under the governor's direct control. The breaking up of the provinces was compensated for by their regrouping into a dozen dioceses, under equestrian vicars who were responsible to the emperor alone. The two praetorian prefects had less military power but played an important role in legislative, judicial, and above all, financial matters: the administration of the *annona,* which had become the basis of the fiscal system, in fact gave them management of the entire economy. Within the central administration the number of offices increased, their managers being civilians who carried out their functions as a regular career. All officials were enrolled in the *militia,* whose hierarchy was to be outlined during the 4th century.

Persian opposition

Great efforts were devoted to strengthening the borders, and the *limes* were outfitted with fortresses *(castella)* and small forts *(burgi),*notably in Syria. The army's strength was increased to 60 legions (but with reduced personnel); and, in principle, each border province received a garrison of two legions complemented by subsidiary troops. Adopting one of Gallienus' ideas, Diocletian created an embryonic tactical army under the direct orders of the emperor whose escort *(comitatus)* it formed. The troops were most often commanded by *duces* and *praepositi,* rather than by provincial governors, and were mainly recruited from among the sons of soldiers and from barbarians who enlisted individually or by whole tribes. In addition, the landowners had to provide either recruits or a corresponding sum of money. All of these reforms were instituted gradually, during defensive wars whose success demonstrated the regime's efficiency. Constantius put down Carausius' attempted usurpation and fought the Alamanni fiercely near Basel; Maximian first hunted down the Bagaudae (gangs of fugitive peasant brigands) in Gaul, then fought the Moorish tribes in Africa, in 296–298, triumphing at Carthage; and on the Danube, Diocletian, and later Galerius, conquered the Bastarnae, the Iazyges, and the Carpi, deporting them in large numbers to the provinces. In the East, however, the opposition of the Persians, led by the enterprising Narses, extended from Egypt to Armenia. The Persians incited uprisings by both the Blemmyes nomads in southern Egypt and the Saracens of the Syrian desert, and made use of anti-Roman propaganda by the Manichaeans and Jews. Diocletian succeeded in putting down the revolt in Egypt and fortified the south against the Blemmyes. But in 297, Narses, the heir to Shāpūr's ambitions, precipitated a war by taking Armenia, Osroene, and part of Syria. After an initial defeat, Galerius won a great victory over Narses, and in 298 the peace of Nisibis reinstated a Roman protege in Armenia and gave the empire a part of Upper Mesopotamia that extended even beyond the Tigris. Peace was thus assured for some decades.

The wars, the reforms, and the increase in the number of officials were costly, and inflation reduced the resources of the state. The *annona,* set up by Septimius Severus, had proved imperfect, and Diocletian now reformed it through the *jugatio-capitatio* system: henceforth, the land tax, paid in kind by all landowners, would be calculated by the assessment of fiscal units based on extent and quality of land, type of crops grown, number of settlers and cattle, and amount of equipment. The fiscal valuation of each piece of property, estimated in *juga* and *capita* (interchangeable terms whose use varied by region and period of time), required a number of declarations and censuses, similar to those practiced long before in Egypt. Each year, the government established the rate of tax per fiscal unit; and every 15 years, beginning in 312, taxes were reassessed. This complicated system was not carried out uniformly in every region. Nevertheless, it resulted in an improved accounting of the empire's resources and a certain progress in fiscal equity, thus making the administration's heavy demands less unbearable. In addition, Diocletian wrshed to reorganize the coinage and stabilize inflation. He thus minted improved sterling coins and fixed their value in relation to a gold standard. Nevertheless, inflation again became disturbing by the end of the century, and Diocletian proclaimed his well-known Edictum de Maximis Pretiis, fixing price ceilings for foodstuffs and for goods and services, which could not be exceeded under pain of death. The edict had indifferent results and was scarcely applied, but the inscriptions revealing it have great economic interest.

Social effects of Diocletian's reforms

Diocletian's reforms adumbrated the principal features of late empire society—the lack of social mobility and the omnipotence of the central administration. Fiscal considerations required that the *annona* be collected regularly and that all land be productive. The tendency was to bind the settlers more and more to their land, to prevent villagers from leaving their homes, and to fix everyone to his place of birth. The curiales and their properties were bound to their cities, the corporation members to their trades, and the sons of soldiers were obliged to take their fathers' places in the army. The upper classes were less tied, having the advantages derived from public careers, whether legally or fraudulently acquired. The senators, now usually called *clarissimi* ("most honorable"), held the major part of landed wealth, and their authority weighed heavily on the peasants. Many of them never attended the Senate in Rome, which had lost its political role. The senatorial class was renewed by the admission of high officials to the clarissimate, including praetorian prefects and numerous provincial governors. The lower appointments were held by the equites, the *perfectissimi* ("the most excellent"), whose place in the social hierarchy depended on their rank in the *militia.* In general, society was becoming hierarchical, and the official who was well paid, both in money and kind, acquired an influence that would only increase during the 4th century.

Persecution of the Christians. After a period of initial indrfference, Diocletian ended his reign by unleashing against the Christians the last and most violent of their persecutions. Decius and Valerian had already dealt severely with them in the middle of the 3rd century and for the first time had directly struck the church's clergy and property by requiring all inhabitants of the empire to sacrifice to the state gods. There were numerous martyrs, but the persecutions came to a quick end upon the death of Decius in 251 and the capture of Valerian in 260. Gallienus published an edict of tolerance, which enabled the church to resume its progress, thanks to the glory of its martyrs and "confessors," the abundance of its donations (for the rich were now also being reached), and the development of the episcopal organization. At the end of the century the new religion was predominant in certain regions of Asia Minor (Cappadocia), well represented in the East, and even powerful in Rome, where the apostle Peter's successors affirmed their episcopal authority. In Italy, Gaul, and Africa around Carthage, the bishoprics were growing in number, and some councils (or synods) assembled dozens of bishops. The progress of a religion that could not accept the religious basis of the tetrarchy, and certain of whose members were imprudent and provocative, as in the incidents at Nicomedia (where a church was built across from Diocletian's palace), finally aroused Galerius' fanaticism. He, in turn, brought pressure on his Augustus: in 303–304 several edicts, each increasingly stringent, ordered the destruction of the churches, seizure of sacred books, and imprisonment of the clergy, and a sentence of death for all those who refused to sacrifice to the Roman gods. In the East, where Galerius was imposing his ideas more and more on the aging Diocletian, the persecution was extremely violent, especially in Egypt, Palestine, and the Danubian regions.

In Italy, Maximian, zealous at the beginning, quickly tired; and in Gaul, Constantius merely destroyed a few churches without carrying reprisals any further. The persecution lasted as late as 312 in the East. Nevertheless, Christianity could no longer be eradicated, for the people of the empire and even some officials no longer felt the blind hatred for Christians that had typified previous centuries.

The seven competing emperors

Struggle for power. The first tetrarchy had ended on May 1, 305; the second did not last long. After Constantius died at Eboracum in 306, the armies of Britain and Gaul, without observing the rules of the tetrarchic system, had hastened to proclaim Constantine, the young son of Constantius, as Augustus. Young Maxentius, the son of Maximian (who had never wanted to retire), thereupon had himself proclaimed in Rome, recalled his father into service, and got rid of Severus. Thus, in 307–308 there was great confusion. Seven emperors had, or pretended to have, the title of Augustus: Maximian, Galerius, Constantine, Maxentius, Maximinus Daia, Licinius (who had been promoted Augustus in 308 by Galerius against Constantine), and, in Africa, the usurper Domitius Alexander also took the purple.

This situation was clarified by successive eliminations. In 310, after numerous intrigues, old Maximian was killed by his son-in-law Constantine, and in the following year Alexander was slain by one of Maxentius' praetorian prefects. In 311 Galerius died of illness a few days after having admitted the failure of his persecutions by proclaiming an edict of tolerance. There remained, in the West, Constantine and Maxentius and in the East, Licinius and Maximinus Daia. Constantine, the best general, invaded Italy with a strong army of faithful Gauls and defeated Maxentius near the Milvian Bridge, not far from Rome. While attempting to escape, Maxentius drowned. Constantine then made an agreement with Licinius, and the two rallied the Eastern Christians to their side by guaranteeing them religious tolerance in the Edict of Milan (313). This left Maximinus Daia, now isolated and regarded as a persecutor, in a weak position; attacked by Licinius near Adrianople, he fell ill and died soon afterward, in 313. This left the empire with two leaders, Constantine and Licinius, allied in outward appearances and now brothers-in-law as a result of Licinius' marriage to Constantine's sister.

The reign of Constantine. Constantine and Licinius soon disputed among themselves for the empire. Constantine attacked his adversary for the first time in 316, taking the dioceses of Pannonia and Moesia from him. A truce between them lasted ten years. In 316 Diocletian died in Salona, which he had never felt a desire to leave despite the downfall of his political creation. Constantine and Licinius then reverted to the principles of heredity, designating three potential Caesars from among their respective sons, all still infants, with the intention of securing their dynasties (two sons of Constantine and one of Licinius). The dynastic concept, however, required only a single emperor, imposing his own descendance. Although Constantine favoured the Christians, Licinius resumed the persecutions; and in 324 war erupted once again. Licinius, defeated first at Adrianople and then in Asia Minor, was obliged to surrender and, together with his son, was executed. Next, Constantine's third son, Constantius, was in turn named Caesar, as his two elder brothers, Crispus and Constantine the Younger, had been some time before. The second Flavian dynasty was thus founded, and Constantine let it be believed that his father, Flavius Constantius (Chlorus), was descended from Claudius Gothicus.

Constantine's conversion to Christianity

Religious policies. Constantine's conversion to Christianity had a far-reaching effect. Like his father, he had originally been a votary of the Sun and had gone to worship at the Grand Temple of the Sun in the Vosges Mountains of Gaul, where he had had his first vision—a pagan one. During his campaign against Maxentius, he had had a second vision—a lighted cross in the sky—and he had painted on his men's shields a figure that was perhaps Christ's monogram (although he probably had Christ confused with the Sun in his manifestation as *summa divinitas* ["the highest divinity"]). After his victory he declared himself Christian. His conversion, as is so often the case, remains somewhat mysterious, and his contemporaries—Lactantius and Eusebius of Caesarea—are scarcely enlightening and even rather contradictory on the subject. But it was doubtless a sincere conversion, for Constantine had a religious turn of mind. He was also progressive and greatly influenced by the capable bishops who surrounded him from the very beginning.

Until 320–322 solar symbols appeared on his monuments and coins, and he was never a great theologian. But his favourable policy toward the Christians never faltered. Christianity was still a minority religion in the empire, especially in the West and in the countryside (and consequently within his own army), thus excluding the possibility of any political calculation on his part. But it was enthusiastically welcomed in the East, and thanks to Constantine the new religion triumphed more rapidly; his official support led to the conversion of numerous pagans, although with doubtful sincerity because they were indifferent in their moral conviction.

The church, so recently persecuted, was now suddenly showered with favours: the construction of magnificent churches (Rome, Constantinople), donations and grants, exemptions from decurial duties for the clergy, juridical competences for the bishops, and exceptional promotions for Christian officials. Pagans were not persecuted, however, and Constantine retained the title of *pontifex maximus.* But he spoke of the pagan gods with contempt and forbade certain types of worship, principally nocturnal sacrifices. In 331 he ordered an inventory of pagan property, despoiled the temples of their treasure, and finally destroyed a few Eastern sanctuaries on the pretext of immorality.

The Arian heresy

The churches were soon to feel the burden of imperial solicitude: the "secular arm" was placed at the service of a changing orthodoxy, for the emperor was impressionable to arguments of various coteries and became quite lost in theological subtleties. In 314 the Council of Arles had tried in vain to stop the Donatist schism (a nationalistic heretical movement questioning the worthiness of certain church officials) that arose in Africa after Diocletian's persecutions. The Arian heresy raised even more difficulties: Arius, an Alexandrian priest and disciple of Lucian of Antioch, questioned the dogma of the Trinity and of the Godhead of Christ, and his asceticism, as well as the sharpness of his dialectics, brought him many followers; he was convicted several times, but the disorders continued. Constantine, solicited by both sides and untroubled by doctrinal nuances that were, moreover, foreign to most Westerners, wished to institute a universal creed; with this in mind he convened the general Council of Nicaea, or Nicene Council, in 325. He condemned Arius and declared, in spite of the Easterners, that Jesus was "of one substance" with God the Father. Nevertheless, the heresy continued to exist, for Constantine changed his mind several times; he was influenced by Arian or semi-Arian bishops and was even baptized on his deathbed, in 337, by one of them, Eusebius of Nicomedia.

Administrative policies. Between 325 and 337 Constantine effected important reforms, continuing Diocletian's work. The division between the *Iimitanei* border troops and the tactical troops (*comitatenses* and imperial guard) led by *magistri militum* was clarified, and military careers became independent of civil careers. He conferred the clarissimate upon many equestrian officials, created new titles that recorded personal fidelity (counts and patricians: companions and relatives of the emperor), and gave first rank in the central administration to the palace questor, the *magister officiorum,* and the counts of finance *(comes sacrarum largitionum, comes rei privatae).* The diocesan vicars were made responsible to the praetorian prefects, whose number was increased and whose jurisdictions were now vast territories: the prefectures of Gaul, Italy, Illyricum, and the East. To this unification of political power there corresponded a decentralization of administration.

In order to reorganize finances and currency he minted two new coins: the silver *miliarensis* and, most importantly, the gold *solidus,* whose stability would make it the Byzantine Empire's basic currency. And by plundering Licinius' treasury and despoiling the pagan temples, he was able to restore the finances of the state. Even so, he still had to create class-taxes: the *gleba* for senators, and the *chrysargyre,* which was levied in gold and silver on merchants and craftsmen in the towns.

Founding of Constantinople

Constantine's immortality, however, rests on his founding of Constantinople. This "New Rome," established in 324 on the site of Byzantium and dedicated in 330, rapidly increased in population as a result of favours granted to immigrants. A large number of churches were also built there, even though former temples were not destroyed; and the city became the administrative capital of the empire, receiving a senate and proconsul. This choice of site was due not to religious considerations, as has been suggested, but rather to reasons that were both strategic (its proximity to the Danube and Euphrates frontiers) and economic (the importance of the straits and of the junction between the great continental road, which went from Boulogne to the Black Sea, and the eastern commercial routes, passing through Asia Minor to Antioch and Alexandria). Constantine died on May 22, 337.

THE ROMAN EMPIRE UNDER THE 4TH-CENTURY SUCCESSORS OF CONSTANTINE

The rule of Constantine's sons. After some months of confusion, Constantine's three surviving sons (Crispus, the eldest son, had been executed in mysterious circumstances in 326), supported by the armies faithful to their father's memory, divided the empire among themselves and had all the other members of their family killed. Constantine II kept the West, Constantius the East, and Constans, the youngest brother, received the central prefecture (Italy, Africa, and Illyricum). In 340, Constantine II tried to take this away from Constans but was killed. For the next ten years there was peace between the two remaining brothers, and Constans won acceptance for a religious policy favourable to the Niceans, whose leader, Athanasius, had received a triumph in Alexandria. In 350 a mutiny broke out in Autun; Constans fled but was killed in Lyons by Magnentius, a usurper who was recognized in Gaul, Africa, and Italy. Constantius went out to engage Magnentius, and the Battle of Mursa (351) left the two strongest armies of the empire, those of Gaul and of the Danube, massacred, thus compromising the empire's defense. Magnentius retreated after his defeat and finally committed suicide in 355.

The reign of Constantius

Thenceforth, Constantius reigned alone as Augustus, aided by a meddlesome bureaucracy in which mission deputies *(agentes in rebus),* informers, and spies played an important role. He named two Caesars in succession, his two young surviving cousins, Gallus in the East and Julian in Gaul, both of whom were apparently considered only "employees," or *apparitores.* Constantius eventually had to get rid of Gallus, who proved incompetent and cruel and soon terrorized Antioch. Julian, however. was a magnificient success, a fact that aroused Constantius' jealousy and led to Julian's usurpation, for the latter was proclaimed Augustus, in spite of Constantius' opposition, at Lutetia in 361. Civil war was averted when Constantius died in November 361, leaving the empire to Julian, the last ruler of the Constantinian family.

Wars against the Persians and barbarians. At the time of his death in 337, Constantine had been preparing to go to war against the Persians. This legacy weighed heavily on the shoulders of Constantius, a military incompetent when compared to the energetic Sāsānian king Shāpūr II. Nearly every year the Persians attacked and pillaged Roman territory; the Mesopotamian towns were beseiged, and Nisibis alone resisted. There was a lull between 350 and 357, while Shāpūr was detained by troubles in the eastern regions of his own kingdom. The war resumed, however, and Mesopotamia was partly lost when the emperor had to go off to fight Julian. Constantius had fought Shāpūr conscientiously, but his generals were mediocre, except for Urisicinus, and he himself was clumsy. In the meantime, the Rhine and the Danube, the troops of which had been withdrawn in favour of the East, were threatened frequently; and Constantius had made a mistake in sending Chnodomar, the Alamannic king, against Magnentius in 351, for his tribes had gone on to ravage Gaul. Julian, however, soon revealed himself to be a great military leader by winning several well-fought campaigns between 356 and 561, most notably at Strasbourg in 357, and by restoring approximately 70 plundered villages. Constantius defeated the Quadi and the Goths on the Danube in 359, but court intrigues, Magnentius' usurpation, and the interminable war against the Persians allowed the barbarians to wreak great havoc.

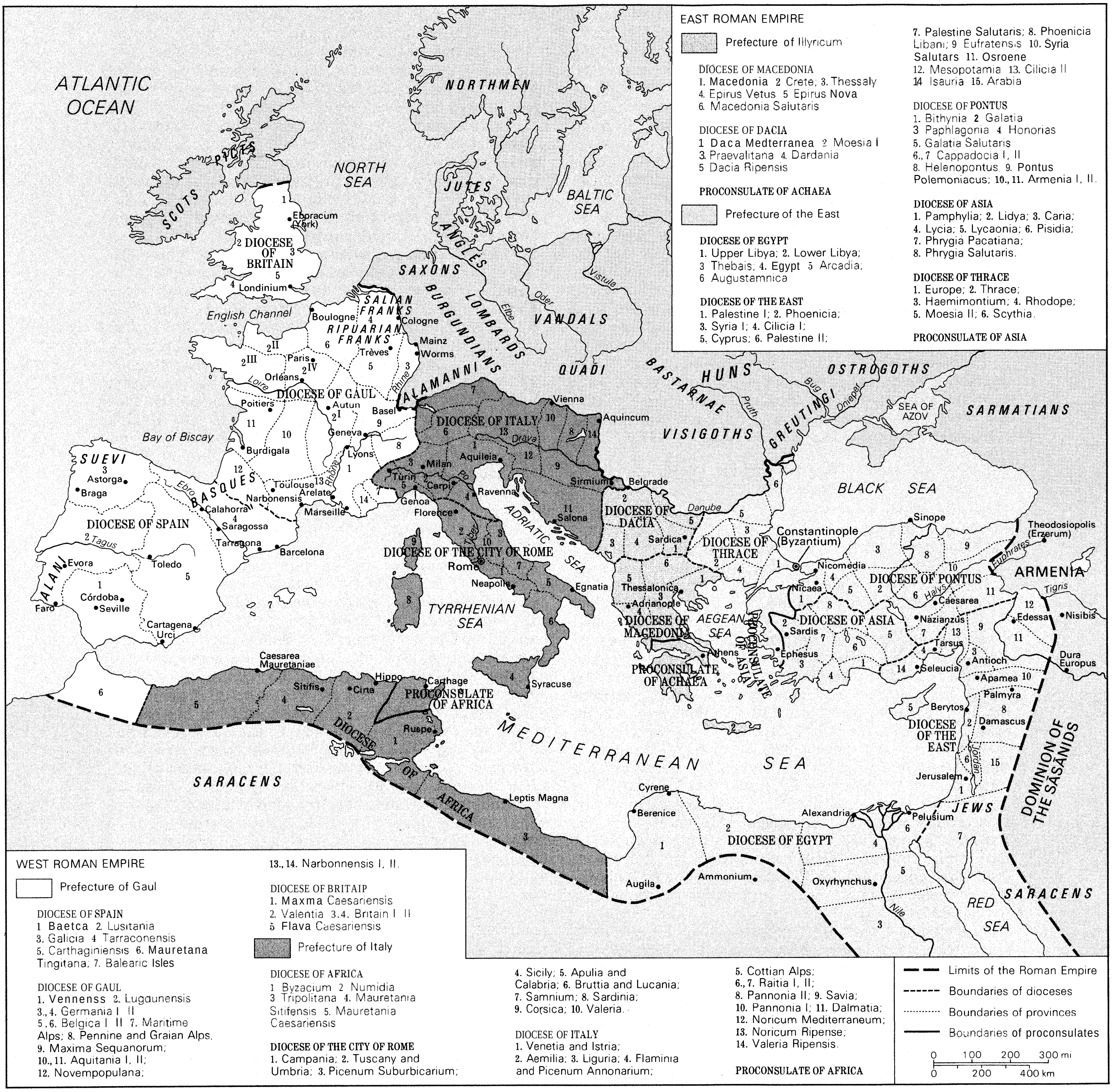
Constantius' Arianism

Constantius' religious policies. Constantius was primarily interested in religious affairs, and his interventions created a "caesaro-papism" that was unfavourable to the church, for after the Battle of Mursa the Emperor had become violently Arian. The Christological problem had moved to the forefront. In 360 Constantius obtained a new creed by force from the Council of Constantinople, which, rejecting the notion of "substance" as too risky, declared only that the Son was like the Father and thus left the problem unresolved. Pagans as well as orthodox Nicaeans (Homoousians) and extremist Arians (Anomoeans) were persecuted, for in 356–357 several edicts proscribed magic, divination, and sacrifices and ordered that the temples be closed. But when Constantius visited Rome in 357, he was so struck by its pagan grandeur that he apparently suspended the application of these measures.

The reign of Julian. Julian, who had been spared because of his tender age from the family butchering in 337, had been brought up far from the court and was undoubtedly intended for the priesthood. Nevertheless, he had been allowed to take courses in rhetoric and philosophy at Ephesus and later at Athens; he developed a fondness for Hellenic literature, and he secretly apostatized around 351. When he became sole emperor at the end of 361, he proclaimed his pagan faith, ordered the restitution of the temples seized under Constantius, and freed all the bishops who had been banished by the Arians, so as to weaken Christianity through the resumption of doctrinal disputes. His paganism was irrational, mystical, and complex. Privately, he worshiped the Sun (Helios-King) and the figure of Cybele, the pagan equivalent of the Virgin Mary. Deeply attached to the traditional Hellenism of the Sophists, who ceaselessly invoked Homer and Plato, he was especially influenced by the last of the Neoplatonists, Iamblichus and Maximus of Ephesus, who instructed him in the techniques of ecstasy and direct communication with the gods (theurgy). The gods spoke to him, and he heard their exhortations. Julian sincerely believed that it was his mission to re-establish their worship.

In time Julian would undoubtedly have become a persecutor, for in 362 he had already forbidden Christians from teaching, under the pretext that their faith made them betray the intentions of the ancient authors. This ostracism aroused even the indignation of certain tolerant pagans. His restoration of the pagan temples provoked disorders and judicial violations among Christians. He gave the pagans a monopoly on high posts, and as a consequence, numerous self-seeking Christians became pagans. Within a short while Julian was successful enough in his undertaking to have aroused the fear and hatred of the Christians, who for a long time thought of him as the Antichrist.

In the political realm, Julian wished to return to the liberal principate of the Antonines—to a time before the reforms of Diocletian and Constantine, whom he detested. He put an end to the terrorism of Constantius' eunuchs and *agentes in rebus* and reduced the personnel and expenditures of the court, while he himself lived like an ascetic. And in the provinces, he lightened the financial burden on individuals by reducing the *capitatio;* that on cities, by reducing the *aurum coronarium* and restoring the municipal properties confiscated by Constantius. On the other hand, he increased the number of curiales by reinstating numerous clerks in an attempt to return the



The Roman Empire in the late 4th century AD.
From W. Shepherd. Historical Atlas, Harper & Row, Publishers (Barnes & Noble Books). New York: revision Copyright © 1964 by Barnes & Noble, Inc.

ancient lustre to municipal life. Thus, he earned the gratitude of pagan intellectuals, who were enamoured of the past of free Greece; and Ammianus made him the central hero of his history.

Julian's attempt to defeat Persia

Taking up Trajan's dream, Julian wished to defeat Persia definitively by engaging the empire's forces in an offensive war that would facilitate a national reconciliation around the gods of paganism. But his army was weak—corrupted perhaps by large numbers of hostile Christians. After a brilliant beginning, he was defeated near Ctesiphon and had to retrace his steps painfully; he was killed in an obscure encounter on June 26,363.

Julian's successor, Jovian, chosen by the army's general staff, was a Christian, but not a fanatic. He negotiated a peace with Shāpūr, by which Rome lost a good part of Galerian's conquests of 298 (including Nisibis, which had not surrendered) and abandoned Armenia. He also restored tolerance in religious affairs, for he neither espoused any of the heresies nor persecuted pagans. In February 364 he died accidentally.

The reign of Valentinian and Valens. Once again the general staff unanimously chose a Pannonian officer—Valentinian, an energetic patriot and, like Jovian, a moderate Christian—but he had to yield to the rivalry of the armies by dividing authority. Taking the West for himself, Valentinian entrusted the East to his brother Valens, an inexperienced man whom he raised to the rank of Augustus. For the first time the two parts of the empire were truly separate, except for the selection of consuls, in which Valentinian had precedence.

Although he served the state with dedication, Valentinian could be brutal, choleric, and authoritarian. His foreign policy was excellent: all the while he was fighting barbarians (the Alamanni in Gaul, the Sarmatians and

Quadi in Pannonia) and putting down revolts in Britain and Africa (notably that of the Berber Firmus) with the aid of his top general, Theodosius the Elder, he was taking care to improve the army's equipment and to protect Gaul by creating a brilliant fortification. His domestic measures favoured the curiales and the lower classes: from then on, taxes would be collected exclusively by officials; the protection of the poor was entrusted to "defenders of the plebs," chosen from among retired high officials (honorati). Nevertheless, needs of state obliged him to accentuate social immobility, to reinforce corporation discipline and official hierarchization, and to demand taxes ruthlessly. At first he was benevolent to the Senate of Rome, supervised the provisioning of the city, and legislated in favour of its university, the nursery of officials (law of 370). But beginning in 369, under the influence of Maximin, the prefect of Gaul, he initiated a period of terror, which struck the great senatorial families. Meanwhile, religious peace reigned in the West, tolerance was proclaimed, and after some difficulty, Rome found a great pope in Damasus, who, beginning in 373, actively supported the new bishop of Milan, St. Ambrose, an ardent defender of orthodoxy.

Religious peace in the West, persecutions in the East

In the East, Valens, who was incapable and suspicious, had fallen under the influence of legists, such as the praetorian prefect Modestus. The beginning of Valens' reign was shadowed by the attempted usurpation of Procopius (365–366), a pagan relative of Julian's who failed and was killed by the army, which remained faithful to Valens. Modestus instituted harsh persecutions in Antioch of the educated pagan elite. Valens was a fanatic Arian, who exiled even moderate Nicaean bishops and granted to Arians favours that aroused violent reactions from the orthodox, whose power had increased in the East. Valens' policies made the East prey to violent religious passions.

On the Danube, Valens fought the Visigoths and made a treaty with their king, Athanaric, in 369; but in 375 the Ostrogoths and the Greutingi appeared on the frontiers, pushed from their home in southern Russia by the powerful Huns. In 376 Valens authorized the starving masses to enter Thrace; but being exploited and mistreated by the officials, they soon turned to uncontrollable pillaging. Their numbers continually increased by the addition of new bands, until finally they threatened Constantinople itself. Valens sent for aid from the West; but without waiting for it to arrive, he joined battle and was killed in the Adrianople disaster of 378, which to some critics foreshadowed the approaching fall of the Roman Empire.

The Goths, who were also stirring up Thrace and Macedonia, could no longer be driven out. The provinces remained insecure until 382, when Theodosius, after fighting many hard battles with only an inconclusive result, resigned himself to settling the Goths legally in the Balkan Peninsula as foederati (soldiers in the service of Rome, receiving subsidies and furnishing recruits); there they formed barbarian enclaves.

The reign of Gratian and Theodosius I. Following Valentinian's sudden death in 375, the West was governed by his son Gratian, then 16 years old, who had been given the title of Augustus as early as 367. The Pannonian army, rife with intrigue, quickly proclaimed Gratian's half-brother, Valentinian II, only four years old. The latter received Illyricum under his older brother's guardianship, and this arrangement satisfied everybody. Valentinian's advisers were executed: Maximian was sacrificed to the spite of the Senate, and Theodosius the Elder became the victim of personal jealousies. Gratian announced a liberal principate, supported in Gaul by the wealthy family of the Bordeaux poet Ausonius and in Rome by the Symmachi and the Nicomachi Flaviani, representatives of the pagan aristocracy. His generals defeated the Alamanni and the Goths on the Danube but arrived too late to save Valens.

On January 19, 379, before the army, Gratian proclaimed Theodosius, the son of the recently executed general, as Eastern emperor. Theodosius was chosen for his military ability and for his orthodoxy (Gratian, extremely pious, had come under the influence of Damasus and Ambrose). The East was enlarged by the dioceses of Dacia and Macedonia, taken from Valentinian II. Gratian and Theodosius agreed to admit the Goths into the empire, and Gratian applied the policy also to the Salian Franks in Germany. Theodosius soon dominated his weak colleague and entered the battle for the triumph of orthodoxy. In 380 the Arians were relieved of their churches in Constantinople, and in 381, the Nicaean faith was universally imposed by a council whose canons established the authority of the metropolitan bishops over their dioceses and gave the bishop of the capital a primacy similar to that of the bishop of Rome.

Influx of Goths and Franks

In ecclesiastical affairs, the separation between East and West was codified. The Westerners bowed to this policy, satisfied with the triumph of orthodoxy. Gratian then permitted Ambrose and Damasus to deal harshly with the Arians, with the support of the state. Paganism also was hounded: following Theodosius' lead. Gratian refused the chief priesthood, removed the altar of Victory from the hall of the Roman Senate, and deprived the pagan priests and the vestals of their subsidies and privileges. The pagan senators were outraged, but their protests were futile because Gratian was watched over by Arnbrose.

This militantly orthodox policy aroused the displeasure of the pagans and of the Western Arians: thus, when Gratian left Trier for Milan, the army of Gaul and Britain proclaimed its leader, Maximus, in 383. He conquered Gaul without difficulty, and Gratian was killed in Lyons. Maximus, who, like Theodosius, was Spanish and extremely orthodox, was recognized by the latter. In the meantime, the third Augustus, Valentinian II, had taken refuge in Milan after suffering defeat in Pannonia. He was effectively under the domination of his mother, Justina, an Arian who sought support for her son among the Arians and pagans of Rome and even among the African Donatists. In 388 Maximus, after arriving in Italy, first expelled Valentinian and then prepared to attack Theodosius. The latter, accepting the inevitability of war, strengthened his resolve and gained several victories. Maximus was killed at Aquileia in 388, and thenceforth Theodosius ruled both West and East; he was represented in the East by his son Arcadius, an Augustus since 383. Valentinian II was sent to Trier, accompanied by the Frankish general Arbogast to control him.

Sole rulership of Theodosius

After a few years' respite, during the prefectureships of Nicomachus Flavianus in Rome and Tatian in the East, paganism waged its last fight: Theodosius, influenced by Ambrose, who had dared to inflict public penance on him in 390 after the massacre at Tnessalonica, had determined to eliminate the pagans completely. After a few hostile clashes, the law of November 8, 392, proscribed the pagan religion. Then Arbogast, after Valentinian II's death in 392 under shadowy circumstances, proclaimed as emperor the rhetorician Eugenius. When Theodosius refused to recognize him, Eugenius was thrown into the arms of the pagans of Rome. But this last "pagan reaction" was short-lived; in 394, with his victory at the Frigidus River, between Aquileia and Emona, Theodosius put an end to the hopes of Eugenius and his followers. His intention was to place his son Honorius, proclaimed Augustus in 393, over the West, while returning his eldest son, Arcadius, to the East. But Theodosius' sudden death, in January 395, precipitated the division of the empire.

Theodosius had finally eliminated the clanger of the Goths, although not without taking risks, and had both established a dynasty and imposed the strictest orthodoxy. A compromise peace with the Persians had given Rome, in 387, a small section of Armenia, where he had founded Theodosiopolis (Erzurum). Internally, there were few reforms; and wars, wastefulness, and the burden of administration had caused a noticeable increase in taxation. The ancient authors all paint a gloomy picture of the condition of the settlers, the curiales, and the artisans of the towns. Meanwhile, the laws were becoming increasingly harsh, and revolts (Antioch in 387, Thessalonica in 390) were ruthlessly suppressed. Theodosius owed his success to the support of the orthodox bishops —who took advantage of his faith to introduce “clericalism" into the state—and to the strength of his army, hereafter dominated by barbarian elements: the Franks and Alamanni in the West; the Goths and even the federated Huns in the East. The majority of commanding officers were also barbarians, including such Franks as Richomer, Merovech, Arbogast, and the half-Vandal Stilicho, who through his marriage to Serena, Theodosius' niece, had entered the imperial family.

Social and economic conditions. During the 4th century the emperor's power was theoretically absolute, the traditions of the principate having given way to the necessities of defense.

Hierarchized bureaucracy

Administration. The emperor, who was most often a soldier, was both heir to the Hellenistic basileus (absolute king) and the anointed of the deity. Pagans and Christians alike considered him "emperor by the grace of God," which, strictly speaking, rendered the imperial cult unnecessary. Indeed, only military victory, willed by God, distinguished the legitimate sovereign from the "tyrant." After Constantine, administration was decentralized and headed by praetorian prefects with enormous power. A gargantuan and strictly hierarchic bureaucracy restricted direct action by the emperor, who was separated increasingly from his subjects by a hierarchical court, particularly under Constantius and Theodosius. This overrun administration was recruited from among students of literature and of law for higher posts, and for lesser posts from among "notaries," who knew how to write shorthand. Under Constantius the latter attained even the highest positions, in spite of their modest origins. There was superabundant legislation as well, which tended to increase fiscal income by limiting exemptions and privileges, and to keep each person in the rank assigned him at birth, for taxes were based on social classes. Tax abuses provoked uprisings by the lower classes as well as recourse to patronage and attempts to escape one's class by entering the administration and, above all, to get into the clarissimate, which was lucrative and rich in privileges.

Apart from the tiny elite of Roman senators, for whom public office was only a last resort, and the great landholders sheltered in their villas, where in the West at least they were able to live in independence, officials everywhere formed the upper class of society. The equites occupied the inferior and beginning positions, for gradually during the course of the century all the high officials attained the clarissimate. A career in public service could no longer be compared to the old magistracies. Public service was a powerful factor in social advancement (like the army in the 3rd century), but it was also a costly and sometimes corrupt superstructure, whose upkeep drained the empire. To protect themselves, the poor and weak took refuge in the patronage (*patrocinium*) of the powerful (potentes), a method which, while it shielded them from the authority of the state in spite of the laws, also placed them in positions of personal dependence, which foreshadowed the social arrangements of the Middle Ages. The bishops, who held spiritual authority over growing portions of the population and had available to them both the power of their office and large revenues, began to concentrate the defense of local interests in their own hands and to compete with the rich for the protection of the weak. In the 5th century, they would be the masters and protectors of the cities, particularly in the West.

Settler system

Economy. Economic life was based on ownership of land, and the number of free peasants and middle-sized landowners was decreasing, for both groups were burdened by heavy taxes and threatened by the expansionism of the rich estate (latifundium) owners, the senators and high functionaries. Some free villages (metrocomiae) continued to exist, notably in Syria, and were protected by the state. In areas where urbanization was sparse, the latifundium dominated and was worked by settlers. The udscripticius settler was subordinated completely to his master (*dominus*) and owed him forced labour, whose name (obsequia servilia) showed its true nature. The tributarius settler was a freeholder but still had to look to the patronage of the powerful man whose land he cultivated. The laws specified that the settler could not leave his land, and a recaptured fugitive was treated like a slave, even though he remained a free man in principle. This evolution toward serfdom was ancient; but it had been accelerated by the fiscal reforms (the jugatio, the taxes in kind, and the increases in amounts of forced labour due). In sum, the settler system was one aspect of the general social immobility of the late empire, and it resulted from the enormous needs of the state.

The economy was of three types, which were unequally distributed: 1) the closed economy of the large estates, which dominated in the few regions where urbanization was weak; 2) the produce state economy, in which taxes, wages, and salaries were paid in kind so as to reduce the disadvantages of inflation for the treasury—but the possibility, often resorted to, of replacing payment in kind with payment in money (adaeratio) probably allowed the state to speculate on monetary fluctuations and to satisfy one or the other social category; 3) the monetary economy, the most widespread, which had been restored by Constantine: the gold standard (the *solidus,* or gold pound) was used to establish prices, to pay certain taxes and services, and to assist in the amassing of wealth (coins, bars, ingots); the regular currency, the *nummus,* was a silver-covered copper coin, having various names and values.

Although there was considerable inflation (culminating under Theodosius), in spite of a deflationary fiscal policy, commercial transactions ignored barter and were based instead on currency throughout the empire at the end of the century. The economy was partially under state direction, which was applied to agriculture through bias toward the settler system on imperial estates and to industry through the requisitioning of corporations (artisans, merchants, carriers) and the creation of state workshops (especially for manufacturing military goods). Opinions differ on the intensity of trade, but there was certainly clear progress in comparison to the 3rd century.

Comparison between East and West

It is above all necessary to distinguish between the East and West. The West was in decline: the rural areas had taken over from the cities, where activity had become anemic; and autonomous blocs had formed in Gaul (with the relative prosperity of Britain), Africa, Italy (which suffered a marked decline, except in Cisalpina), and the Danubian countries. The great West–East route of the early empire (from Boulogne to Byzantium, by way of the Rhine and the Danube), as a result of threats by barbarians, had been displaced toward the south, through Lyons, Milan, and Aquileia. On the other hand, the East was less affected by invasions; there, urban life was traditionally more active, and the peasantry resisted the settler system and patronage more effectively. Although the Balkans were devastated after Adrianople (378), elsewhere the wealth of the towns and of the officials sustained active trading. There were three great commercial arteries: 1) the "wheat" road, from Alexandria to Constantinople by way of Antioch; 2) the Far Eastern trade routes, through Osroene (Batnae) or Cappadocia (Somoata) to Antioch and the capital; and 3) the maritime route linking East to West, from Seleucia (the port of Antioch) to Rome. In addition, there were the more dangerous land routes, such as the one across the Balkans (the ancient Via Egnatia through Macedonia) or the Danube route (Sardica, Sirmium, Aquincum, Carnuntum). Although Egypt was impoverished from having been exploited for so long, Asia Minor around Constantinople was prosperous, having numerous towns as well as well-populated rural areas. And of the Eastern provinces, Syria was the richest—thanks to a balanced agriculture, a large and stable population, the industries of Antioch, the textiles of Apamea, Damascus, and Berytus, and above all thanks to commerce, which was centred in Antioch and its port, Seleucia, rebuilt by Constantius. All of this accounts for the intensity of urbanization.

The remnants of pagan culture. The spread of Christianity in no way harmed the flourishing of pagan literature. Instruction in the universities (Rome, Milan, Carthage, Bordeaux, Athens, Constantinople, Antioch, Alexandria) was still based on rhetoric, and literature received the support of senatorial circles, especially in Rome (for example those of the Symmachi and the Nicomachi Flaviani). Latin literature was represented by Symmachus and the poet Ausonius. The last great historian of Rome was Ammianus Marcellinus, a Greek who wrote in Latin for the Roman aristocracy; of his *Res gestae*, the most completely preserved part describes the period from 353 to 378. The works of the abridgers Aurelius Victor and Eutropius are fairly accurate and more reliable than the *Scriptores kistoriae Augustae*, a collection of imperial biographies of unequal value, undoubtedly composed under Theodosius but for an unknown purpose. Erudition was greatly prized in aristocratic circles, which, enamoured of the past, studied and commented on the classic authors (Virgil) or the ancestral rites (the *Saturnalia* of Macrobius). Greek literature is represented by the works of philosophers or sophists: Themistius, a political theoretician who advocated absolutism; Himerius of Prusias; and above all Libanius of Antioch, whose correspondence and political discourses from the Theodosian period bears witness to his perspicacity and, often, to his courage.

The Christian Church. From the time of Constantine the church had perfected its organization, modelled on that of the state. Each city had its bishop; each provincial capital had its Metropolitan, and some had a recognized "primacy": the patriarchs of Alexandria, Antioch, and Constantinople and the bishops of Milan and Rome, the latter claiming an *auctoritas* over the others, which was a heritage of classic Rome. The church had become extremely wealthy and had obtained immunities for its clergy and its bishops. The emperors had assumed the right to interfere in religious affairs, but under Theodosius, Pope Damasus and St. Ambrose reacted: the state was to restrict itself to furnishing the "secular arm," while the church, in the name of evangelical ethics, claimed the right to judge the emperors, a policy that had grave implications for the future. The "caesaro-papism" of Constantius would gain adherents under the Byzantine emperors. Paganism, officially suppressed by Theodosius, survived well into the 5th century in the rural areas, where missionary activity (such as that of St. Martin in Gaul) was only recent, and among the intellectuals, teachers, and senators of Rome. Orthodoxy triumphed over the great heresies (Donatism, Arianism), but other heresies continued to emerge. In the meantime, the Goths had been converted to Arianism by Ulfilas during the period of Constantius and Valens, thus presaging conflicts that would come after the great invasions. Orthodox missionaries had converted Osroene, Armenia, and even some countries on the Red Sea. And in Egypt at the end of the 3rd century there appeared the monasticism of isolated hermits (St. Anthony) and of cenobites grouped in communities (Pachomius). Monasticism rapidly spread to Syria and Asia Minor by toning down its ascetic severity, and in the West St. Martin founded the monasteries of Ligugé and Marmoutier.

Triumph of orthodoxy

The Christian literature of the 4th century is remarkable. Its first representative is Eusebius of Caesarea, a friend and panegyrist of Constantine and a church historian whose creation of a "political theology" sealed the union between the Christian emperor and the church. St. Athanasius wrote apologetic works and a life of St. Anthony. Also prominent were the great Cappadocians: St. Basil of Caesarea, St. Gregory of Nazianms, St. Gregory of Nyssa, and St. John Chrysostom of Antioch, the greatest preacher of his time. The Westerners, too, had great scholars and brilliant writers: St. Hilary of Poitiers, enemy of the Arians and of Constantius; St. Ambrose, administrator and pastor, whose excessive authority was imposed on Gratian and even on Theodosius; and St. Jerome, a desert monk and confessor of upper class Roman ladies, a formidable polemicist who knew Greek and Hebrew and made the first faithful translation of the Old and New Testaments (the Vulgate) as well as of a chronicle of world history, which was a translation and continuation of the work of Eusebius. Finally, St. Augustine, the bishop of Hippo, was a great pastor, a vigorous controversialist, a sensitive and passionate writer (the *Confessions)*, and the powerful theologian of *The City of God*. The century that developed these great minds cannot be considered decadent.

THE ECLIPSE OF THE ROMAN EMPIRE IN THE WEST (C. 395-500) AND THE GERMAN MIGRATIONS

Invasions in the early 5th century. After the death of Theodosius the Western empire was governed by young Honorius. Stilicho, an experienced statesman and general, was charged with assisting him and maintaining unity with the East, which had been entrusted to Arcadius. The Eastern leaders soon rejected Stilicho's tutelage. An anti-barbarian reaction had developed in Constantinople, thus impeding the objectives of the half-Vandal Stilicho. He wanted to intervene on several occasions but was prevented from doing so by a threat from the Visigoth chieftain Alaric, whom he checked at Pollentia in 402, then by the Ostrogoth Radagaisus' raid in 406, and finally by the great invasion of the Gauls in 407. The following year he hoped to restore unity by installing a new emperor in Constantinople, Theodosius II, the son of Arcadius, who had died prematurely; but he succumbed to a political and military plot in August 408. The division of the two *partcs imperii* was now a permanent one.

Honorius. seated in Ravenna, a city easier to defend than Milan, had only incompetent courtiers surrounding him, themselves animated by a violent hatred of the barbarians. Alaric soon reappeared, at the head of his Visigoths, demanding land and money. Tired of the Romans' double-dealing, he descended on Rome itself. The city was taken and pillaged for three days, thus putting an end to an era of Western history (August 410). An Arian, Alaric spared the churches. He died shortly thereafter in the south; and his successor, Athaulf, left the peninsula to march against the Gauls.

Sack of Rome by the Visigoths

Fleeing from the terrifying advance of the Huns, on December 31, 406, the Vandals, Suevi, and Alani, immediately followed by the Burgundians and bands of Alamanni, crossed the frozen Rhine and swept through Gaul, effortlessly throwing back the federated Franks and Alamanni from the frontiers. Between 409 and 415 a great many of these barbarians arrived in Spain and settled in Lusitania (Suevi) and in Baetica (Vandals, whence the name Andalusia). As soon as Gaul had become slightly more peaceful, Athaulf's Visigoths arrived, establishing themselves in Narbonensis and Aquitania. After recognizing them as "federates," Honorius asked them to go to Spain to fight the Vandals. Meanwhile, the Roman general Constantius eliminated several usurpers in Gaul, confined the Goths in Aquitania, and reorganized the administration (the Gallic assembly of 418). But he was unable to expel the Franks, the Alamanni, and the Burgundians, who had occupied the northern part of the country, nor to eliminate the brigandage of the Bagaudae. He was associated with the empire and was proclaimed Augustus in 421, but he died shortly afterward. His son, Valentinian III, succeeded Honorius in 423 and reigned until 455.

The beginning of Germanic hegemony in the West. During the first half of the 5th century the barbarians gradually installed themselves, in spite of the efforts of the Roman general Aetius at the head of a small army of mercenaries and of Huns. Aetius took back Arles and Narbonne from the Visigoths in 436, either pushed back the Salian and Ripuarian Franks beyond the Rhine or incorporated them as federates, settled the defeated Burgundians in Sapaudia (Savoy), and established the Alani in Orléans. The other provinces were lost: Britain, having been abandoned in 407 and already invaded by the Picts and Scots, fell to the Angles, Saxons, and Jutes; a great Suevi kingdom, officially federated but in fact independent, was organized in Spain after the departure of the Vandals, and it allied itself to the Visigoths of Theodoric I, who were settled in the country around the Garonne.

Loss of provinces

The Vandal seizure of North Africa. In 428 the Vandal Gaiseric led his people (80,000 persons, including 15,000 warriors) to Africa. St. Augustine died in 430 in besieged Hippo, Carthage fell in 435, and in 442 a treaty gave Gaiseric the rich provinces of Byzacena and Numidia. From there he was able to starve Rome, threaten Sicily, and close off the western basin of the Mediterranean to the Byzantines.

Invasion by the Huns. Shortly afterward, in 450, Attila's Huns invaded the West—first Gaul, where, after having been kept out of Paris, they were defeated by Aetius on the Campus Mauriacus (near Troyes), then Italy, which they evacuated soon after having received tribute from the pope, St. Leo. Attila died shortly afterward; and this invasion, which indeed left more legendary memories than actual ruins, had shown that a solidarity had been created between the Gallo-Romans and their barbarian occupiers, for the Franks, the Alamanni, and even Theodoric's Visigoths had come to Aetius' aid.

The end of the Roman Empire of the West

The last emperors in the West. After the death of Aetius, in 454, and of Valentinian III, in 455, the West became the stake in the intrigues of the German chiefs Ricimer, Orestes, and Odoacer, who maintained real power through puppet emperors. In 457–461 the energetic Majorian re-established imperial authority in southern Gaul until he was defeated by Gaiseric and assassinated shortly afterward. Finally, in 476, Odoacer deposed the last emperor, Romulus Augustulus, had himself proclaimed king in the barbaric fashion, and governed Italy with moderation under the theoretical tutelage of the emperor of the East. The end of the Roman Empire of the West passed almost unperceived.

Barbarian kingdoms. Several barbarian kingdoms were then set up: in Africa, Gaiseric's Vandals; In Spain and in Gaul as far as the Loire, the Visigothic kingdom; and farther to the north, the Salian Franks and the Alamanni. The barbarians were everywhere a small minority. They established themselves on the great estates and divided the land to the benefit of the federates without doing much harm to the lower classes or disturbing the economy. The old inhabitants lived under Roman law, while the barbarians kept their own "personality of laws," of which the most well-known is the judicial composition, the Wergild. Romans and barbarians coexisted but uneasily. Among the obstacles to reconciliation were differences in mores; social and political institutions (personal monarchies, fidelity of man to man); language, although Latin was still used in administration; and above all religion: the Arianism of the barbarians permitted the Catholic bishops to retain their hold over their flocks. The only persecution, however, was under the Vandals, whose domination was the most harsh.

Two great kingdoms marked the end of the 5th century. In Gaul, Clovis, the king of the Salian Franks (reigned 481/482–511), expelled Syagrius, the last Roman, from Soissons, took Alsace and the Palatinate from the Alamanni (496), and killed Alaric II, king of the Visigoths, at Vouillé (507). His apparently sudden conversion to Catholicism assured him the support of the bishops, and Frankish domination was established in Gaul. At the same time, Theodoric, king of the Ostrogoths, was reigning in Italy. He had been charged by the emperor Zeno to take back Italy from Odoacer in 488, and in 494 he had himself proclaimed king at Ravenna. His Goths, few in number, were established in the north; elsewhere he preserved the old imperial administration, with senators as prefects. Externally, he kept Clovis from reaching the Mediterranean and extended his state up to the valley of the Rhône. Theodoric died in 526. Ten years later Justinian charged his general Belisarius with the reconquest of Italy, a costly, devastating, and temporary operation that lasted from 535 to 540.

Analysis **of** the decline **and** fall. The causes of the fall of the empire have aroused a good deal of discussion. The reasons most often invoked are not definitive; the economic decline of the Roman world was not obvious in the 4th century, nor was the decay of the form of government or the defects of the administration. Nevertheless, the empire was yielding under the weight of its state superstructure and of taxation that was too heavy for the ancient world, where production and the standard of living were both very low. The army, which was extremely expensive, was not large enough; and even the population seemed to have declined. But it was really the massive invasions, which had become incessant during the second half ,of the century, that toppled the empire. It is essential to see why the East resisted, whereas the West was overcome. The East during this period was rich and well populated, and its heavy state structure was healthier than that of the West. The latter was in the hands of the great landholders, who exploited their settlers, who, in turn, had no recourse. The East was also less threatened from the outside and easier to defend: only the lower Danube was vulnerable, with Persia in fact representing no great danger, whereas in the West long frontiers required large numbers of troops and ruinously expensive fortifications. Once the borders were broken through, the vital regions of Gaul and Italy were left defenseless. Even if the East had come to its aid instead of growing increasingly aloof after the failure of Stilicho, the barbarians would undoubtedly still have triumphed; for the population, although it dreaded the barbarians and hated their religion, showed a surprising apathy: bureaucratic absolutism and the patronage of the powerful in the West had killed the idea of the state and the patriotism of ancient Rome. **(P.P.)**

BIBLIOGRAPHY

Origins to 264 BC: P.G. GIEROW, *The Iron Age Culture of Latium:* vol. 1, *Classification and Analysis* and vol. 2, *Excavations and Finds* (1964), a systematic view of the archaeological discoveries made in Latium, dating from the early Iron Age; EINAR GJERSTAD, *Early Ronze: vol. 1, Stratigraphical Researches in the Forum Romanum and Along the Sacra Via* (1953); vol. 2, *The Tombs* (1956); vol. 3, *Fortifications, Domestic Architecture, Sanctuaries, Stratigraphical Excavations* (1960); vol. 4, *Synthesis of Archaeological Evidence* (1966), a monumental work on all the archaeological remains from ancient Rome—the investigation of the material is outstanding, although the historical conclusions have been contested; PIETRO DE FRANCISCI, *Primordia Civitatis* (1959), and RAYMOND BLOCH, *Les Origines de Rome,* rev. ed. (1959; Eng. trans., *The Origins of Rome,* 1960), two overall views of the early centuries of Rome internally and externally; ANDRAS ALFOLDI, *Early Rome and the Latins* (1965), a first-rate study of the relations between Rome and the Latins, up to the time of Rome's subjugation of Latium; T.R.S. BROUGHTON, *The Magistrates of the Roman Republic,* 2 vol., and one suppl, vol. (1951–52), an excellent view of the Roman magistracies of the republican era; PIERRE LEVEQUE, *Pyrrhos* (1957), a monograph (in French) that deals with the personality and ambitions of Pyrrhus, Rome's formidable opponent at the beginning of the 3rd century Bc; RUDI THOMSEN, *Early Roman Coinage,* 3 vol. (1957–61), a balanced, well-done history of the oldest Roman currency.

Rome—264 to 133 BC: Important sections, together with inscriptions, are published in English translation by NAPHTALI LEWIS and MEYER REINHOLD (eds.), *Roman Civilization Sourcebook,* vol. *1, The Republic* (1966); and A.H.M. JONEs, *A History of Rome Through the Fifth Century,* vol. 1, *The Republic* (1968). *(General histories of the period): The Cambridge Ancient History,* vol. 7–8 (1928–30); G. DE SANCTIS, *Storia dei Romani,* 2nd ed., vol. 3–4 (1967–69); H.H. SCULLARD, *A History of the Roman World from 753 to 146 B.C.,* 3rd ed. (1969). *(Roman constitutional history):* A.H.J. GREENIDGE, *Roman Public Life* (1901); W.M.F. JASHEMSKI, *The Origins and History of the Pro-Consular and Pro-Praetorian Imperium to 27 B.C.* (1950); THEODOR MOMMSEN, *Römische Staatsrecht,* 3 vol. (1871–88, reprinted 1952); E.T. SALMON, *Roman Colonization Under the Republic* (1969); L. ROSS TAYLOR, *Roman Voting Assemblies* (1966) and *The Voting Districts of the Roman Republic* (1960); A.N. SHERWIN WHITE, *The Roman Citizenship* (1939). *(Social and economic history):* TENNEY FRANK, *An Economic History of Ronze fo the End of the Republic,* 2nd ed. (1927), and (ed.), *An Economic Survey of Ancient Rome,* 6 vol. (1933–40); M.I. ROSTOVTZEFF, *The Social and Economic History of the Hellenistic World* (1941); ARNOLD TOYNBEE, *Hannibal's Legacy,* 2 vol. (1965). *(Provincial government and imperialism):* E. BADIAN, *Foreign Clientelae, 264–70 B.C.* (1958); TENNY FRANK, *Roman Imperialism* (1914); DAVID MAGIE, *Roman Rule in Asia Minor to the End of the Third Century After Christ, 2* vol. (1950); G.H. STEVENSON, *Roman Provincial Administration till the Age of the Antonines* (1939); C.H.V. SUTHERLAND, *The Romans in Spain, 217 B.C.–A.D. 177* (1939).

Rorne—133 to 31 BC: The chief ancient sources for the period are the works of CAESAR and CICERO and SALLUST'S two monographs on Jugurtha and Catiline; the biographies of prominent Romans of the period in PLUTARCH, *Parallel*

Lives; and the historical narratives of APPIAN *(The Civil Wars)* and of CASSIUS DIO *(Roman History).* All are available in the original, with facing English translation, in the "Loeb Classical Library," the first four in many other translations. An important source collection for the period it covers is A.H.J. GREENIDGE and A.M. CLAY, *Sources for Roman History,* 133–70 *B.C.,* 2nd ed. rev. by E.W. GRAY (1960). The best outline in English is the first half of H.H. SCULLARD, *From the Gracchi to Nero,* 3rd ed. (1970), with excellent notes and bibliography. The classic reference work is WILHELM DRUMANN, *Geschichte Roms in seinem Übergange von der republikanischen zur monarchischen Verfassung,* 2nd ed. by P. GROEBE, 6 vol. (1899), giving biographies (with full source material) of all prominent figures of the period, arranged by families. For a survey of scholarship, with critical bibliography, see E. BADIAN, "From the Gracchi to Sulla (1940–59)," *Historia,* 11:197–245 (1962); part 2 of Badian's *Foreign Clientelae,* 264–70 *B.C.* (1958), treats the political history down to 70 BC. For the social background, see the fundamental work of MATTHIAS GELZER, *Die Nobilitiit der römischen Republik* (1912; Eng. trans., *The Roman Nobility,* 1969); and, on special aspects of it, SUSAN TREGGIARI, *Roman Freedmen During the Late Republic* (1969); T.P. WISEMAN, *New Men in the Roman Senate,* 139 *B.C.–A.D.* 14 (1971); and E. BADIAN, *Publicans and Sinners* (1972). L. ROSS TAYLOR, *Party Politics in the Age of Caesar* (1949), gives the best analysis of the working of the political system after Sulla; RONALD SYME, *The Roman Revolution* (1939), is the classic study of the social and political transformation of the republic into the principate. Special aspects of particular interest are studied in JACQUES HARMAND, *L'Armée et le soldat a Rome de* 107 *à* 50 *avant notre Ére* (1967); E. BADIAN, *Roman Imperialism in the Late Republic,* 2nd ed. (1968); E.S. GRUEN, *Roman Politics and the Criminal Courts,* 149–78 *B.C.* (1968); and the last chapters of E.T. SALMON, *Roman Colonization Under the Republic* (1969).

Rome—31 BC to AD 193: TACITUS' *Annals* and *Histories,* describing the period from AD 14 to 96, are largely lost, including his account of the Flavian period, much of which he himself witnessed. The extant portions, however, written from the viewpoint of a disillusioned member of the Roman Senate, form the authoritative basis for any account of Tiberius, Claudius, Nero, and the year 69. The surviving ancient text that comes closest to covering the whole period is DIO CASSIUS' history of Rome to AD 222, in Greek. It is extant for the years 31 BC to AD 46 (and, for those from 46 to 193, there are epitomes and fragments, some of considerable length). These historical works are supplemented by imperial biographies: SUETONIUS' *Lives of the Caesars* (Julius Caesar to Domitian inclusive), PLUTARCH'S *Galba* and *Otho,* and the *Historia Augusta* (Hadrian and his successors). Suetonius is very informative and the *Historia Augusta,* a late compilation of dubious authorship and varying value, is a main source for the 2nd century. A selective, but very good collection of source material for the period in English is provided by NAPHTALI LEWIS and MEYER REINHOLD (eds.), *Roman Civilization,* 2 vol. (1951–55). *(Modern works): The Cambridge Ancient History,* vol. 10–11 (1934–36), with serviceable bibliographies of earlier work, is still a most helpful study in English. The best one-volume general accounts, both of them recent and with first-class bibliographies, are ALBINO GARZETTI, *L'Impero da Tiberio agli Antonini* (1960), particularly good for political and social developments; and PAUL PETIT, *La Paix romaine* (1967), particularly good on military and economic aspects. For the foundation of the principate, see D.C. EARL, *The Age of Augustus* (1968), a very realistic account; and for the development of the principate, MASON HAMMOND, *The Antonine Monarchy* (1959), careful, but somewhat optimistic. For imperial biographies, see under the individual emperors. Two recent ones, ANTHONY BIRLEY, *Marcus Aurelius* (1966); and B.H. WARMINGTON, *Nero* (1969), are very informative, the latter on the first century and the former on the second. For the economy: TENNEY FRANK (ed.), *An Economic Survey of Ancient Rome,* 6 vol. (1933–40)—vol. 2–5 are fundamental for the study of this period; M.I. ROSTOVTZEFF, *The Social and Economic History of the Roman Empire,* 2nd ed. rev. by P.M. FRASER, 2 vol. (1957), brilliant, but often subjective. For social life: JEAN GAGE, *Les Classes sociales dans l'Empire romain* (1964), comprehensive. For the provinces: THEODOR MOMMSEN, *Römische Geschichte,* vol. 5 (1886; Eng. trans., *The Provinces of the Roman Empire,* 2 vol., 1909), a classic; its European sections have now been edited by T.R.S. BROUGHTON, 1968; G.H. STEVENSON, *Roman Provincial Administration till the Age of the Antonines* (1939), sound. For the provincial communities: F.F. ABBOTT and A.C. JOHNSON, *Municipal Administration in the Roman Enzpire* (1926, reprinted 1968), a standard work; A.H.M. JONES, *The Cities of the Eastern Roman Provinces,* 2nd ed. rev. (1971), essential for the Greek-speaking East. For the armed services: H.M.D. PARKER, *The Roman Legions,* ed. by G.R. WATSON (1958); and C.G. STARR, *The Roman Imperial Navy,* 31 *B.C.–A.D.* 324, 2nd ed. (1960), both standard works. For general culture: JEAN BEAUJEU, *La Religion romaine à l'apogée de l'Empire* (1955), a very useful survey; J.A. CROOK, *Law and Life of Rome* (1967), sound and readable; W.L. MACDONALD, *The Architecture of the Roman Empire* (1965), excellent; RONALD SYME, *Tacitus,* 2 vol. (1958), a mine of information on many aspects of the period; J.M.C. TOYNBEE, *The Art of the Romans* (1965), comprehensive and sensible.

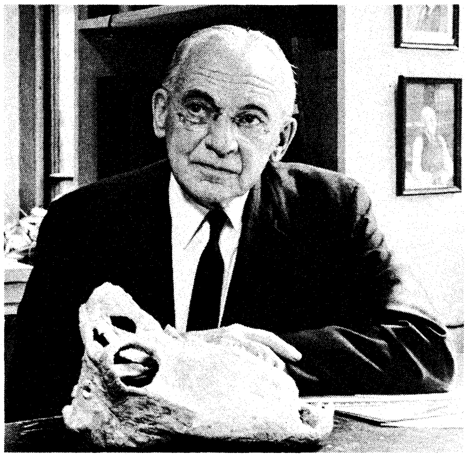
The Later Roman Empire—AD 193 *to c.* 500 *in the West, to c.* 395 *in the East:* The general works treating all aspects of the period are a little old: *The Cambridge Ancient History,* vol. 12, *The Imperial Crises and Recovery, A.D. 193–*324 (1939); and *The Cambridge Medieval History,* vol. 1, *The Christian Roman Empire nnd the Foundation of the Teutonic Kingdoms* (1936). A very detailed (except with regard to events) and excellent work is A.H.M. JONES, *The Later Roman Empire, 284–602: A Social, Economic and Administrative Survey,* 3 vol. (1964). On the problems of civilization and the passage from the Roman to the barbarian worlds, see JOSEPH VOGT, *Der Niedergang Roms* (1965; Eng. trans., *The Decline of Rome: The Metamorphosis of Ancient Civilization,* 1967), a completely remarkable work in all respects. In the absence of recent works on the Severan dynasty, that of G.J. MURPHY, *The Reign of the Emperor L. Septimius Severus, front the Evidence of the Inscriptions* (1945), assembles the epigraphic sources. L. ROSS TAYLOR, *The Divinity of the Roman Emperor* (1931); and FRANZ CUMONT, *Les Religions orientales dans le paganisme romain,* 4th ed. (1929), are classic and reliable. See also the great literary histories of ALBIN LESKY, *Geschichte der griechischen Literatur,* 2nd ed. (1963; Eng. trans., *A History of Greek Literature,* 1966); and H.J. ROSE, *A Handbook of Latin Literature,* 2nd rev. ed. (1949). The best work on ancient "paideia" is that of H.I. MARROU, *Histoire de l'éducation dans l'antiquité,* 5th ed. (1960; Eng. trans., *A History of Educatiort in Antiquity,* 1956). Economic aspects are discussed in F.M. HEICHELHEIM, *An Ancient Economic History,* vol. 3 (1970); and above all by M.I. ROSTOVTZEFF, *The Social and Economic History of the Roman Empire,* 2nd ed. rev. by P.M. FRASER, 2 vol. (1957). On the transformations of society, C.W. KEYES, *The Rise of the Equities in the Third Century of the Roman Empire* (1915); and the most recent and complete work of JEAN GAGE, *Les Classes sociales dans l'Empire romain* (1964), are recommended. On municipal life, F.F. ABBOTT and A.C. JOHNSON, *Municipal Administration in the Roman Empire* (1926, reprinted 1968), includes an excellent text and a wide choice of documents. The religious problems created by Constantine's conversion and the birth of the Christian empire are treated in the classic works of ANDRAS ALFOLDI, *The Conversion of Constantine and Pagan Rome* (1948); and of A.H.M. JONES, *Constantirze and the Conversion of Europe* (1949). Questions of culture and mind of the 4th century are dealt with in the essays of several scholars and brought together by ARNALDO MOMIGLIANO in *The Conflict Between Paganism and Christianity in the Fourth Century* (1963). A lively and well-documented work is SAMUEL DILL, *Roman Society in the Last Century of the Western Empire,* 2nd ed. (1958): and for the history of ideas, see SANTO MAZZARINO, *La fine del mondo antico* (1959; Eng. trans., *The End of the Ancient World,* 1966). The very important work of VOGT (cited above) gives an excellent bibliography (pp. 319–328) on all the problems treated in this section.

(R.Bl./J.P.V.D.B./E.Ba./E.T.S./P.P.)

Romer, Alfred Sherwood

Alfred S. Romer's concepts of the nature and order of the evolutionary history of vertebrate animals have strongly affected the outlook of all evolutionary biologists since the early 1930s. This U.S. vertebrate paleontologist and comparative anatomist is most widely known for his books *Vertebrate Paleontology, The Vertebrate Body,* and *Marl and the Vertebrates.* The explicit use of comparative anatomy and embryology in studies of fossil vertebrates underlies his major contributions to biology.

Romer was born in White Plains, New York, on December 28, 1894. His early life and schooling gave no indication of the direction his career was to follow. His father, a newspaper man, moved frequently, and young Alfred's early days were unsettled. In 1909, when he returned to White Plains to live with his father's mother, a more



Romer, *1965.*
By *courtesy of the Harvard University News Service*

stable phase of his youth began, though the family was poor. A rather timid boy, he found adjustment to the new environment difficult. In high school, however, he found a place among his peers. A trait that was to dominate his life emerged—an urge to improve things by becoming involved.

In 1913, after a year of doing odd jobs to earn money for college, Romer, entirely on his own and against family tradition, entered Amherst College. He financed his way in college with the help of a scholarship, various kinds of work, and loans from his college fraternity. History and German literature were his major subjects of study, but he liberally sampled courses in other disciplines. Early in his career, while in New York City, young Romer visited the American Museum of Natural History, and its dinosaur exhibits sparked his interest in fossils. After taking a course in evolution under Frederic B. Loomis to fulfill a science requirement, Romer knew what he wanted to do.

Graduation in 1917, during World War I, found him restless. He joined the American Field Service in France and in the autumn of that year enlisted in the United States Army. On his return to the United States in 1919, he entered graduate school at Columbia University to work under William K. Gregory. Romer completed the work for his Ph.D. in two years and produced a thesis that remains a classic in comparative myology, the study of musculature. With others who were students at Columbia at about this time, Romer was deeply influenced by Gregory and helped to spread his ideas, establishing a school of comparative anatomy, functional anatomy, and evolution that still prevails.

After two years as instructor of anatomy at Bellevue Hospital Medical College in New York, Romer became associate professor in the department of geology and paleontology at the University of Chicago. There he encountered fine collections of Late Paleozoic fossil fishes, amphibians, and reptiles, which set the course for much of his later fieldwork and scientific studies.

Early in his career at Chicago, Romer met and married Ruth Hibbard, who bore him three children and was his companion in fieldwork and travel. *Vertebrate Paleontology* appeared in 1933. In its three editions, this book shaped much of the thinking in the subject for several decades. After 11 years at Chicago, marred only by the problem of training biologically oriented graduate students in a geology department, Romer went to Harvard University as professor of biology in 1934.

His early propensity for leadership emerged when he became director of the biological laboratories in 1945 and of the Museum of Comparative Zoology in 1946. The museum flourished under his guidance.

Romer's scientific career prospered at Harvard with the publication of *Review of the Pelycosauria* (1940), *The Vertebrate Body* (1949), *The Osteology* of *the Reptiles* (1956), and numerous research papers. *The Vertebrate Body* is a comprehensive textbook of comparative anatomy, widely used in colleges and universities throughout the United States. The *Pelycosauria* and *Osteology* treat various aspects of reptilian evolution. Honours began to come to him—honorary degrees, memberships in foreign societies, and medals. Most important to him, however, was the less renowned Society of Vertebrate Paleontology, which he nurtured through its infancy, becoming its first president after it had become formally established in 1940. Later he initiated action that led to the establishment of the Vertebrate Morphology Section of the American Society of Zoologists. He accepted the presidency of the XIVth International Zoological Congress, which met in Washington, D.C., in 1963. He had long since been elected a member of the National Academy of Sciences of the United States. In 1929 and in 1956, he made trips to South Africa and Argentina to collect fossils. Despite the pace of his life—teaching, travel, administration, scientific study, and honours—Romer remained a relaxed, affable man, with a quiet wit and charm, always in demand as a lecturer. His rigorous dedication and depth of knowledge often emerged from light, colloquial presentations and an informal style.

Romer retired from Harvard in 1965 and, with characteristic energy, used his leisure for world travel, fossil collecting, lecturing, attending conferences, and visiting museunis. Scientific papers continued to flow from his pen. His studies extended the understanding of the course of evolution of the lower vertebrates. His basically conservative approach focussed attention upon the significance of form and function in relation to environment as the key to evolutionary progress. Documentation of the slowly changing flow through time, as envisaged by Darwin, emerges clearly from Romer's studies of the evolution of extinct fishes, amphibians, and reptiles.

Romer devoted a lifetime of research to the evolution of the vertebrates, with evidence drawn from comparative anatomy, embryology, and particularly paleontology. For many years he collected fossils of fishes, amphibians, and reptiles from geological deposits in Texas dating to the Permian Age. Close attention to anatomical adaptations as responses to environmental change characterized his method. He traced fundamental changes in structure and function that have occurred during the evolution of fishes to primitive land vertebrates and during the subsequent radiation of these vertebrates to modern amphibians, reptiles, and mammals. He died in Cambridge, Massachusetts, during the night of November 5, 1973.

(E.C.O.)

Rommel, Erwin

A German general of exceptional gifts of initiative and improvisation, Erwin Rommel became the most popular general at home and gained the open respect of his enemies with his spectacular victories as commander of the Afrika Corps in World War II. When, in 1944, his role in the unsuccessful conspiracy to remove Hitler as Germany's rulei was discovered, he was forced to kill himself.

Erwin Johannes Eugen Rommel was born on November 15, 1891, in Heidenheim in the kingdom of Württemberg. His father was a teacher, as his grandfather had been, and his mother was the daughter of a senior official. A career as an army officer began to be fashionable, even among middle class south Germans, after the establishment of the German Empire in 1871; and thus, notwithstanding the absence of a military tradition in his family, Erwin Rommel in 1910 joined the 124th Wiirttemberg Infantry Regiment as an officer cadet.

Early career

In World War I Rommel fought as a lieutenant in France, Romania, and Italy. His deep understanding of his men, his unusual courage, and his natural gift of leadership quite early showed promise of a great career. In the Prussian-German army, a career on the general staff was the normal avenue for advancement, yet Rommel declined to take that road. Both in the Reichswehr of the Weimar Republic and in Hitler's Wehrmacht, he remained in the infantry as a front line officer. Like many great generals, he possessed, possibly gotten from his father, a pronounced talent for teaching and was accord-

Rornrnei, 1941.
Ullstein Bilderdienst

ingly appointed to posts at various military academies. The fruit of his battle experiences in World War I, combined with his ideas on training young soldiers in military thinking, formed the main components of his military textbook *Infanterie greift an* ("Infantry Attacks"), originally published in 1937, which received high initial esteem.

In 1938, after Austria's annexation by Germany, Colonel Rommel was appointed commandant of the officers' school in Wiener Neustadt, near Vienna. At the beginning of World War II he was appointed commander of the troops guarding the Fuhrer's headquarters—not a very satisfying post for an enthusiastic front line soldier. Rommel's chance to prove himself came in February 1940 when he assumed command of the 7th Panzer Division. He had never commanded armoured units before, yet he quickly grasped the tremendous possibilities of mechanized and armoured troops in an offensive role. His raid on the channel coast provided the first proof of his boldness and initiative.

Commander of Afrika Corps

A year later, in February 1941, Rommel was appointed commander of the German troops dispatched to aid the all but defeated Italian army in Libya. The deserts of North Africa became the scene of his greatest successes—and of his defeat at the hands of a vastly superior enemy. In the North African theatre of war the "Desert Fox," as he came to be called by both friend and foe because of his audacious surprise attacks, acquired a formidable reputation, and soon Hitler promoted him to field marshal. He found it difficult, however, to get along with his Italian allies; basically, the British were more to his liking. He was an avid reader of the book on warfare written by Sir Archibald Wavell, the first Allied commander in chief of the Middle East, and Rommel's opponent in North Africa.

Rommel had difficulties not only with his Italian allies but with his own supreme command as well. North Africa was, in Hitler's view. onlv a sideshow. Nonetheless, despite the increasing difficulties of supply and Rommel's request to withdraw his exhausted troops, in the summer of 1942, Hitler ordered an attack on Cairo and the Suez Canal. Rommel and his German-Italian army were stopped by the British at el-Alamein (al-ʿAlamayn, Egypt) 60 miles (96 kilometres) from Alexandria. At that time Rommel won astounding popularity in the Arab world, where he was regarded as a "liberator" from British rule. At home, the propaganda ministry portrayed him as the invincible "people's marshal."

But the offensive against Egypt had overtaxed his resources. At the end of October 1942, he was defeated in the second Battle of el-Alamein and had to withdraw to the German bridgehead in Tunis. In March 1943 Hitler ordered him home. In 1944 Rommel was entrusted with the defense of the French channel coast against a possible Allied invasion. The master of the war of movement then developed an unusual inventiveness in the erection of coastal defense works. Nevertheless, his recommendation to prevent the enemy by all possible means from establishing large bridgeheads, his insistence that strong forces should be kept in reserve immediately behind the coastal defense line for counterattacks, and his prophecy that, unless the enemy could be successfully driven back into the sea, the fate of the invasion battle would be decided on the first day all fell on deaf ears.

Conspiracy against Hitler

As early as the fall of 1943, Rommel, a purely professional soldier whose judgment was not swayed by political predilections, had been convinced that the war could no longer be won and that Hitler was neither prepared to face that fact nor to draw the inevitable conclusion—the necessity of making peace with the Western powers. In the spring of 1944 some of Rommel's friends who had joined the clandestine opposition to Hitler approached Rommel and suggested to him that it was his duty to take over as head of state after Hitler had been overthrown. Rommel did not reject the suggestion; but the men who wanted to extricate Germany from the war never revealed to Rommel that they planned to assassinate Hitler. They knew that Rommel did not accept the idea of murder for political ends; he had invariably disregarded any execution orders given by Hitler. When the invasion began, Rommel tried on several occasions to point out to Hitler that the war was lost and that he should come to terms with the Western powers. On July 17, 1944, at the height of the invasion battle, Rommel's car was attacked by British fighter-bombers and forced off the road. It somersaulted and Rommel was hospitalized with serious head injuries. In August he had recovered sufficiently to be able to return to his home to convalesce. In the meantime, after the failure of the attempt on Hitler's life on July 20, 1944, Rommel's contacts with the conspirators had come to light. Hitler did not want the "people's marshal" to appear before the court as his enemy and thence be taken to the gallows. He sent two generals to Rommel to offer him poison with the assurance that his name and that of his family would remain unsullied if he avoided a trial. On October 14, 1944, Rommel took poison, thus ending his life. He was later buried with full military honours.

BIBLIOGRAPHY. ERWIN ROMMEL, *Infanterie greift an* (1937), is a collection of tactical studies and notes based upon his own experiences in World War I. After World War II his widow and one of his former chiefs of staff published his personal papers and notes: LUCIE-MARIA ROMMEL and FRITZ BAYERLEIN, *Krieg ohne Hass* (1950; Eng. trans., *The Rommel Papers,* 1953). RONALD LEWIN, *Rommel As Military Commander* (1968), is the best study to date; DESMOND YOUNG, *Rommel* (1950), is of great value as the first biography by a British author. FRIEDRICHRUGE, *Rommel und die Invasion: Erinnerungen* (1959); HANS SPEIDEL, *Invasion 1944: Ein Beitrag zu Rommels und des Reiches Schicksal* (1949; Eng. trans., *Invasion 1944: Rommel and the Normandy Campaign,* 1950; British title, *We Defended Normandy,* 1951); and SIEGFRIED WESTPHAL, *Heer in Fesseln: Aus den Papieren des Stabschefs von Rommel, Kesselring und Rundstedt* (1952), are of great value as memoirs.

(W.Go.)

Ronsard, Pierre de

The most eminent and prolific poet of the French Renaissance, Pierre de Ronsard produced verse that epitomized his time, reflecting the diverse topics of interest, tastes, and moods of Renaissance France. To the 20th-century reader he is perhaps most appealing when celebrating his native countryside, reflecting on the brevity of youth and beauty, or voicing the various states of unrequited love, though he is also effective when identifying himself imaginatively with some classical mythological character or situation and when expressing sentiments of fiery patriotism or deep humanity. He was a master of lyric themes and forms, and his poetry remains attractive to composers; some of his odes, such as "Mignonne, allons voir si la rose . . . ," were set to music half a dozen times in the 16th century alone and have become as familiar to the general public in France as folksongs.

He was born on September 11, 1524, at his father's manor house, La Possonnière, near the village of Couture

Ronsard, portrait after an engraving by L. Gaultier, 1557.

By courtesy of the trustees of the British Museum; photograph, J.R. Freemen & Co. Ltd.

in the *département* of Loir-et-Cher. As a younger son of a noble family of the county of Vendôme, he entered the service of the royal family as a page in 1536 and accompanied Princess Madeleine to Edinburgh after her marriage to James V of Scotland. On his return to France two years later, a court appointment or a military or diplomatic career seemed to be open before him, and in 1540 he accompanied the diplomat Lazare de Baïf on a mission to an international conference at Haguenau in Alsace. An illness contracted on this expedition left him partially deaf, however, and his ambitions were deflected to scholarship and literature. For someone in his position, the church provided the only future, and he accordingly took minor orders, which entitled him to hold ecclesiastical benefices, though he was never an ordained priest. A period of enthusiastic study of the classics followed his convalescence; during this time he learned Greek from the brilliant tutor Jean Dorat, read all the Greek and Latin poetry then known, and gained some familiarity with Italian poetry. With a group of fellow students he formed a literary school that came to be called the Pléiade, in emulation of the seven ancient Greek poets of Alexandria: its aim was to produce French poetry that would stand comparison with the verse of classical antiquity.

Poetry of the Pléiade

The title of his first collection of poems, *Odes* (4 books, 1550), emphasizes that he was attempting a French counterpart to the odes of the ancient Roman poet Horace. In *Les Amours* (1552) he also proved his skill as an exponent of the Italian *canzoniere*, animating the compliments to his beloved, entreaties, and lamentations traditional to this poetic form by the vehemence of his manner and the wealth of his imagery. Always responsive to new literary influences, he found fresh inspiration in the recently discovered verse of the Greek poet Anacreon (6th century BC). The more playful touch encouraged by this model is to be felt in the *Bocage* ("grove") of poetry of 1554 and in the *Meslanges* ("miscellany") of that year, which contain some of his most exquisite nature poems, and in the *Continuation des amours* and *Nouvelles Continuations*, addressed to a country girl, Marie. In 1555, he began to write a series of long poems, such as the "Hymne du Ciel" ("Hymn of the Sky"), celebrating natural phenomena, abstract ideas like death or justice, or gods and heroes of antiquity; these poems, published as *Hymnes* (following the 3rd-century-BC Greek poet Callimachus, who had inspired them), contain passages of stirring eloquence and vivid description, though few of them can hold the modern reader's interest from beginning to end. Reminiscences of his boyhood inspired other poems, such as his "Complainte contre fortune," published in the second book of the *Meslanges* (1559), which contains a haunting description of his solitary wanderings as a child in the woods and the discovery of his poetic vocation. This poem is also notable for a celebrated denunciation of the colonization of the New World, whose people he imagined to be noble savages living in an unspoiled state of nature comparable to his idealized memories of childhood.

Court poetry

The outbreak of the religious wars found him committed to an extreme royalist and Catholic position, and he drew upon himself the hostility of the Protestants. To this period belong the *Discours des misères de ce temps* (1562) and other *Discours*, attacking his opponents, whom he dismissed as traitors and hypocrites, with ever-increasing bitterness. Yet he also wrote much court poetry during this period, encouraged by the young king Charles IX, a sincere admirer; and on the King's marriage to Elizabeth of Austria in 1571, he was commissioned to compose verses and plan the scheme of decorations for the state entry through the city of Paris. If he was by now in some sense the poet laureate of France, he made slow progress with *La Franciade*, which he intended to be the national epic; this somewhat half-hearted imitation of Virgil's great Latin epic, the *Aeneid*, was abandoned after the death of Charles IX, the four completed books being published in 1572. After the accession of Henry III, who did not favour Ronsard so much, he lived in semiretirement, though his creativity was undiminished. The collected edition of his works published in 1578 included some remarkable new works, among them the so-called "Elegy Against the Woodcutters of Gâtine" ("contre les bucherons de la forêt de Gastine"), lamenting the destruction of the woods near his old home; a sequel to *Les Amours de Marie;* and the *Sonnets pour Hélène*. In the latter, which is now perhaps the most famous of his collections, the veteran poet demonstrates his power to revivify the stylized patterns of courtly love poetry. Even in his last illness, Ronsard still wrote verse, sophisticated in form and rich with classical allusions. His posthumous collection, *Derniers Vers*, poignantly expresses the anguish of the incurable invalid in nights spent alone in pain, longing for sleep, watching for the dawn, and praying for death. He died December 27, 1585, at the priory of Saint-Cosme, near Tours.

Ronsard perfected the 12-syllable or alexandrine line of French verse, hitherto despised as too long and pedestrian, and established it as the classic medium for scathing satire, elegiac tenderness, and tragic passion. He was a man of powerful intellect and wide learning, and in his lifetime he was recognized as the prince of poets, as well as a figure of national significance. This prominence, scarcely paralleled until Victor Hugo in the 19th century, faded into relative neglect in the 17th and 18th centuries; but his reputation was reinstated by the Romantic critic Sainte-Beuve, and it has remained secure for the last 100 years.

MAJOR WORKS

Les Quatre Premiers Livres des Odes, 4 books (1550), also including the first *Le Bocage; Les Amours* (1552), with 5th book of *Odes; Livret de Folastries* (1553); *Les Amours*, 2nd ed. (1553), with 39 new sonnets; *Le Bocage* (1554); *Les Meslanges* (1554, dated 1555); *Continuation des amours* (1555); *Les Hymnes*, 2 books, (1555–56); *Nouvelle Continuation des amours* (1556); *Les Oeuvres de P. de Ronsard*, 4 vol. (1560), first collected works; subsequent editions added to by Ronsard appeared in 1567, 1571, 1572–73, 1578, and 1584; *Institution pour l'adolescence du roy trhs Chrestien Charles neufviesme de ce nom* (1562); *Discours des mishres de ce temps à la Royne nzhre du Roy* (1562); *Remonstrance au peuple de France* (1562); *Résponce . . . aux injures et calomnies* (1563); *Élégies, mascarades et hergerie* (1565); *Les Quatre Premiers Livres de La Franciade*, 4 books (1572); *Les Oeuvres de P. de Ronsard*, 5th ed. (1578), including *Les Amours diverses*, *Les Amours de Marie: seconde partie, sur la mort de Maria*, and the *Sonnets pour Hélène; Les Derniers Vers de P. de Ronsard* (1586).

There are translations of the *Sonnets pour Hélène* by Humbert Wolfe (1934) and W. Van Wyck (1932). See also *Songs and Sonnets*, trans. by Curtis Hidden Page (1903); and *Salute to Ronsard*, ,selected poems trans. by E.J. Dennis (1960).

BIBLIOGRAPHY. The major collected edition of Ronsard's works to be published in this century is the *Oeuvres complètes*, critical edition with introduction and commentary by PAUL LAUMONIER for the *Société des Textes Français Modernes*, 18 vol. (1914–70); the volumes after Laumonier's death were completed by R. LEBEGUE and ISIDORE SILVER. This gives the original text of each poem (with the variants, due to subsequent revision by Ronsard, in footnotes) and in the order originally published. It is indispensable for the study of Ronsard's development as a poet. The edition by GUSTAVE COHEN, in the "Pléiade" series, 2 vol. (1938), offers a reliable text (following the edition of 1584) in a compact form. HUMBERT WOLFE, *Pierre de Ronsard: Sonnets pour Hélène, with English Renderings* (1934), is a complete translation, in sonnet form; there are some misunderstandings, but the renderings are often felicitous. The *Lyrics of Pierre de Ronsard Chosen and Translated by Charles Graves* (1967), presents in a pleasant and scholarly form a selection of the odes and sonnets, of the Elegy to Mary, Queen of Scots, and of the Epitaph upon himself. CLAUDE BINET, *Discours de la vie de Pierre Ronsard* (1586), is the basis of all study of Ronsard's biography; it is best read in the critical edition by PAUL LAUMONIER (1910). PAUL LAUMONIER, *Ronsard, poète lyrique*, 2nd ed. (1923), is a complete survey of the themes and forms used in Ronsard's lyric poetry; PIERRE DE NOLHAC, *Ronsard et l'humanisme* (1921), is a full examination of Ronsard's relations with classical scholars. MORRIS G. BISHOP, *Ronsard, Prince of Poets* (1940), is a useful biography without being definitive; D.B. WYNDHAM LEWIS, *Ronsard* (1944), should be used with caution, the approach being very personal, the quotations sometimes inaccurate, and the translations of them often incorrect. Recent studies, in English, of important aspects of Ronsard's poetry include: D.B. WILSON, *Ronsard, Poet of Nature* (1961); DONALD STONE, JR., *Ronsard's Sonnet-Cycles: A Study in Tone and Vision* (1966); ELIZABETH ARMSTRONG, *Ronsard and the Age of Gold* (1968); and ISIDORE SILVER, *The Intellectual Evolution of Ronsard*, vol. 1 (1969). A complete bibliography of Ronsard up to about 1950 may be found in A. CIORANESCU, *Bibliographie de la littérature française du seizième siècle*, pp. 609–625 (1959).

(A.E.Ar.)

Roosevelt, Franklin D.

Franklin Delano Roosevelt, the 32nd president of the United States, held office from 1933 to 1945 during the New Deal era and World War II. The modern role of the United States government, in both its domestic and foreign policies, owes much to the changes that Roosevelt helped bring about. To counter the Great Depression of the 1930s he enlisted the powers of the federal government to promote the economic welfare of the U.S. people. He was a leader in the Allied struggle against the Axis powers in World War II, preparing the way for the United States to assume a continuing role in world security. He was the only president to be re-elected three times.

UPI Compix

Franklin D. Roosevelt, 1937.

EARLY LIFE AND POLITICAL GROWTH

Roosevelt was born at Hyde Park, New York, on January 30, 1882, the only son of James and Sara Delano Roosevelt. The Roosevelts lived in unostentatious and genteel luxury, dividing their time between the Hudson River Valley and European resorts. They often took young Franklin to Europe; he was taught privately at home and was reared to be a gentleman, responsible toward those less fortunate. At 14, Roosevelt, a rather shy youth, entered Groton School (Groton, Massachusetts), modelled after the great public schools of England, where wealthy young men were trained to exercise Christian stewardship through public service.

Schooling and training

After he entered Harvard in 1900, Franklin Roosevelt threw himself into undergraduate activities. His strenuous extracurricular and social life left him relatively little time for his studies, in which his record was undistinguished. He was, however, influenced by his economics professors, who modified traditional laissez-faire views with advocacy of government regulation; but, even more, Roosevelt fell under the spell of the progressive president, his glamorous distant relative Theodore Roosevelt, a fifth cousin.

During his final year at Harvard, Franklin became engaged to Theodore Roosevelt's niece, Eleanor Roosevelt, who was then active in settlement work in New York City; they were married on March 17, 1905. Eleanor helped open young Roosevelt's eyes to the deplorable living conditions of the underprivileged in the slums.

New York social life interested Roosevelt more than did his studies at Columbia University School of Law. As soon as he passed the New York bar examination, he discontinued his schooling. This indifference to the legal profession carried over into Roosevelt's years as a clerk with the distinguished Wall Street firm of Carter, Ledyard & Milburn, defense counsel in several spectacular antitrust cases.

Early political activities. His admiration for his cousin Theodore, who continued to urge young men of substance to enter public service, led Roosevelt toward politics. His opportunity came in 1910 when the Democratic leaders of Dutchess County, New York, persuaded him to undertake an apparently futile campaign for the state senate. Roosevelt, whose branch of the family had always been Democratic, hesitated only long enough to make sure his distinguished Republican relative would not speak against him.

State senator. He campaigned so strenuously that, with the aid of a Republican schism and his famous name, he won the election. Roosevelt, not quite 29, quickly won statewide and even some national attention by leading a small group of Democratic insurgents who refused to vote for the nominee of Tammany Hall, the New York City organization. For three months Roosevelt helped hold the insurgents firm, until Tammany switched to another candidate.

First political campaign

Roosevelt became the foremost champion in the New York Senate of the upstate farmers, and in the process he converted to the full program of progressive reform. From the New York City legislators, whom he had earlier scorned and now continued to fight, he learned much of the give-and-take of politics. Among them were James J. Walker, later mayor of New York City; Robert Ferdinand Wagner, who became a leading U.S. senator; and Alfred E. Smith, later governor of New York. Roosevelt gradually abandoned his patrician airs and attitude of superiority.

Before the end of 1911, Roosevelt supported the presidential boom for Gov. Woodrow Wilson of New Jersey, the leading Democratic progressive. An attack of typhoid fever kept Roosevelt from participating in the 1912 campaign, but, even without making a single public appearance, he was re-elected to the state senate. This was because of publicity by an Albany newspaperman, Louis McHenry Howe, who saw in the tall, handsome young Roosevelt a promising politician. Howe served Roosevelt for the rest of his life with a jealous loyalty.

Assistant secretary of the navy. For his work on behalf of Wilson, Roosevelt was rewarded in March 1913 with an appointment as assistant secretary of the navy under Josephus Daniels. Roosevelt loved the sea and naval traditions, and he knew more about them than did his superior, with whom he was frequently impatient. Roosevelt tried with mixed success to bring reforms to the navy yards, which were under his jurisdiction, meanwhile learning to negotiate with labour unions among the civilian employees. After war broke out in Europe, Roosevelt became a vehement advocate of preparedness; following U.S. entrance, he built a reputation as an effective administrator. In the summer of 1918 he made an extended tour of naval bases and battlefields overseas. During much of his seven years as assistant secretary, he had been less than loyal to Daniels, but in the end he came to appreciate his superior's skill in dealing with Southern congressmen and his solid worth as an administrator.

Vice-presidential nomination

Paralytic attack. At the 1920 Democratic convention Roosevelt was nominated for vice president. He campaigned vigorously with the presidential nominee, James M. Cox, on behalf of U.S. entrance into the League of Nations. After a Republican landslide, Roosevelt became a vice president of the Fidelity & Deposit Company of Md., a bonding company, entered into numerous business schemes (some of a speculative nature), and remained active in Democratic politics. Suddenly, in August 1921, while on vacation at Campobello Island, New Brunswick, Roosevelt was severely stricken with poliomyelitis. He suffered intensely and for some time was almost completely paralyzed, but he soon began predicting (as he did for some years) that he would quickly regain the use of his legs. His mother wished him to retire to Hyde Park, but his wife and his secretary, Louis Howe, felt it essential to his morale that he remain active in his career and in politics. Because Roosevelt could not himself go to political gatherings, his wife attended for him, acting as his eyes and ears (a service she frequently performed for him during the rest of his life). Under the tutelage of Howe, she overcame her shyness and became an effective political worker and speaker. Because he could not run for office for the time being, Roosevelt was able to function effectively as a sort of premature "elder statesman," trying to promote unity between the urban and rural wings of the Democratic Party. Himself a rural Democrat, he nominated Gov. Al Smith of New York, the favourite of the city faction, at the 1924 and 1928 Democratic conventions.

Smith urged Roosevelt to run for governor of New York in 1928 to strengthen the ticket. Roosevelt was reluctant; he still could not walk without braces and assistance. In the years since 1921 he had worked incessantly to try to regain the use of his legs—for several winters he swam in warm Florida waters and, beginning in 1924, in the mineralized water at Warm Springs, Georgia. Wishing to share with others the beneficent effect of the warm water and a systematic program of therapy, Roosevelt in 1927 established the Warm Springs Foundation, a nonprofit institution for the care of polio victims. He wished to develop Warm Springs further and to continue treatments in the hope of regaining full use of his legs.

Governor of New York. Nevertheless, despite these concerns and his feeling that 1928 was not a propitious year to run on the Democratic ticket, Roosevelt succumbed to strong persuasion and accepted the nomination. When he began campaigning by automobile, he demonstrated that he had retained his youthful buoyance and vitality; he also showed that he had matured into a more serious and human person. Opponents raised the question of his health, but his vigorous campaigning effectively disposed of the issue. Smith was defeated in Herbert Hoover's landslide, and he failed to carry New York state; but Roosevelt won by 25,000 votes.

Program as governor

Succeeding Smith as governor, Roosevelt decided he must establish his own type of administration. He did not keep Smith's closest adviser nor did he depend upon Smith for advice. Smith, already stung by his defeat for the presidency, was hurt and alienated. Whereas Smith had built his reputation on administrative reform, Roosevelt concentrated upon a program to give tax relief to farmers and to provide cheaper public utilities for consumers. The appeal of this program in upstate New York, coupled with the effects of the deepening Depression, led to Roosevelt's re-election in 1930 by the overwhelming plurality of 725,000 votes.

During his first term as governor, Roosevelt's policies, except on the power issue, were scarcely further to the left than those of President Hoover in Washington, D.C. But during Roosevelt's second term, as the Depression became more catastrophic in its effects, he acted to mobilize the machinery of the state government to aid the economy. In the fall of 1931 he obtained legislation establishing the Temporary Emergency Relief Administration, the first of the state relief agencies. Throughout his four years, he was successful in most of his bouts with the Republican legislature, sharpening skills that would prove vital in the future. And, increasingly, beginning with some slight speculation in November 1928, he was being talked of as the most likely Democratic presidential nominee in 1932. After his spectacular victory in 1930, he was so conspicuous a target for the Republicans and for rival Democratic aspirants that he had no choice but to begin immediately and quietly to obtain support for the convention. Because it then took a two-thirds vote in the Democratic convention to nominate, a leading contender could be stopped with relative ease. It soon became apparent that Roosevelt's strongest opposition would come from urban and conservative Eastern Democrats still loyal to Smith; his strongest support was in the South and the West.

Progressives and intellectuals found Roosevelt's overall program attractive, but many feared that he was weak because he sidestepped Republican challenges to oust corrupt Democratic officials in New York City. The opposition became stronger when John Nance Garner of Texas, speaker of the House of Representatives, won the California primary.

PRESIDENCY

At the 1932 convention Roosevelt had an early majority of the delegates but seemed blocked by a combination of the Smith and Garner forces. On the third ballot, Garner allowed his delegates to be thrown to Roosevelt; in return, Garner was nominated for the vice presidency.

Election to the presidency

First term. In the campaign of 1932 the Depression was the only issue of consequence. Roosevelt, displaying smiling confidence, campaigned throughout the country, outlining in general terms a program for recovery and reform that came to be known as the New Deal. In a series of addresses carefully prepared by a team of speech writers, popularly called the brain trust, he promised aid to farmers, public development of electric power, a balanced budget, and government policing of irresponsible economic power. He declared in his most notable speech in San Francisco: "Private economic power is . . . a public trust as well." His program appealed to millions who were nominally Republicans, especially Western progressives. Roosevelt received 22,822,000 popular votes in the election to Hoover's 15,762,000; the electoral vote was 472 to 59. The Democrats also won substantial majorities in both houses of Congress.

Inauguration as president. Roosevelt took office on March 4, 1933. Following the election, President Hoover had sought Roosevelt's cooperation in stemming the deepening economic crisis that culminated in the closing of banks in several states during February 1933. But Roosevelt refused either to accept responsibility without the accompanying power or to subscribe to Hoover's proposals for reassuring business; Hoover himself granted that his proposals would mean "the abandonment of 90 per cent of the so-called new deal."

When Roosevelt took office, most of the nation's banks were closed, industrial production was down to 56 percent of the 1929 level, 13,000,000 or more persons were unemployed, and farmers were in desperate straits. Even the congressional leaders were so shaken that for the time being they were ready to follow Roosevelt's recommendations.

In his inaugural address Roosevelt promised prompt, decisive action and somehow conveyed to the nation some of his own unshakable self-confidence. "This great Nation will endure as it has endured, will revive and will prosper," he asserted, ". . . the only thing we have to fear is fear itself." For the moment, people of all political views were Roosevelt's allies, and he acted swiftly to obtain enactment of the most sweeping peacetime legislative program in U.S. history.

The Hundred Days. Through a broad array of measures, Roosevelt first sought quick recovery and then reform of the malfunctions in the economic system that he thought had caused the collapse. He tried to aid each of the main interest groups in the U.S. economy and, at a time when the Democrats were the minority politi-

Attempts to maintain a consensus

cal party, to hold the backing of many who were previously Republicans. His choice of Cabinet members indicated his efforts to maintain a consensus; it was geographically and politically balanced, containing both liberal and conservative Democrats, three Republicans, and, for the first time, a woman--Secretary of Labor Frances Perkins.

He also directed his legislative program toward a broad constituency. The prelude was the enactment of several conservative measures, to inspire confidence among businessmen and bankers. First Roosevelt ended depositors' runs on banks by closing all banks until Congress, meeting in special session on March 9, could pass a cautious measure allowing those in a sound condition to reopen (Roosevelt also strongly favoured banking reform, but it came later). In March, Roosevelt redeemed one of his most important campaign pledges by introducing a program of drastic government economy. He firmly believed in economy and never became a convert to the Keynesian views so often attributed to him. The emergency banking and economy acts brought him the enthusiastic support of an overwhelming proportion of the electorate but, he pointed out at the time, could do little to bring real recovery.

Roosevelt was already preparing, and he soon sent to Congress, a series of messages and draft bills proposing the program that comprised the early New Deal. Roosevelt first obtained from Congress federal funds for the relief of human suffering. Congress established the Federal Emergency Relief Administration (FERA), which granted funds to state relief agencies for direct relief. It also established a Civilian Conservation Corps (CCC), which at its peak employed 500,000 young men in reforestation and flood-control work; it was a favourite project of Roosevelt's and remained popular through the New Deal. Mortgage relief aided other millions of persons, both farmers and homeowners. The key loan agency of the New Deal was the previously established Reconstruction Finance Corporation (RFC), the powers of which were broadened so that it could make loans to small enterprises as well as to large. Although at the time Roosevelt did not envisage public spending as the primary role of these relief agencies, the agencies poured so much money into the economy that within several years they were stimulating recovery.

Recovery measures. The two key recovery measures of the New Deal were acts to restore farm prosperity and to stimulate business enterprise. The first act, in 1933, established the Agricultural Adjustment Administration (AAA), the objective of which was to raise farm prices and increase the proportion of the national income going to farmers. The principal means was through subsidies given to growers of seven basic commodities in return for their willingness to reduce production. The subsidies were to be paid from a processing tax on the commodities. Roosevelt accepted this scheme as a temporary expedient, which Congress would enact because a majority of farm organization leaders favoured it. He also hoped to raise farm prices through mild inflation. Roosevelt envisaged a program, following farm recovery, of extensive rural planning—moving farmers from submarginal to better lands and luring some of the unemployed from metropolises to rural and village life. In 1935 Roosevelt obtained the Resettlement Administration, which gave some aid to smaller, poorer farmers. When the Supreme Court invalidated the processing tax in 1936, he switched the AAA program to one of soil conservation. Nevertheless, throughout the New Deal, farm leaders and Congress succeeded in maintaining an agricultural program the major emphasis of which was to raise farm prices. Thanks to this legislation and several years of drought, production fell, and farm income gradually improved. But not until 1941 did it reach even the inadequate level of 1929.

National Industrial Recovery Act

The demand of businessmen for government stabilization and of labour for a shorter workweek led Roosevelt to recommend to Congress the National Industrial Recovery Act (NIRA) of 1933. It was a two-pronged program. On one side was a $3,300,000,000 appropriation for public works. Had this money been poured into the economy rapidly, it would probably have done much to bring recovery, but Roosevelt wanted to be sure it would be spent soundly on self-liquidating public works, through the Public Works Administration (PWA). Because careful planning took time, the PWA did not become an important factor until late in the New Deal. On the other side of the NIRA was a National Recovery Administration (NRA), to administer codes of fair practice within given industries. At first under a "blanket code," then under specific codes negotiated by representatives of each industry and labour, minimum wages, maximum hours, and fair trade practices were established within each industry. The codes were designed to stabilize production, raise prices, and protect labour and consumers. Consumers received scant protection, but labour received guarantees on wages and hours and also the right to bargain collectively. During the summer of 1933 there was a quick flurry of recovery as manufacturers produced goods in anticipation of sale at higher prices under the codes, the boom collapsed by fall because prices had risen faster than purchasing power.

By February 1934 the code making was over, but far too many—557 basic codes and 208 supplementary ones—had come into existence, containing innumerable provisions that were difficult to enforce. By 1935 the business community, which had demanded the NRA at the outset, was becoming disillusioned with it and blaming Roosevelt for its ineffectiveness. In May the Supieme Court, in the Schechter decision, invalidated the code system. Despite shortcomings, however, the NRA had aided several highly competitive industries, such as textiles, and brought important reforms that were re-enacted in other legislation: federal wages-and-hours regulation, collective-bargaining guarantees, and abolition of child labour in interstate commerce.

In the fall of 1933 Roosevelt had already turned to other expedients for bolstering the economy. He experimented with "managed currency," driving down the gold content of the dollar and tripling the price of silver through large purchases. These efforts brought only small price increases at home, but they improved the position of the United States in foreign trade by making dollars cheaper abroad. In January 1934 Roosevelt stabilized the gold content of the dollar at 59.06 percent of its earlier value. Managed currency created a significant precedent, even though it did little to bring recovery at the time.

Opposition from the right and left

Altogether, by the fall of 1934 Roosevelt's program was bringing a limited degree of recovery, but it was alienating conservatives, including many businessmen. They contended that much of the program was unconstitutional, that it created uncertainties for business that hampered recovery, and that the lowering of the gold content of the dollar had deprived holders of government obligations of their just return. At the same time, many of the underprivileged who were still in serious difficulties felt that the New Deal had not gone far enough. They were ready to listen to demagogic leaders offering still more. In the 1934 mid-term election they voted overwhelmingly for Democratic candidates for Congress; but there was a danger that in the 1936 presidential election they might vote for a third-party candidate to the left of Roosevelt.

Reform measures. To meet the threat to his political coalition from the left, Roosevelt emphasized reform in his annual message to Congress in January 1935. This was less a shift from a first to a second New Deal than it was a rush to enact reform measures that Roosevelt had long been planning. In 1933 he had obtained the Tennessee Valley Authority (TVA), to provide flood control, cheap hydroelectric power, and regional planning for an impoverished region. At his recommendation also, Congress had enacted two laws to protect investors: the Truth-in-Securities Act of 1933 and an act establishing the regulatory Securities and Exchange Commission (SEC) in 1934.

Additional legislation in 1935 did much to undermine the appeal of demagogues to the needy, especially the Social Security Act, which included unemployment insurance and old-age insurance. For workers still unemployed, Congress created the Works Progress Administration (WPA), to provide relief that would stem the erosion of their skills and self-respect (between 1935 and 1941 the WPA employed an average of 2,100,000 workers, and by the end of 1935 it was already bringing a marked measure of recovery by pouring billions of dollars into the economy). For workers who were employed, the National Labor Relations Act (the Wagner Act), only belatedly accepted by Roosevelt, strengthened the government guarantees of collective bargaining and created a National Labor Relations Board (NLRB) to adjudicate labour disputes. The Public Utility Holding Company Act, also of 1935, regulated the control holding companies had over operating public utility companies. A new 1935 tax measure, labelled by its opponents the "soak-the-rich" tax, raised the levies on persons with large incomes and on big corporations and became a significant factor in redistributing U.S. income.

Second term. These measures effectively undercut the left-wing opposition to Roosevelt, but they further alienated conservatives. He ran for re-election in 1936 with the firm support of farmers, labourers, and the underprivileged; and the epithets that the extreme right hurled at him merely helped unify his following. The Republican nominee, Gov. Alfred Mossman Landon of Kansas, a moderate, could do little to stem the Roosevelt tide. Roosevelt received 27,752,000 popular votes to Landon's 16,680,000 and carried every state except Maine and Vermont.

Re-election in 1936

Supreme Court fight. The only hope of conservatives to thwart the New Deal was for the Supreme Court to invalidate its key measures. Following the Schechter decision, the court in 1936 ruled against the AAA processing taxes, and cases challenging the Social Security Act and the Wagner Act were also pending. Roosevelt, beginning his second term with a massive mandate, was determined to remove this threat. Believing that the measures were well within the scope of the Constitution and that the reasoning of the justices was old-fashioned and at fault, he proposed early in 1937 the reorganization of the court, including the appointment of as many as six new justices. The proposal, labelled by opponents as a court-packing scheme, touched off a vehement debate in which many of Roosevelt's previous supporters in and out of Congress expressed their opposition. Meanwhile, in the spring of 1937, the Supreme Court upheld both the Wagner Act and the Social Security Act. With the need for the court plan dissolving, its enemies managed by summer to bring about its defeat. This was a severe political blow for Roosevelt, even though the new decisions by the court opened the way for almost unlimited government regulation of the economy.

Growing opposition. Roosevelt's prestige dropped further in the summer of 1937, when much of the public blamed him for labour difficulties that grew out of organizing drives in the steel, automobile, and other mass-production industries. Operating under the protection of the Wagner Act, the unions engaged in strikes that often resulted in violence. Roosevelt himself preferred paternalistic government aid to all workers, such as the wages-and-hours guarantees of the Fair Labor Standards Act of 1938. But union membership jumped to about 9,500,000 by 1941, while most middle class people returned to the Republican Party.

A sharp economic recession in the fall of 1937 added to Roosevelt's troubles. There had been substantial recovery by 1937; but Roosevelt, wishing to balance the budget, had curtailed government spending drastically, sending the economy plummeting back toward 1932 levels. Businessmen blamed the New Deal spending policies; Roosevelt blamed the businessmen and inaugurated an anti-monopoly program. In October 1937 massive government spending began again, and by June 1938 the crisis was past.

From 1938 on, many of the conservative Southern Democrats heading key congressional committees openly opposed the New Deal. In 1938 Roosevelt tried unsuccessfully to defeat several of them in the primaries and was inveighed against as a dictator trying to conduct a "purge." Democrats won the November elections, but the Republicans gained 80 seats in the House and seven in the Senate, permitting a coalition of Republicans and conservative Democrats that could thwart the President.

Nevertheless, the second Roosevelt administration saw the passage of some notable reform legislation, extending and improving earlier legislation and moving into some new fields. These years also saw the development of soil conservation to stem erosion and the large-scale construction of public works, including public housing and slum clearance. Many New Deal innovations, such as social security, the agricultural program, the TVA, and the SEC, had now become accepted as permanent functions of the federal government.

Foreign policy. By 1939 foreign policy was overshadowing domestic policy. Even before taking office, Roosevelt had endorsed Hoover's refusal to recognize Japanese conquests in Manchuria. From the outset of his administration, Roosevelt was deeply involved in foreign-policy questions, mostly relating to the Depression. In the early summer of 1933 he refused to support international currency stabilization at the London Economic Conference, but by 1934 he had stabilized the dollar and had begun helping France and Great Britain to keep their currencies from being undermined by dictator nations. In November 1933 Roosevelt recognized the government of the Soviet Union in the mistaken hope that he could thus promote trade. Greater opportunities seemed to exist in negotiating reciprocal trade agreements with numerous nations—a program that began in 1935—and in fostering more cordial relations with Latin American nations. In his first inaugural address Roosevelt had pledged himself to the "policy of the good neighbor." Secretary of State Cordell Hull had interpreted this to mean no unilateral U.S. intervention in Latin America; but, gradually, as European war became imminent, the Good Neighbor Policy led to collective-security and mutual-defense agreements.

In the early New Deal years, Roosevelt not only pursued programs of economic nationalism but, like most Americans, was also intent upon keeping the United States out of any impending war. He thus supported a series of neutrality laws, beginning with the Neutrality Act of August 1935. Roosevelt moved toward a new policy in 1937, after Japan began a major thrust into northern China. In October, speaking in Chicago, he proposed that peace-loving nations make concerted efforts to quarantine aggressors. He seemed to mean nothing more drastic than the breaking off of diplomatic relations, but the proposal created such national alarm that during ensuing months he was slow to develop a collective-security position. He quickly accepted Japanese apologies when the U.S. gunboat "Panay" was sunk on the Yangtze River in December 1937. Relations between the United States and Japan gradually worsened, but the rapid domination of Europe by Adolf Hitler of Germany was more threatening.

The outbreak of war. When World War II began in Europe in September 1939, Roosevelt called Congress into special session to revise the Neutrality Act to permit belligerents to buy arms on a "cash-and-carry" basis. With Hitler's aggressions and the fall of France in the spring and early summer of 1940, Roosevelt and Congress turned to defense preparations and "all aid short of war" to Great Britain. Roosevelt even gave Great Britain 50 overage destroyers in exchange for eight Western Hemisphere bases. Isolationists, fearing U.S. involvement in the war, debated hotly with those who felt the national self-interest demanded aid to Britain.

The third and fourth terms. In the 1940 presidential campaign the Republicans nominated Wendell L. Willkie, who agreed with Roosevelt's foreign policy. Both candidates pledged to keep the nation out of foreign war; but isolationists tended to support Willkie, while those favouring strong measures against Hitler swung toward Roosevelt. By a closer margin than before—27,244,000 to 22,305,000 popular votes and 449 electoral votes to 82—Roosevelt was elected to an unprecedented third

Re-election in 1940

Through 1941 the nation moved gradually closer toward actual belligerency with Germany. After a bitter debate in Congress, Roosevelt in March 1941 obtained the Lend-Lease Act, enabling the United States to finance aid to Great Britain and its allies. Preventing submarines from sinking goods en route to Europe gradually involved more drastic protection by the U.S. Navy; in the fall Roosevelt authorized the navy to "shoot on sight" at German submarines. Meanwhile, in August, on a battleship off Newfoundland, Roosevelt met with Prime Minister Winston Churchill of Great Britain and signed a joint press release proclaiming an Atlantic Charter to provide national self-determination, greater economic opportunities, freedom from fear and want, freedom of the seas, and disarmament.

U.S. entry into the war. Yet it was in the Pacific that war came to the United States. Japan, bound in a treaty of alliance with Germany and Italy, the so-called Axis, extended its empire in East Asia. Roosevelt, viewing these moves as part of Axis world aggression, began to deny Japan supplies essential to its war making. Throughout 1941 the United States negotiated with Japan, but proposals by each side were unsatisfactory. Roosevelt did not want war with Japan in the fall of 1941, but he miscalculated in thinking the Japanese were bluffing. By the end of November he knew that Japanese fleet units and transports were at sea and that war was imminent; an attack in Southeast Asia and perhaps on the Philippines seemed likely. To Roosevelt's angered surprise, the Japanese, on December 7, 1941, struck Pearl Harbor, Hawaii. On December 8, at Roosevelt's request, Congress voted a war resolution within four hours; on December 11, Germany and Italy declared war on the United States.

Roosevelt made concessions to the conservatives in Congress in order to obtain support in prosecuting the war. Several New Deal agencies were abolished. At a press conference Roosevelt asserted that "Dr. Win the War" had replaced "Dr. New Deal" but that this was to be only for the duration of the war. Roosevelt also fought resourcefully although not always successfully against inflationary pressures.

One of the immediate problems after Pearl Harbor was to build up massive production for war. Roosevelt had begun experimenting in 1939 with various defense agencies to mobilize the economy. Eventually, a workable organization had evolved. At the time of Pearl Harbor, U.S. war production was already nearly as great as that of Germany and Japan combined; by 1944 it was double the total of all Axis nations.

Relations with allies. During the war, Roosevelt concentrated upon problems of strategy, negotiations with the nation's allies, and the planning of the peace. From the outset, he took the lead in establishing a grand alliance among all countries fighting the Axis.

Wartime leader

Roosevelt met with Churchill in a number of wartime conferences at which differences were settled amicably. Debate at the earlier conferences centred upon the question of a landing in France, which the British succeeded in postponing repeatedly; the great Normandy invasion was finally launched in June 1944. Meanwhile, the United States had followed the British lead in invading North Africa in November 1942, Sicily in July 1943, and Italy in September 1943. At one of the most significant of the meetings, at Casablanca, Morocco, in January 1943, Roosevelt, after previous consultation with Churchill, proclaimed the doctrine of unconditional surrender of the Axis. He seemed to want to avoid the sort of differences of opinion among the Allies and misunderstanding by the Germans that had made trouble at the time of the 1918 Armistice. There is no tangible evidence that the doctrine in any way lengthened the war.

Relations with the Soviet Union posed a difficult problem for Roosevelt. Throughout the war the U.S.S.R. accepted large quantities of lend-lease supplies but seldom divulged its military plans or acted in coordination with its Western Allies. Roosevelt, feeling that the maintenance of peace after the war depended upon friendly relations with the Soviet Union, hoped to win Joseph Stalin's confidence. Roosevelt seemed to get along well with Stalin when he and Churchill first met with the Soviet leader at Teheran, Iran, in November 1943. In their optimism, Roosevelt and Churchill seemed not to see realistically that the sort of peace being foreshadowed at Teheran would leave the U.S.S.R. dominant in Europe.

Meanwhile, the Axis had been suffering serious defeats in both Europe and the Pacific. By February 1945, when the Big Three met again at Yalta in the Crimea, the war seemed almost over in Europe. As for Japan, the United States expected a last-ditch defense that might require another 18 months or more of fighting. Work in developing an atomic bomb was well advanced, but its power was expected to be only a fraction of what it actually turned out to be. Consequently, Roosevelt and his military advisers were eager to obtain Soviet aid in Asia; and, in return for Stalin's promise to enter the war against Japan, Roosevelt and Churchill offered concessions in the Far East. As for eastern Europe, earlier decisions were ratified, and plans were made for the establishment of democratic governments. Had the arrangements for eastern Europe been followed by Stalin in the manner expected by Roosevelt and Churchill, there would have been little room for criticism. But the understandings were not precise enough, and they received different interpretations in the U.S.S.R. By mid-March false Soviet accusations against the United States led Roosevelt to send a sharp telegram to Stalin.

Declining health and death. Roosevelt hoped that the establishment of an effective international organization, the United Nations, could maintain the peace in years to come. He planned to attend a conference of 50 nations at San Francisco, opening April 25, 1945, to draft a United Nations charter. But, since January 1944, his health had been declining. His political opponents had tried to make much of this during the campaign of 1944, when he ran for a fourth term against Gov. Thomas E. Dewey of New York. A final burst of vigour on Roosevelt's part, however, seemed to refute the rumours. Roosevelt won by 25,602,000 to 22,006,000 in popular votes and 432 to 99 in electoral votes. But his address to Congress after he returned from Yalta had to be delivered sitting down. He went to Warm Springs for a rest, and there, on April 12, 1945, he died of a massive cerebral hemorrhage.

Re-election in 1944

EVALUATION

During Roosevelt's years as president, he had relatively little time for personal life. He continued his interest in his lands at Hyde Park. His zest for sailing and his enjoyment in collecting stamps and naval books and prints continued unabated. His tight schedule and the incessant publicity imposed upon him limited the time he could give to his wife, who became an important figure in her own right, and to his five children: Anna Eleanor, James, Elliott, Franklin D., Jr., and John A. As a public figure he was, at the same time, one of the most loved and most hated men in U.S. history. Opponents ascribed to him shallowness, incompetence, trickiness, and dictatorial ambitions. His supporters hailed him as the saviour of his nation's economy and the defender of democracy not only in the United States but throughout the world. It was generally conceded that as a political leader he was unexcelled in winning and holding popular support and in retaining, in his administration, leaders of diverse views. Many experts have expressed the opinion that despite occasional confusion and overlapping authority, his administration was unusually effective. He brought even more than this to the office: in 1932 he stated what remained his view through peace and war, "The Presidency . . . is pre-eminently a place of moral leadership."

BIBLIOGRAPHY. S.I. ROSENMAN (ed.), ***The Public Papers and Addresses of Franklin D. Roosevelt,*** 13 vol. (1938–50, reprinted 1969), contains official statements. J.M. BURNS, ***Roosevelt,*** 2 vol. (1956–70); and FRANK B. FREIDEL, ***Franklin D. Roosevelt,*** 3 vol. of a projected six-volume work (1952–), are the most detailed biographies. A.M. SCHLESINGER, JR., ***The Age of Roosevelt,*** 3 vol. to date (1957–), is a brilliant survey of both the man and era; W.E. LEUCHTENBERG, ***Franklin D. Roosevelt and the New Deal, 1932–1940*** (1963), is an authoritative brief account.

(F.Fr.)

Roosevelt, Theodore

Theodore Roosevelt, 26th president of the United States, author, explorer, and amateur soldier, left a permanent mark upon American politics. He made the presidency a more powerful office, persuaded Congress to regulate the railroads, challenged the power of large industries, and engaged the nation diplomatically in Asia and Europe. Roosevelt was the first president who envisioned the federal government as a protector of the public interest, and as an umpire in the growing conflicts between big business and big labour.

By courtesy of the Library of Congress, Washington, D.C.

Theodore Roosevelt.

Roosevelt was born in New York City on October 27, 1858, of a moderately wealthy family of Dutch ancestry; his mother, Martha Bulloch of Georgia, was of Scots-Irish and Huguenot descent. He received an excellent education from private tutors and at Harvard College; he was one of the very few presidents endowed with an encompassing intellectual curiosity. In 1880 he entered Columbia University Law School. But historical writing and politics soon lured him away from a legal career. During the same year he married Alice Hathaway Lee of Boston and after her death, in 1884, married Edith Kermit Carow, with whom he lived for the rest of his life near Oyster Bay, Long Island.

Early political and military career. A physical weakling during his youth, by persistent exercise Roosevelt developed a rugged physique and became a lifelong advocate of strenuous activity. He was a born competitor against both nature and his fellowman, and he used the same enormous energy in public life. Conflict, which he thoroughly enjoyed, came easily to this man endowed, according to a friend, the historian Henry Adams, with one of the attributes of divinity, that of being "pure act" (which apparently meant that he could not sit still and that he did not need to think). Another of his friends remarked that Roosevelt was compelled to be the centre of every stage, the bride at every wedding, the corpse at every funeral. At the age of 23 he successfully ran for the New York State Assembly, in which he soon became one of the Republican leaders, known for his opposition to corrupt, party-machine politics. Misfortune then struck in the form of three successive political defeats. But after two years spent ranching in the Dakota Territory he re-entered public life and continued his reform activities as a member of the U.S. Civil Service Commission (1889–95) and as the president of the New York City Board of Police Commissioners (1895–97). As assistant secretary of the navy under Pres. William McKinley he vociferously advocated war with Spain. When war was declared in 1898, he abruptly resigned, organized the 1st Volunteer Cavalry, known as the "Rough Riders," and took them to Cuba that year. Roosevelt's leadership was spectacular. Disdaining army red tape and even orders, his colourful exploits, especially his impulsive charge up San Juan Hill, made him something of a national hero.

"Rough Riders"

Roosevelt returned home just when Thomas C. Platt, the Republican boss of New York, was looking for a respectable candidate for governor. Platt distrusted him as the "perfect bull in a china shop"; but upon Roosevelt's promise that he would not attack the machine, he was easily elected. An excellent governor, he removed several corrupt politicians from office and over Platt's opposition secured a corporation franchise tax and a civil service system. Enraged, Platt manoeuvred Roosevelt into the 1900 nomination for vice president on the McKinley ticket and thus secured his elimination from state politics.

Accession to the presidency. Roosevelt's enemies compared his election to the vice presidency to a nun's "taking the veil," and the prediction soon proved accurate. He found himself completely bored by this powerless office until September 14, 1901, when McKinley died after being shot by an assassin, and he himself became president. He now had the opportunity to put into effect his political ideas, some of which differed considerably from those of his Republican predecessors.

Roosevelt viewed the presidency not only as an office of tremendous power for formulating legislative and administrative policy but also as a major force in determining the very quality of American life. Abraham Lincoln and Andrew Jackson, both of whom had used the presidency to initiate great changes in American society, were his favorite presidents. Not since Lincoln's day was the power of the office used with such vigour. Some of his opponents called him "Theodore Rex" and likened him to Oliver Cromwell and Napoleon.

Although Roosevelt announced that there would be no change in policy, it soon became apparent that a new life-style had been introduced at the White House. Guest lists there were expanded to include cowboys, prizefighters, explorers, and some of the more distinguished artists of the nation. Young, college-educated men were appointed to administrative positions. Presidential speeches overflowed with indignation and moral righteousness. It soon became apparent that Teddy—as he was known nationwide—was enjoying himself immensely. Whether campaigning, advising Congress, or belabouring a hapless opponent, few presidents have pursued their way with more personal zest.

But for all the office's satisfactions, Roosevelt also had reasons for subdued reflection. He was always conscious that he had become president by accident, and his chief ambition was to be elected in 1904, and he was continuously worried about the possibility of defeat. A highly sensitive politician, he was aware that William Jennings Bryan's defeat for the presidency in 1896 had not quieted the popular demands he represented for control of the trusts, regulation of railroads, and a reduction of import duties. But he also knew that both houses of Congress were controlled by conservative Republicans bitterly opposed to all reforms. He met this perplexing situation by asking for little legislation and by using executive power in appeasing the rising popular discontent—hence the format of the "Square Deal."

The Square Deal. In 1902 Roosevelt took three steps that virtually assured his re-election. From Congress he asked for the establishment of a Bureau of Corporations with powers to inspect the books of all businesses engaged in interstate commerce. Even this limited measure was resisted by leading Republican conservatives; the President secured its passage only by promising not to ask for any further regulatory measures. But this bargain did not keep Roosevelt from further executive actions, and on February 18, in one brilliant stroke, he revived the all but forgotten Sherman Anti-Trust Act by bringing successful suit against the Northern Securities Company, a giant combine of railroads put together by four of the greatest industrialist capitalists, J.P. Morgan, John D. Rockefeller, Edward H. Harriman, and James Hill. Roosevelt pursued his policy of "trust-busting" by bringing suit against 43 other major corporations during the following seven years. Although he himself did not place much faith in breaking up the trusts, preferring their regulation, the Northern Securities action met with tremendous popular approval. Because of the prominence

Attacks on industrialists

of the men involved, its symbolic value was evident; and in the popular view big business, which had threatened to devour the small man, had finally met its master.

In the fall of 1902 Roosevelt again set an important precedent by intervening in the anthracite coal strike. Roosevelt's social views were not particularly advanced for his day. He accepted the legality of big labour unions as he accepted that of big business and felt that both contributed to the general welfare. But he believed the rights of the general public as represented by the government were superior to either. Thus, when the strike threatened to result in cold homes, schools, and hospitals, he requested that representatives of capital and labour meet in the White House and accept mediation. By threatening to use the army to operate the mines he won an arbitration agreement that included a modest pay increase for the miners. Never before had the federal government intervened in a labour struggle except to assure the operation of a governmental service or to protect property. Roosevelt promptly labelled his actions against industry and indirectly for labour a manifestation of a "Square Deal" between labour and capital. In the long run, however, the most significant aspects of his actions were the precedents they set for governmental intervention in the affairs of business and labour for the public interest.

Once overwhelmingly elected in 1904 as president "in his own right," Roosevelt immediately asked Congress for substantial powers to regulate interstate railroad rates. The Hepburn Act of 1906, giving the Interstate Commerce Commission authority to set maximum rates, created the first of the government's regulatory commissions and thus was a milestone on the long road to the modern social-service state.

Consumer protection and conservation laws

Roosevelt's pressure on Congress also led to the passage of the Pure Food and Drug and the Meat Inspection acts (1906) which laid the basis for the modern concept of consumer protection. But when asked after retirement which of his many achievements he considered most important to the national welfare Roosevelt named his conservation policy. Responding to the rapid disappearance of the federal domain, Congress had empowered the President 15 years before to convert portions of the remaining land to national forests. Under Roosevelt's three predecessors only some 40,000,000 acres (16,000,000 hectares) had been transferred. Roosevelt and his chief conservation agents, Secretary of the Interior James R. Garfield and Chief Forester Gifford Pinchot, not only rapidly accelerated the pace but broadened the powers of the act to reserve for future generations parks, and mineral, oil, and coal lands, as well as waterpower sites. In seven years, 194,000,000 additional acres (78,000,000 hectares) of the federal domain were closed to commercial development. Had it not been for Roosevelt's vigorous use of a slumbering statute, today's ecological problems would be infinitely more severe.

Foreign policy. In international affairs Roosevelt believed that strong nations survived while weak ones perished. He also sensed that the relatively peaceful period that had preceded his administration was being replaced by one in which force was the principal arbiter. Every year he asked for larger naval appropriations, and to induce Congress to grant him new ships he occasionally exaggerated the seriousness of international incidents. By the end of this term Roosevelt had built the U.S. Navy into a major sea force. He also encouraged his secretary of war, Elihu Root, to continue his reform of army administration.

In the developing world power struggle Roosevelt regarded Germany as the chief potential threat to the United States. The other dominant naval power, Britain, had international aims very close to those of the U.S. But threatened German expansion, particularly in the Caribbean and in Latin America, was probably the main cause of Roosevelt's armed intervention in the Caribbean. It also prompted his extension of the Monroe Doctrine—according to which the United States regarded European interference in any American nation as an unfriendly act—and even his hesitant participation in European affairs at the Algeciras Conference of 1906.

Roosevelt Corollary to the Monroe Doctrine

Twice during Roosevelt's years in office European powers threatened intervention in Venezuela and once in the Dominican Republic, presumably to collect debts owed to their nationals. To meet a threat of possible permanent intervention Elihu Root and Roosevelt framed a policy statement in 1904 that eventually became known as the Roosevelt Corollary to the Monroe Doctrine. The corollary stated that not only would the United States prohibit non-American intervention in Latin American affairs but it would police the area and guarantee that these nations met their international obligations. The corollary sanctioning American intervention was to be applied in 1905 when, without Congressional approval, Roosevelt forced the Dominican Republic to accept the appointment of an American "economic advisor," who quickly became the financial director of the small state.

Quoting an African proverb, Roosevelt once said that the proper way to conduct foreign affairs was to "speak softly and carry a big stick." Roosevelt was to use big-stick diplomacy again in the acquisition of the Panama Canal Zone from Colombia in 1903, in the formation of a provisional government in Cuba in 1906, and to some extent in the quarrel with Canada over the Alaskan and Canadian border. During the Canal affair he actually wrote an undelivered message to Congress recommending that the nation take possession of the isthmus "without any further parley with Colombia." He also played a notable part in inspiring the subsequent Panamanian revolution that assured American control of the zone and enabled the United States to start construction of the canal before the presidential election of 1904.

If Roosevelt's dealings with small countries were often brusque, his negotiations with major powers were characterized by far more caution. The American Pacific position, he said in 1903, "is such as to insure our peaceful domination of its waters." But the steadily rising powei of Japan caused him to revise that estimate. His efforts to resolve the Russo-Japanese War of 1904–05 included bringing both countries to the Portsmouth Peace Conference and mediating between them. His direct motive, however, was to construct a balance of power in East Asia that might peacefully aid American interests. The friction caused by the anti-Japanese sentiment in California he helped to allay by the so-called Gentlemen's Agreement of 1907, restricting Japanese immigration to the United States. By another informal executive agreement Japan accepted the American position in the Philippines while the U.S. recognized the Japanese conquest and occupation of Korea. Later, in 1910, Roosevelt became convinced that the Philippines were indefensible against a Japanese thrust and that there was no hope of American "dominance" in East Asian waters.

European policy

During his last years as president, Roosevelt was worried by the possibility of a general European war. Since he saw British and American interests generally coinciding, he was strongly inclined to support Britain whenever it would not jeopardize official neutrality, violation of which would have brought strong protest from Congress and the country. The secret instructions given to the American representatives to the Algeciras Conference of 1906, called to prevent a European war over Morocco, were theiefore ambiguous. The envoys were told to maintain American neutrality but also to do nothing that would imperil the existing Franco-British understanding, the continuation of which was "to the best interest of the United States. . . ." But for all the talk of neutrality Roosevelt had in effect deviated from the traditional position of neutrality in non-American affairs. American representatives had attended a strictly European political conference, their actions favoured Britain and France as against Germany, and by signing the agreement the United States presumably undertook to sustain it. Algeciras pointed unerringly toward United States entry into World War I on the side of the allied powers.

Last years as president. The end of Roosevelt's presidency was anything but calm. His crusade against "race suicide," prompted by his alarm at the decreasing birthrate, his public indictment of amateur naturalists, and his order to the government printers to use a simplified system of spelling all developed into national arguments. Especially after the financial panic of 1907 his quarrels with Congress became more vehement. His rather high-handed disciplining of a Negro regiment involved in a riot at Brownsville, Texas, and his suggestion that congressmen who were opposed to increasing the secret-service funds had something to hide, all produced bitter controversy. But most of the congressional trouble came from the split that had developed in his party between the Roosevelt progressives and the party's conservatives, who blamed the financial panic of 1907 on Roosevelt's attacks on big industry. Roosevelt did more than leave a divided party behind him; when he chose William Howard Taft as his legatee and assured the country that Taft would carry out "my policies," he made sure that the party schism would continue into the next administration.

Participation in 1912 election

Later years. After leaving the White House in March 1909, Roosevelt took a ten-month hunting trip through central and northern Africa and made a grand tour of Europe. On his return, he was reluctantly drawn into politics. Though he attempted to support both his old progressive friends and Taft, who had often allied himself with the conservative leadership of the Republican Party in order to carry out progressive policies, the two men soon were violently opposed over policy matters. The conflict became personal in October 1910 when Taft authorized an anti-trust suit against the United States Steel Corporation regarding a merger to which Roosevelt as president had tacitly agreed. Personal animosity and the developing split in the Republican Party finally prompted Roosevelt to contest Taft's 1912 renomination. The resulting bitter campaign and convention practically insured a Democratic victory. Roosevelt himself made that outcome inevitable by founding the Progressive Party and running for president as an independent after he had lost the Republican nomination. In seeking votes, the former president, through both logic and necessity, was forced to radical proposals. Both the Progressive platform and its candidates' campaign for a "New Nationalism" looked forward to a powerful regulatory and social-service state, prefiguring the New Deal policies of President Franklin D. Roosevelt. The results of the campaign were as expected, with Woodrow Wilson, the Democratic candidate, winning by a large electoral vote.

Since the Progressive Party had managed to elect only a handful of candidates to minor offices, Roosevelt knew immediately that it was doomed. He kept it alive for bargaining purposes and, in the meantime, occupied himself with an expedition into the Brazilian jungles and with writing. After World War I broke out he became a strong partisan of the Allied cause, demanding a stiffer foreign policy toward Germany and national preparedness at home. Although ambitious for the 1916 Republican nomination, he was ready to support almost any candidate who opposed Wilson and who was not personally involved in his own defeat in 1912. Amid much bitterness he abandoned the Progressive Party and vigorously supported the Republican candidate, Charles Evans Hughes, but again his efforts ended in failure. His anger against Wilson increased when his offer to lead a volunteer division to France was rejected. Although he had previously supported an international peace-keeping organization, he was adamantly opposed to Wilson's League. By 1918 he felt the Republicans might nominate him for president in 1920. But the years of inordinate activity had taken their toll, and he died suddenly in his sleep on January 6, 1919.

Self image

Evaluation. Few presidents have reflected the dominant temper of their times as vividly as Theodore Roosevelt. America at the turn of the century was prosperous, confident, energetic, expansionist, and experimentally minded. So was Roosevelt. But though a compulsive activist, Roosevelt the politician was usually very cautious. He viewed himself as a middle-of-the-roader and a "half loaf" man, settling for what he thought was possible. Consequently, his tendency to compromise was criticized by contemporary radicals of both the left and the right.

Roosevelt alienated many people by the difference between his rhetoric, invariably highly seasoned with morality, and his performance, usually tempered by expedience. His penchant for impulsive actions and his bullying of small nations and individuals too helpless to reply infuriated others. In particular, he has been criticized by historians for his part in bringing on the Spanish-American War and the acquisition of the Philippines, and for his high-handed treatment of Colombia over the Panama Canal. Some historians have labelled his business policy as one that favoured the small man oratorically but in reality benefitted the large corporations. The majority of authorities, however, impressed with the new spirit he brought to the federal government and with the legislative achievements that laid the groundwork for the solution of many modern problems, have accorded him a more favourable judgment. If not ranked among the four or five great American Presidents, he is placed among the very small group of the near great.

MAJOR WORKS

HISTORY: *The Naval War of* 1812; *or the History of the United States Navy During the Last War with Great Britain* (1882); *The Rough Riders* (1899).

POLITICS: *Essays on Practical Politics* (1888); *American Ideals and Other Essays, Social and Political* (1897); *The Strerzuous Life: Essays and Addresses* (1900); *The New Nationalism* (1910); *Progressive Principles* (1913); *America and the World War* (1915); *Fear God and Take Your Own Part* (1916); *The Foes of Our Own Household* (1917); *The Great Adventure: Present-Day Studies in American Nationalism* (1918).

BIOGRAPHY: *Life of Thomas Hart Benton* (1887); *Gouverneur Morris* (1888); *Oliver Cromwell* (1900).

AUTOBIOGRAPHY: *Theodore Roosevelt: An Autobiography* (1913).

OTHER WORKS: *Hunting Trips of a Ranchman* (1885); *Ranch Life and the Hunting-Trail* (1888); *New York* (1891); *The Wilderness Hunter* (1893); *History As Literature, and Other Essays* (1913); *Through the Brazilian Wilderness* (1914); *A Book-Lover's Holidays in the Open* (1916).

BIBLIOGRAPHY. WILLIAM HENRY HARBAUGH, *Power and Responsibility: The Life and Times of Theodore Roosevelt* (1961), is the most objective and best of the one-volume biographies. HENRY F. PRINGLE, *Theodore Roosevelt,* rev. ed. (1956), although brilliantly written, is perhaps prejudiced against its subject. The most comprehensive work on the early life is CARLETON PUTNAM, *Theodore Roosevelt: The Formative Years,* 1858–1886 (1958), the first of a projected four-volume work. Particularly brilliant short interpretations are JOHN M. BLUM, *The Republican Roosevelt* (1954; with new preface, 1962); and G. WALLACE CHESSMAN, *Theodore Roosevelt and the Politics of Power* (1969). Other special studies of value are GEORGE E. MOWRY, *Theodore Roosevelt and the Progressive Movement* (1946); G. WALLACE CHESSMAN, *Governor Theodore Roosevelt* (1965); HOWARD K. BEALE, *Theodore Roosevelt and the Rise of America to World Power* (1956); and STEFAN LORANT, *The Life and Times of Theodore Roosevelt* (1959).

For a general history of Roosevelt's times, see GEORGE E. MOWRY, *The Era of Theodore Roosevelt,* 1900–1912 (1958); and HAROLD U. FAULKNER, *The Quest for Social Justice, 1898–*1914 (1931). Some of Roosevelt's 2,000 published works including books and hundreds of articles on history, politics, travel, and natural history, and perhaps a few of his personal letters numbering over 150,000 should be consulted. The most comprehensive collection of his works is the *Memorial* edition, 24 vol. (1923–26). A superb collection of his more important letters appears in ELTING E. MORISON, *The Letters of Theodore Roosevelt,* 8 vol. (1951–54).

(G.E.M.)

Ropes and Cables

Rope is an assemblage of fibres, filaments, or wires compacted by twisting or braiding (plaiting) into a long, flexible line. Wire rope is often referred to as cable. The basic requirement for service is that the rope structure remain firmly compacted and structurally stable, even while the rope is bent, twisted, and pulled. The prime property of a rope is its strength; *i.e.*, its ability to sustain a pulling force that may be exerted in order to move or resist the force of a weighty object.

The texture and the nature of the rope is determined by the fibres or filaments used in its construction. Natural fibres vary in colour, fineness, stiffness, strength, and stretchability. Ropes of such fibres will vary accordingly.

Cotton ropes, for example, are softer, weaker, and stretchier than manila or sisal ropes. Manila ropes are stronger, for a given size, than hemp or jute ropes. Sisal ropes are usually lighter in colour and somewhat stronger than henequen ropes. Since even short fibres are adaptable to an operation in which numbers of fibres are bunched together, overlapped or twisted around one another, and in compacted form extended into long flexible yarns, practically any fibre can be made into a rope. The rope strength attainable with any given fibre can be regulated by the size (bulk) of the rope.

Ropes made of filaments drawn from thermoplastic materials, such as ropes made from metallic wires, differ from natural-fibre ropes in that any particular filament contributing to the rope form runs throughout the entire length of the rope and reflects its characteristics more readily. A thick filament or wire is stiffer than a fine one of the same material; its use would then result in a less flexible rope for a given size.

Rope structure

The basic twisted rope structure requires at least two stages of twisting, the second in the reverse direction from the first in order to give torsional stability. The braided or plaited rope structure provides torsional balance by crossing and recrossing rope components in maypole fashion, such that each component passes alternately over and under, and at opposing angle to, one or more of the others.

Natural-fibre, man-made filament and wire ropes ordinarily start with a diameter of 3/16 inches (about 5 mm). Structurally stable rope structures smaller in size are designated cords, and when torsional balance is of no consequence, the designation may be twine or yarn. Cables are torsionally balanced structures formed by twisting several ropes together.

Compared to fibre ropes of a given size, natural or man-made, wire ropes are considerably stronger, will respond to load with little stretch, are much stiffer, and are proportionately heavier. Man-made filament ropes are considerably stronger than natural-fibre ropes but are generally stretchier under load.

FIBRE ROPES

History. Implements and weapons from the Paleolithic Period suggest that knowledge of twisting or plaiting strips or thongs from animal hides into cords, and even coarser vines and fibrous barks into thicker, stronger rope structures, had already been known to primitive man. The animal drawings in European caves and the carvings ascribed to the late Paleolithic Period even show what are perhaps twisted rope bridles made from skin strips or tendons. When Neolithic man began to build with stone and bronze, facing the problems of moving and lifting heavy objects on land and on water, he would have applied such knowledge to develop even stronger and longer twisted and plaited rope structures.

The progressive art of rope making accordingly represents the evolution of a process beginning with plaiting thongs into cords in the distant past to twisting heavy rope structures requisite for the development of massive construction and sea transportation in the eastern Mediterranean.

To rope making, too, may be ascribed the beginning of other developed arts. Before the end of the Paleolithic Period, thong plaiting had already engendered mat and basket weaving, later spinning and weaving of fine twisted yarns modelled after the coarse yarn structures of ropemaking. The knowledge of how to transform short fibres or strips into extensive lengths by twisting or plaiting was ready as growing civilizations discovered and learned to use fibres, animal and vegetable hairs, silken filaments, plant barks, and stems.

The method of making coir yarns and cords, practiced even today in Southeast Asia, demonstrates the fundamental nature of this art. Where coconut palms grow in abundance, the natives still occasionally practice a primitive technique of twisting. The fibres are rolled together with the palm of a hand on the bare thigh into a twisted yarn form, and, with a similar motion, two such yarns are rolled into a twisted cord.

Egyptian rope making

The rope maker, as an established artisan, attained a height of skill with the age of the pyramids in Egypt. Beginning about the 4th dynasty (3500 BC), enormously heavy stone blocks and colossal statues were dragged and moved by slaves, backs bent to heavy, strong ropes. Scenes in pyramids and tombs ascribed to later dynasties depict a well-developed rope-making process, which involved twisting strips cut from leather hides or fibres from papyrus reeds. The twisting whorl had already been established for twisting strands, cords, and even small ropes. Whether the largest ropes were twisted from strands or plaited is not clear.

An account by Herodotus (Book VII, 34 and 37) of the invasion of Greece in 480 BC by Xerxes discloses further advance in the art of rope making, as described by the bridge of boats used to cross the Hellespont. Many small boats or barges were lashed together by six cables—two of white flax of Phoenician make and four of papyrus of Egyptian manufacture--each about 1.5 miles (2.4 km) long. The flax cable was stated to have weighed one talent per cubit, corresponding to 38 pounds per foot, or 57 kilograms per metre, with a possible diameter as much as 14 inches (about 36 cm). It is problematical whether each cable was a single-twisted rope structure in the sense of modern technology, or a series of ropes lashed together to form the cable.

The knowledge of the art of rope making apparently was not limited to the Mediterranean countries. Rope making in India was such a specialized trade that as early as the 4th century BC, ropes were being made to suit specific uses. One may anticipate that jute, indigenous to the area, was used together with coir. In China, during the Han period (206 BC to AD 221), ropes made of silk filaments had already been fashioned for the emperor's funeral carriage. According to tradition, the emperor Shen Nung had encouraged the growth of hemp as early as the 28th century BC. This fibre, originating in the temperate zones of Asia, was discovered later in Europe. By 200 BC, Mediterranean ships were rigged with rope made of hemp grown in the valley of the Rhine. By the end of the 1st century AD, hemp had become the prime cordage material—a position held until the middle of the 19th century.

Abaca fibre

In 1830, when the Philippine Islands were opened to foreign commerce, American rope makers found that abaca fibre could be made into ropes (called manila) that were even stronger than hemp and more suited for marine work than any other fibre rope made theretofore. The use of this fibre expanded at the expense of hemp, giving manila ropes a position of superiority that lasted until the introduction of nylon for rope making in the 1950s, followed by polyester and polypropylene. These three man-made fibres constitute the main fibres of this category used for high-strength rope today.

Natural fibres. The natural fibres used for ropes are classified according to the nature of the plant and the location of the fibres within the plant structure. Cotton is a fine seed hair. Coir, on the other hand, is the coarse seed hair of the coconut, also used for rope. Hemp, jute, flax, kenaf, and ramie are fibres constituting the inner bark of stems or main stalks of reedlike plants. These fibres, long and multiple celled, are designated bast fibres. Abaca (manila), sisal, henequen, and phormium (New Zealand flax) are the coarse fibres extending lengthwise through the pulpy tissues of long leaves or leaf stems of tropical plants, which are also long and multiple celled.

The process for obtaining the fibre from the plant structure is determined by the difficulty of separating it from the adhering plant tissue. The bast fibres are quite adherent and are therefore separated by fermentation—a process called retting—followed by mechanical scraping, washing, drying, and baling. Hemp is derived from the plant *Cannabis sativa.* The coarse variety used for cordage is generally dew retted; *i.e.*, left in open fields exposed to the weather after being cut and stacked like bundles of grain. For finer grades, the plants are water retted; *i.e.*, immersed in natural bodies of water or tanks, to accelerate the process. Jute from the plants *Corchorus capsularis* and *Corchorus olitorius* is water retted, as is kenaf, a jutelike fibre from the plant *Hibiscus cannabinus.*

The leaf fibres manila, sisal, and henequen, generally used for the largest and strongest ropes, are readily removed from the leaf by a mechanical scraping process called decortication. Manila, also designated abaca, is the fibre from the leaf stem of the plant ***Musa textilis***, growing mainly in the Philippines. Since this fibre is removed readily from the plant, it is generally stripped away with a knife by hand, rather than by machine decortication. Sisal, from the plant ***Agave sisalana***, cultivated extensively in Africa and Brazil, and the coarser, darker variety Henequen from the plant ***Agave fourcroydes***, grown mainly in Mexico, are machine decorticated.

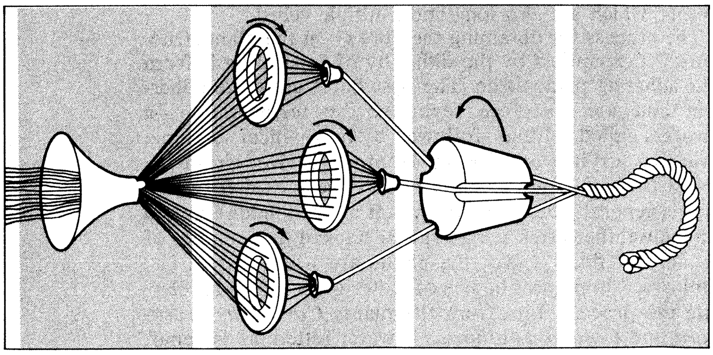
Man-made rope fibres. The man-made fibres used for rope are generally supplied in the form of yams built up by combining fine, high-strength filaments suited to textile spinning and weaving. Nylon, polyester, and polypropylene polymers are extruded through fine-holed dies while molten, then, while the extruded filaments are hot and plastic, they are stretched (drawn) to requisite fineness. The stretching process rearranges the molecules, orienting them to develop the high strength requisite for rope. The fineness of the filaments is expressed in denier; their strength in grams per denier. Nylon yarn in multiples of six denier filaments (a 9,000-metre length weighing six grams) develops a strength of at least 6.5 grams per denier, which is suitable for rope manufacture. Polypropylene rope is made of coarser filaments generally manufactured by the rope manufacturer. Such filaments may be as coarse as 600 denier in size (0.012-in., or 0.3-mm diameter).

Extrusion process

Nylon is the generic term for the polymer obtained from hexamethylene diamine and adipic acid (nylon 6,6) or the caprolactam (nylon 6). Polyester identifies the filaments obtained from the polymeric ester of ethylene glycol and terephthalic acid, marketed under various trade names such as Dacron, Terylene, etc. Polypropylene is the generic term for the polymer of propylene, a product from the petroleum industry. Both nylon and polyester filaments are supplied by chemical manufacturers. Because of its lower melting temperature, polypropylene, on the other hand, is generally extruded into filament form by the rope manufacturer and when used for rope is generally more wirelike in form and stiffness. Such filaments are designated as monofilaments. Finer filaments supplied by the chemical manufacturers are designated multifilament yarns.

Glass, which in recent years has received attention as a rope material because of its extremely high strength and low stretchability, is a form of silica that constitutes the composition of other glass forms. Because of extreme brittleness, the glass filaments must be well lubricated to retain their form through handling in subsequent manufacturing and service operations. The glass-fibre rope used commercially is protected from bending and abrasion stresses by bonding the glass filaments with resins.

Manufacturing process. Rope making is divided into four phases: (1) The fibres or filaments are prepared for spinning (twisting) into yarns. (2) The fibres or filaments are spun or bunched into yarns and yarns into cords for the manufacture of man-made filament ropes. (3) A number of yarns are twisted into strands (forming). (4) Three or more strands are twisted into rope (laying).



Processes through which fibre passes in the manufacture of rope.

Ordinarily, each phase is accomplished with the twist in the reverse direction from the twist in preceding phase. When the rope twist is seen to spiral in direction upward to the right as the rope is held vertically, it is designated as right-laid or Z-twist; if upward to the left, as left-laid (S-twist).

Rope twists

Preparing the fibres (slivering). The goal of fibre preparation is to convert the fibres, which are hairlike in form and several feet in length, into endless ribbons of parallel fibres, wherein each fibre is overlapped uniformly by adjoining fibres. The ribbon or stream of fibres is called the sliver.

The fibres taken from the bales as bundles, called hanks, are made parallel on combing (hackling) machines or, if very short, on carding machines. As the fibre bundles are spread to enter the machine, faster moving steel pins comb through the fibres in the direction of the fibre bundle, arranging the fibres in parallel order and removing extraneous debris. The fibres are at first retained while being combed, but soon moving at the faster speed of the combing pins, some fibres are pulled through the bundle faster than others, thereby extending and overlapping the separate fibre bundles. This process of extending and overlapping is called drafting or drawing. When the fibres emerge from the first combing machine, which is called the breaker, the separate bundles of fibres will have been converted into sliver form. At this stage, the lubricant and preservative liquid is added to the fibre bundles. This not only facilitates the subsequent combining and drafting operations but, retained in the rope, improves its serviceability.

In subsequent combing and drawing operations, emerging slivers are now combined (doubled) and repeatedly combed and drawn in at least four more similar operations, the machines for each operation being provided with combing pins that are progressively finer in size. The doubling operations afford opportunity for mixing the fibres and redistributing them so that on a chance basis the thinner sections of one sliver will coincide with the heavier sections of another and thereby provide uniformity with respect to the number of fibres in the cross-section and the degree to which the ends of the fibres overlap.

Spinning. Spinning is the operation whereby the fibres in sliver form are once more drawn out and, by twisting, converted into a cylindrical form called a yarn, in which form the fibres are compressed against each other. The greater the twist, the greater the friction holding the fibers together, and (until the fibers are damaged by overtwisting), the greater the yarn strength or tenacity. The amount of twist imparted to the yarns in spinning is determined by the nature of the fibres. A rigid fibre will be damaged more readily than a pliant fibre and will withstand less twist. A smooth filament will slide more readily against contacting filaments and will require more twist to compress the fibres adequately.

The three essential operating components of the spinning machine are (a) the chain of pins, or gills, which hold the sliver as the fibres are drawn out by faster moving drawing rollers, (b) the flyer, which twists the yarn about a bobbin retained by a spindle, and (c) a winding mechanism, which winds the twisted yarn onto a large bobbin.

Rope-yarn spinning machines are designated as horizontal or vertical, depending upon whether the twisting flyer rotates about a horizontal or vertical spindle. Machines for spinning rope yarns of coarse natural fibres are also identified as chain gill (spinning jenny) or screw gill; the term "gill" identifying the comb of steel pins retaining the sliver. The screw-gill spinner, of British origin, is a machine in which the combs are moved by a screw thread. In the chain-gill machine, of American origin, the combs are linked together to form an endless chain of pinned combing bars.

In the spinning operation, the sliver is placed on the combing pins, led through retaining rollers or nippers about the flyer, and hence to a wind-up bobbin large enough to provide yarns in continual lengths of several thousands of feet.

The spinning of man-made filament yarns requires no combing pins. The most recently developed spinning machine used in the manufacture of such yarns, called the Twister-Winder, has no mechanical flyer but imparts twist as the yarn loop is spun about in lariat fashion.

Stranding machines

Stranding. Strands, also known as readies, are formed by twisting the yarns, or smaller size cords, together into the visible components of the rope structure. The respective yarns, unwinding from bobbins free turning on a suitable creel or frame, are "bunched" together, twisted, and wound onto large size bobbins. The stranding machines, called formers or bunchers, vary in size and form depending on ability to accommodate continuous strand lengths as well as on production rates, flyer speeds, and related weights. Their components correspond to those in the yarn-spinning machines with respect to the twisting and winding functions. One strand former, designed for heavy rope manufacture, has a full bobbin capacity of 1,500 pounds (about 680 kg), and is used to twist strands ranging in size from 1-inch to 2%-inch (2.54 to 6.67 cm) diameter, the 1,500 pounds of 1-inch diameter strand providing a continuous length of approximately one mile (about 1.6 kilometres).

***Rope** laying.* The twisted rope consists generally of three S-twist strands, twisted (laid) together in the direction of opposing twist (Z-twist), to maintain a torsional balance in the rope structure. The most common, three-strand rope, is also designated plain or hawser-laid rope; a four-strand rope, shroud-laid rope.

The rope-laying operations require machines similar to strand-forming machinery. The strands, on bobbins, are pulled through a compression tube and twisted into rope by a revolving flyer. As twisted, the rope is wound onto a heavy steel bobbin, also turning with the flyer. The three subassemblies of the rope-laying machine, arranged in tandem horizontally, are identified as the foreturn flyers (rotating strand bobbins), the capstan flyer (pulling mechanism), and the receiving flyer (rope-twisting and storage bobbin mechanism). The length of rope twisted in such a laying machine is limited by the dimensions of the receiving flyer.

In another type of horizontal rope-laying machine, the strand bobbins are arranged in tandem within a flyer. As each strand is pulled off its bobbin, it is overtwisted and in this condition combined with its adjacent strands into rope. This machine requires no receiving bobbin in its flyer; the rope is coiled directly into a reel form. The rope length, accordingly, is limited only by the strand length. Such machines may be designed with component parts arranged horizontally or vertically to minimize required floor space.

Rope-walk method

Rope making by the "rope walk" method is practiced extensively throughout the world even today, more than a century after the availability of rope forming and laying machinery. The rope-walk method requires three major pieces of equipment: a traveller with twisting hooks to form strands, a traveller with one or more hooks to twist the rope, and a foreboard to restore strand turns lost in laying the rope. Both travellers ride on rails, moved by a transmission rope drive, which also turns the twisting hooks. The strands formed on the strand traveller and stretched out for their entire length are secured to the twisting hook on the rope twisting traveller, now called the afterturn, at one end of the strand. At the other end of each strand, the strands are individually fastened to the foreboard twisting hooks, now called the foreturn. A grooved block called a top holds the strands separated at the afterturn and guides the strands into the proper positions as the strands make up the rope twist. When three or four strands are twisted together in their entirety, the rope laying is completed. The rope is then removed from the twisting hooks at either end of the rope walk and wound into shipping coil form.

In addition to the twisted structures, ropes in the size range of one-inch to five-inch (2.54-cm to 12.7-cm) diameter are also made in which four sets of strands, equally paired left twist and right twist, are braided into an eight-strand plaited structure. Another braided rope structure, identified as double-braided rope, consists of a layer of heavy twisted yarns, braided about a coarse braided rope core, to envelop and cover the core. Such braided ropes require specialized machinery and are used to best advantage where rope flexibility and torsional balance are prime service requirements.

Applications. To the extent that rope provides a flexible device for transmission of force from a source to a site of application, all ropes are used with a common purpose. Once used, all ropes are subject to deterioration — from mechanical stressing to environmental degradation. The rope user is always cautioned to exercise care to retard deterioration.

Breaking strength

When a rope is pulled, it stretches continuously, reducing in girth until a point is reached when the rope can stretch no longer and it breaks. This point identifies the breaking strength of the rope. The breaking strength, as the major identifiable rope property, with its stretch, is used to gauge the serviceability of the rope. The constructional factor, for a given fibre or filament, that can most influence the strength of the rope is the degree of twist in the rope and strands; the greater the twist, the lower the strength. Repeated loading in tension, short of breaking the rope, will have no adverse effect on the rope strength. Actually, provided the tensile load is not great enough to break inner strand yarns, repeated loadings may result in higher breaking strengths, due to better mutual adjustment of yarn and strand tensions as the rope is repeatedly stretched.

Natural-fibre ropes deteriorate most readily because of fibre degradation associated most commonly with mold growth. Man-made filament ropes deteriorate most readily when exposed to sunlight, elevated temperatures, or damaging chemicals, all of which accelerate filament decomposition.

Rope kinks or strand kinks (cockles) result from an unbalanced twist relationship in the rope structure, the consequence of improper handling. In this respect, the braided or plaited rope structure is superior to the twisted structure.

Considering the characteristics described, the rope user has considerable choice in the selection of a rope suited to optimum serviceability. The marine rope user, who consumes the major portion of manufactured rope tonnage, formerly preferred heavy twisted manila rope but now finds greater serviceablility in nylon or in polypropylene because of its higher strength and lighter weight. If stretch or flexibility are also important considerations, ropes of composite filaments or plaited structures enhance these characteristics.

The foregoing criteria are applied to many rope structures related to specific uses. Water skiing rope, for example, is generally polypropylene of braided structure to minimize kinking; mountain climbing rope, of nylon to develop high strength yet retain firmness; sail rope, of spun (short filament) polyester, braided or twisted to minimize stretch.

WIRE ROPE

History. It seems quite well established that wire rope was developed by Wilhelm Albert (1787–1846), a mining official of Clausthal in the Harz Mountains in Saxony, in a series of experiments extending over the years 1831–34. Even when first tried for hauling and hoisting in his mine, it proved so superior to hemp rope in serviceability and cost that its use soon became widespread in European mining. Known as the Albert Lay, the rope structure he developed was brought to England in 1839 by Lewis Gordon.

In England, wire rope structure and its manufacture became subject to extensive patent litigation. Stranded wire rope had already found extensive use not only in mines but also in railroad transportation and in heavy lifting where hemp ropes were proving too bulky and unwieldy.

Stranded wire rope

In the United States, John A. Roebling (1806–69), who had emigrated from Thuringia in 1831, appears to have been the first to start making stranded wire rope. Such rope replaced the heavy hemp ropes used on the Delaware and Hudson, and the Morris Canal portage inclines. The earliest wire ropes were over a mile (1.6 km) in length and three inches (7.6 cm) in diameter. It was not until after the Civil War that wire rope manufacture was begun on a really extensive industrial scale.

The stranded wire rope structure as developed corresponded to the present wire rope form in that individual wires were twisted about a hemp rope core to form the strand, six such strands then being twisted about a larger hemp rope core in reverse direction to form the rope. Before this, wire rope had already been made in the form of a selvagee—a bundle of individual wires stretched out into a long length and arranged parallel to one another, then bound together and covered with tarred hemp yarns. The stiffness of this structure precluded its adoption for hoisting and bending service, where flexibility and ease in handling are requisite. The selvagee structure, however, was adopted by the British Admiralty in 1838 for the standing rigging for the Navy's wooden ships. Its application for the construction of suspension bridges began in the United States with the erection of a 408-foot- (124-metre-) span footbridge over the Schuylkill River at Philadelphia in 1816. For this purpose, the suspension cables were each made up of six wires, each ⅜ inch diameter. The first wire footbridge in Europe was made at Geneva, over the Fosse River, completed in 1823.

Developments in wire rope making have kept pace with metallurgical developments. The major portion of wire rope manufactured now uses high-strength steel wire. Ropes for marine service, where corrosion resistance is paramount, are also made of phosphor bronze and stainless steel wires.

Manufacturing process. The manufacture of wire rope is basically similar to making rope from the natural yams or man-made filaments. The individual wires are first twisted into strands wound around a core; six strands usually, twisted about a core rope, are then laid into the rope. The cores are cord or rope structures made of steel wires; sisal, manila, henequen, jute, or hemp fibres; or polypropylene monofilaments. The function of the core is to provide a firm cushion for positioning the wires in the strands, to maintain a firm rope structure, and to provide some internal lubrication to extend the rope serviceability, particularly when bending stresses are involved.

The wire, as used in the rope-making operations, is received wound in coils and is first rewound onto spools or bobbins adapted for the strand-forming machines. In this rewinding, the outgoing end of one coil is attached to the new coil, either by twisting or welding, to insure a continuous length of wire in the strand.

Stranding

The design of the stranding machine resembles that of the strand former used in fibre and filament rope making, with the exception that the path of the bends to which the twisted strand will be subject as pulled and wound are of larger radius in order to cope with the stiffer wire structure without damaging or kinking the wires. The spools or bobbins containing the individual wires are arranged horizontally in tandem, within rotating flyers that provide the twist. As formed, the strand is tightly wound on a reel.

The strand diameter is attainable by adjusting the number of wires in relation to their size. Strand flexibility is enhanced when a larger number of finer wires are used. A strand made with coarser wires, however, allows the rope to better withstand abrasion and surface wear. Coarse wire strands are usually twisted with seven wires. The strand for general purpose, flexible wire rope usually requires 19 wires. Strands for especially flexible ropes are generally made with 37 wires.

In addition to these strand structures, there are many special structures combining smaller and larger diameter wires. The Warrington strand interposes *6* fine wires with *6* coarse wires to constitute the 12 surface wires in a 19 wire strand. The Seale strand maintains a coarse wire outer surface of 9 wires, and a concentric inner layer of 9 fine wires about a single coarse wire core, in a 19 wire strand: There are also combinations of Warrington and Seale strands; each structure designed to provide better serviceability under the operating conditions encountered.

The strands, wound on bobbins or reels, are now brought to the closer or layer, where the six strands are twisted together about a rope core, into rope form. These layers, similar to fibre rope layers, are designed with twisting components arranged horizontally or vertically. The largest size layer will provide a rope up to four inches in diameter, in a continuous length of approximately 2,000 feet, and with a weight of 27 tons: in metric units, a diameter of 10 cm, a length of about 600 metres, and a weight of 24.5 metric tons. As the six bobbins rotate within the rotating platform, the strands are compacted and twisted together about the rope core, laying or closing the rope. The strands, as the rope, are pulled through the laying machine by a pair of grooved drums turning as capstans. The finished rope, wound on a slow-turning reel, is now ready for use.

Lang lay ropes

Most wire ropes are layed with a right twist; each of the six strands respectively formed with a left twist. Such ropes are designated "regular Iay" or right, regular lay. Ropes twisted in opposing directions, designated left, regular lay, also find commercial use. In addition to the regular lay, ropes are also made with "Lang" lay. This designates a right or left twist rope lay, with the strand twist corresponding in direction to the rope twist. Lang lay ropes are named after John Lang, the manager of an English rope factory. With this construction, a greater portion of the strand surface wires are exposed, thereby affording greater area to withstand surface wear and to provide longer service life when severe bending stresses are encountered. The disadvantage of this construction lies in its torsional unbalance. In service, the Lang lay rope will untwist much more readily than regular lay rope, if the ends are free to rotate.

With either regular or Lang lay, the rope structure may be identified as "preformed" or "nonpreformed". The nonpreformed rope is twisted as described in the foregoing. In the manufacture of preformed wire ropes, both the strands and their wires are shaped during the respective rope- and strand-twisting operations, so that the helical twisted form to be assumed in the rope structure is preformed within the strander or layer, before the twisted structure is completed. The result is a rope in which the component strands and wires retain no residual stresses that would cause them to spring out of position when the rope is loose. This method also relieves the rope's tendency to rotate as it bends over a sheave. Both these features contribute to longer service life. The preformed rope is easier to splice and is much less likely to form kinks.

Applications. Most wire ropes are used in hoisting and hauling operations and machinery for these purposes, such as cranes, power shovels, elevators, mine hoists, etc. A flexible rope structure to cope with fast movement and bending stresses is an overall consideration. In other uses, such as support guys, stays, and barrier ropes, this property is not so much a consideration, and a 6×7 (six strands of seven wires each) structure will serve, as against a 6×19 rope.

Similar considerations prevail for marine ropes, which are usually galvanized. A stiff 6×7 structure is suited to

Comparative Properties of ½-inch (12.7-mm) Diameter Ropes

	nominal breaking strength (pounds)	approximate weight per foot (pounds)	stretch to breaking point (percent)
Natural fibres (3-strand twisted)			
Manila	2,650	0.075	13
Sisal	2,120	0.075	13
Henequen	1,590	0.075	15
Cotton	1,450	0.073	17
Man-made filaments (3-strand twisted)			
Nylon	6,400	0.065	45
Polyester	6,400	0.080	30
Polypropylene	4,200	0.047	35
Wires (6 x 19, regular lay, uncoated, fibre core)			
Iron	8,400	0.40	0.5
Steel			
Mild	17,000	0.40	0.5
Plow	18,800	0.40	0.5
Improved plow	21,400	0.42	0.5
Phosphor bronze	9,220	0.45	0.5
Corrosion resisting steel	22,800	0.46	0.5

rigging, mooring, and towing; but a 6×37 structure may be required when flexibility and ease in handling become overriding requirements.

Wire rope also finds extensive use in the form of lifting slings, cargo and construction nets, support spans in electric power transmission, and bridge and building construction.

The properties of wire rope in relation to the properties of natural and man-made fibre ropes are given in the Table.

WIRE AND WIRE PRODUCTS

Manufacture. The steel for wire manufacture may be made either by the open-hearth or the electric-furnace process. The steel mill provides the material for wire drawing in the form of hot-rolled rods, in long coiled lengths of 300 to 1,000 pounds, rod diameters ranging from 5 gauge (0.207-inch diameter) to 7/0 gauge (0.490-inch diameter). The metric equivalent would be weights of 140 to 450 kg, in diameters ranging from 5.26 to 12.4 mm.

As the rod is uncoiled, it is progressively descaled, washed, and coated with material to facilitate its being drawn through a die. When drawn, the wire is lubricated or coated with a protective finish, then recoiled and ready for use in wire rope or other manufactured products. Depending upon the grade of steel used, the wire can attain an ultimate strength as high as 250,000 pounds per square inch or 17,600 kg per square centimetre.

The rod as received may require a preliminary heat treatment to improve the grain structure of the steel. In this operation, a number of rods are pulled slowly through a furnace well above a critical temperature established to bring out the required strength and hardness and cooled under controlled conditions in air or in molten lead or salt baths.

Pickling

In the next operation, the rod is immersed in a "pickling" solution to remove dirt and scale. The pickling solution may be a dilute solution of sulfuric acid compounded with an organic inhibitor to allay attack on the cleaned steel. The pickling solution is kept hot. The descaling may be accomplished with other acids, molten alkaline salts, or even by metallic-grit blasting. Descaled, the rod is washed thoroughly to remove and neutralize descaling material and is then immersed in a coating liquid containing lime with other ingredients that coat the rod. It is then baked and allowed to cool. The lime coating serves as a lubricant and flux, protecting the wire from surface damage during drawing.

The wire drawing die provides a conical-shaped hole, a portion of which is bored with precision to correspond to the required wire diameter. The extent of reduction in size from rod to wire bears on the quality of the wire. The wire drawing may accordingly be done in a single step or in multiple steps. In its simplest form, the die may be a conical-shaped hole in a tool steel block. In more involved form, it may consist of a tungsten carbide insert in a steel casing, set into a water-cooled jacket. Wire drawing is essentially a stretching process, the reduction in diameter providing correspondingly greater length. Each drawing further increases the strength and toughness of the wire.

Steps in wire drawing

The wire-drawing process consists of three steps: (1) The rod is first pointed to enable it to pass through the smallest bore die that will be used to attain wire size. (2) The pointed rod is run through the successive dies and drawing blocks; each die also provides prelubrication of the rod or wire, using soap, metallic stearates, or lime with grease or tallow. (3) The drawn wire is coiled on the drawing block, which consists of a rotating drum, water cooled, turning fast enough to take up and coil the wire as it is stretched. The coil is built up on a winding reel (called a stripper), which is a movable component part of the block.

Although the wire is cold-drawn, there is so much heat generated by the drawing friction that both die and water block are water cooled. The lubricant protects against die erosion and wire damage. The conversion of rod to wire is generally done in several successive stages of drawing, as a continuous process; each successive block turning at a progressively higher speed to take up the increased length attained with each drawing. When the proper wire coil length is reached, the stripper reel is lifted from the block and the coil of wire is released.

The wire, now designated bright wire, may be further annealed by heat treatment to relieve drawing stresses or treated with a surface lubricant for strand-forming operations. When a protective finish is required, the wires may be galvanized or tinned in a hot-dip or electrodeposition process. The galvanizing process reduces the tensile strength of the bright wire by 10 percent.

Wire diameter is almost universally given in terms of steel wire gauge number. Sizes range from gauge no. 37, with a diameter of 0.0085 in. (0.22 mm) to gauge no. 7/0, with a diameter of 0.49 in. (12.4 mm).

Other steel wire products. Other steel wire is fabricated into many products, the most important being nails, fencing, springs, reinforcing and prestressing concrete wires, and suspension bridge wire.

Nails. A wide range of sizes and styles of wire nails is now available. Wires are cut and shaped automatically from coils of wire of proper size. The usual bright finish is attained by a tumbling process. Protective treatments are galvanizing, tinning, bluing by oxidation, or cement coating.

Woven fencing. Fence-weaving machines are designed to cross weave horizontal wires (laterals or line wires) with vertical or diagonal wires (stays), in fabric-weaving fashion, so that the wire crossings form triangular, square, rectangular, hexagonal, or diamond-shaped meshes for farm and industrial uses. The wire sizes range from 9- to the 20-gauge wire used for light poultry netting. There are basically two types of woven fencing; the stiff stay, in which the stay is in one length constituting the width of the fencing, and the cut stay in which a short stay is twisted about or welded to the lateral to form the mesh. Fencing of this construction constitutes exceptionally strong reinforcement for concrete.

Chain link fence is twisted from 6- to 11-gauge wires called pickets, which are spirally wound and interwoven to form a continuous link fabric. The ends of the adjoining pickets are twisted or folded together to provide fencing free of knots or ties except at the top and bottom.

Ornamental fencing (lawn fence) uses crimped stays formed into semicircles at one end. The line wires are locked into the crimp of each picket; the ends of the pickets are twisted or welded into twisted 2-wire line wires, called cables.

Barbed-wire fencing. The high-speed machinery producing this fencing twists two wires (line wires) together and, at regular intervals of three to six inches (7.6 to 15.2 cm) apart, winds barbs about one or both of the line wires. The barbs are cut diagonally to produce the points. Most barbed wire is made with no. 12%-gauge line wires and 14-gauge barbs.

Springs. The common requirement for springs is that the wire be in an annealed, hard-drawn, or pretempered condition and that steel carbon content be 0.45 to 0.90 percent. This allows the spring to be fashioned easily, then to be heat-treated to develop elastic characteristics required for service.

Prestressed concrete tendons. Cold-drawn, uncoated steel wires of no. 6 gauge and heavier and twisted wire strands as large as $1\frac{11}{16}$-inch diameter are used as tendons in prestressing concrete. By stressing the tendons, either before or shortly after pouring, the concrete develops a better capability to resist tensile stresses. The wire must be stress relieved to develop proper ductility and have a tensile strength in the range of 235,000 to 250,000 pounds per square inch.

Bridge wire. The manufacture of wire for this purpose requires careful quality control to insure compliance with demanding specifications. For example, the wire must be uniformly galvanized so that the adherent zinc coating will not peel off as the wire is bent. The stretch must be at least 4 percent in a 10-inch (25.4-cm) length. Bridge wire is drawn to 0.196-inch (about 5 mm) diameter, from high-carbon steel rod, and is handled in coils five feet (about 1.5 metres) in diameter.

BIBLIOGRAPHY

Rope and fibres: BRITISH ROPES LTD., Ropes Made *from Nylon* (1966); H.R. CARTER, Modern Flax, Hemp, and Jute Spinning and Twisting (1925), Rope, Twine and Thread Making (1924); R. CHAPMAN, A Treatise on Ropemaking, rev. ed. (1868); D. HIMMELFARB, The Technology of Cordage Fibers and Rope (1957); G. LAWRIE, The Practical *Ropemaker* (1948); J.M. MATTHEW, Textile Fibers, 5th ed. by H.R. MAUERSBERGER (1947); S.B. MCFARLANE (ed.), Technology of Synthetic Fibers (1953); F.I. OAKLEY, *Long* Vegetable Fibres (1928); A.V. PRINGLE, The Mechanics of Flax Spinning (1954); UNITED STATES DEPARTMENT OF AGRICULTURE, Monograph No. 21, Abaca, Cordage Fiber (1954), Misc. Publ. No. 518, Fiber *Production in* the Western Hemisphere (1943); L. WEINDLING, Long Vegetable Fibers: Manila, Sisal, Jute, Flax and Other Related *Fibers* of Commerce (1953).

Wire and wire rope: AMERICAN CABLE CO., Wire Rope Users' Handbook (1932); A. POMP, The Manufacture and Properties of Steel Wire, trans. by C.P. BERNHOEFT (1954); WIRE INDUSTRY LTD., Wire Industry Encyclopedic Handbook (1963).

Wire fencing: J.L. SCHUELER, "The Design and Manufacture of Woven Wire Fence," Wire and Wire Products, 26:29–41 (1951); UNITED STATES DEPARTMENT OF AGRICULTURE, Farmers Bull. No. 1832, Farm *Fences* (1940).

(D.H.)

Rosales

In comparison with other orders of flowering plants, the rose order (Rosales) is not particularly large in terms of either the number of families it includes (three) or in the number of its species (about 3,200). It is, however, one of the most frequently encountered flowering plant groups in the temperate zones of the world, especially in cultivation. Some of the most important food plants occur in the order, and numerous ornamentals are of this alliance, among them roses (Rosa), flowering cherries (*Prunus*), mountain ash (*Sorbus*), spirea (Spiraea), firethorn (Pyracanthn), hawthorn (*Crataegus*), and various species of cinquefoil (*Potentilla*). Members of the order are most characteristic of the upper latitudes or higher altitudes, being common in the temperate zones of both the New and Old World, but there are also important tropical species.

GENERAL FEATURES

Woody nature of the order

Life-form, distribution, and abundance. In life-form, the order is extremely diverse, ranging from small herbs (strawberry and cinquefoil) to shrubs and woody climbers (roses and spirea) to small trees (mountain ash) and giant tropical trees (*Parinari,* Licania). There are many more woody or partly woody species than there are herbaceous kinds. Tropical lowland representatives in the New World are mostly woody shrubs to quite large trees, some more than a hundred feet tall, with extremely hard, often useful wood. These woody kinds are especially characteristic of the family Chrysobalanaceae.

Members of the rose order have a natural distribution over much of the world, but the group is especially abundantly represented in eastern Asia (Japan and China), North America (north temperate areas), and Europe. The largest genera are *Rubus* (blackberries and raspberries), with several hundred species; Crataegus (hawthorns), with a like number of species; *Potentilla* (cinquefoils), with 300 species; and Rosa itself, with about 150 natural kinds. The precise size of these genera is very difficult to state, especially in the blackberry and hawthorn groups. In these, there have been widespread hybridization and recombination of species formerly isolated from each other ecologically. With the destruction of forests and other barriers to interbreeding, two or more species have often come together, bred with one another, and the hybrids have backcrossed with the parents. The consequence is that estimates, for example, of blackberry–raspberry species range from 25 to 1,000 or more different kinds, depending upon the taxonomic recognition given to the many forms. If each variant is considered a species or a subspecific unit of some rank, the number of kinds becomes almost infinite, and such a classification system serves no useful purpose.

Economic importance. Numerous economically important food plants belong to the rose order, including apples (*Malus*), pears (Py*rus*), peaches (Amygdalu), apricots, and plums (all three, *Prunus*). That is, most of the more important fruits produced in the temperate areas belong to this one order and to one of the families, the rose family (Rosaceae). Apple culture, based on crosses and selection from the common crab apples, is the economic base for several parts of the United States, including the northwestern states and New York, Michigan, and Virginia. In the southern U.S., strawberries (*Fragaria*) and peaches replace apples as the major fruit crops of the rose order.

Besides the many food plants of the order, there are also large numbers of ornamentals in temperate gardens. Of these, none is more widesyread, ayyreciated, or diverse in form than the cultivated roses. There are numerous wild species, but, except when planted for livestock enclosures, they are seldom used. The hundreds of cultivated varieties, however—climbers and bushes, dwarfs and standards, singles and doubles, and those with petals of every hue—grace the gardens of cottage and castle, ghetto and suburb. The damask rose is widely grown in Bulgaria for the valued perfume, called attar of rose and rosewater, which is distilled from the fresh flowers. An uncommon but historically interesting economic use was recognized and began to be exploited during World War II, when the fruits of wild rose species were gathered for their very high vitamin C content. The practice has continued among present-day advocates of herbs and health foods.

Many members of the Rosales order have less well known values. Ninebark (*Physocarpus*) is a shrub that in dense stands helps to prevent soil erosion and provides shelter for birds. Mountain ash trees (*Sorbus*) produce orange fruits that feed many birds and can even be used for jelly making. Shadbush (*Amelanchier*) bears fruits used by birds, and its leaves, bark, and young twigs feed deer and rabbits. Hawthorns (Crataegus) provide excellent cover for wild animals, and the fruits make up a substantial part of the diet of some birds. The wood of the black cherry (*Prunus*) is important in cabinetmaking. The heavy, hard, close-grained wood of the hawthorn makes excellent tool handles, canes, and charcoal.

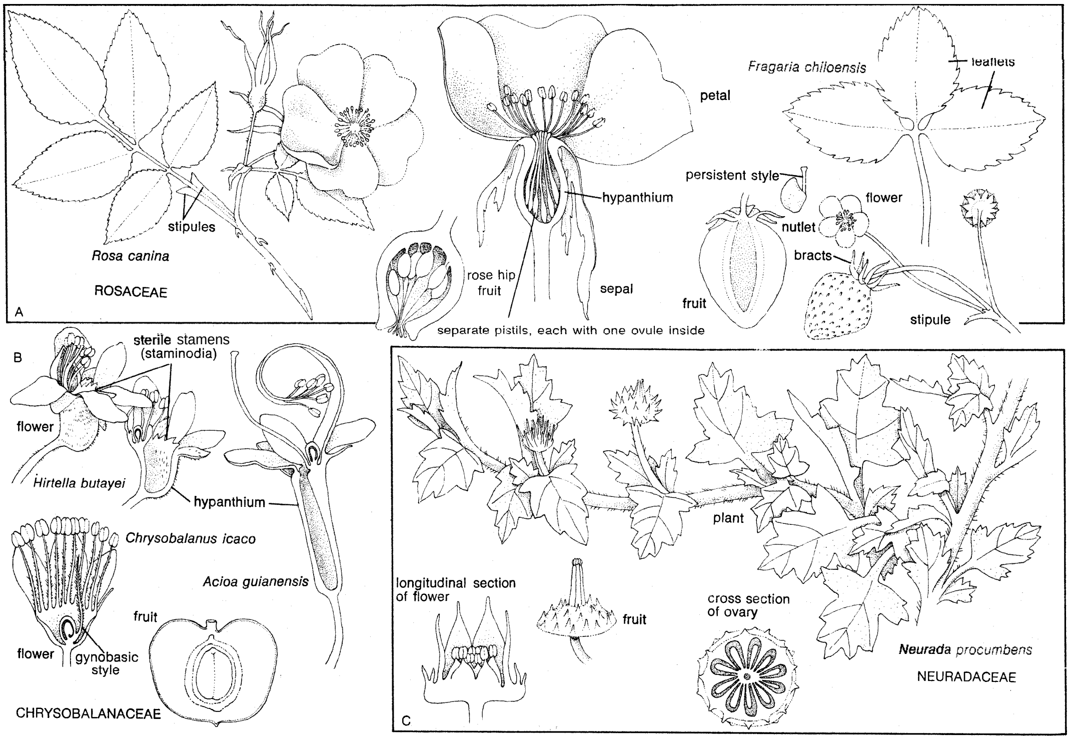
Medicinal uses

Contributions of the order for medicinal purposes are numerous, especially among the herbaceous species. Agrimony (Agrimonia), for example, was looked upon in past ages as a general cure-all for any sort of wound, snakebites, and even for removing warts. It has also been used commonly as a substitute for or as an additive to tea, and the entire plant yields a pale- to intense-yellow dye, depending on the season it is collected. It is still believed to be useful in the treatment of diarrhea, liver ailments, and as a gargle. The active principle is a tannin.

The petals of certain rose species are strongly odorous by virtue of a volatile oil they produce. This property gives rose petals value as flavouring for cough syrups and candies. Dried lose petals are often kept in potpourri jars or among clothing items, where their fragrance is slowly released. Fresh rose petals impressed in the surface of butter contained in a tightly covered container overnight in a cool place impart a delicate rose odour and taste to the butter. This, spread on small thin shapes of bread and garnished with a fresh rose petal, makes rose-petal sandwiches, often served with tea.

The leaves of hawthorn have been used as an adulterant of tea. Its flowers and fruits have astringent properties, which form the basis of their medicinal use as a cardiac tonic and as a diuretic for some kidney ailments. The fruits added to brandy produce a highly regarded liqueur.

In addition to the temperate-zone fruits of the rose family, there are also a few now widely cultivated species from the tropics and subtropics. The loquat (Eriobotrya) originally came from China, but it grows easily in warm places from southern Florida south. The coco plum (Chrysobalanus), a native fruit in the tropics and subtropics, grows along sandy shores from Florida through the West Indies, south to Brazil. The sweet, dry, but rather tasteless fruits are borne on small shrubs or low trees of the strand and riverbanks. In Mexico, the bark, leaves, and rooks are used medicinally as an astringent; the leaves and fruits are the source of a black dye; and the



Vegetative and **floral features** of **the three families** in **the order** Rosales.

Drawing by M. Moran based on (A, left, right) reprinted with permission of Macmillan Publishing Co., Inc. from Taxonomy of Vascular *Plants* by G.H.M. Lawrence. Copyright 1951 by The Macmillan Company, and (A, middle, B) A. Engler, Syllabus *der Pflanzenfamilien,* and (C) A. Engler and K. Pranti, Die *Naturlichen Pflanzenfamilien*

oily pits are used for food, or they are strung on sticks to serve as candles. Oiticica oil is a product expressed from fruits of one of the species of the family Chrysobalanaceae that grows in northeastern Brazil, where it has long been used medicinally and for illumination. Large quantities of the oil are imported by the United States for use in place of tung oil in paints, in making printing inks, and in making linoleum.

NATURAL HISTORY

Life cycle—reproductive features. The evolutionary success of the rose order rests on a variety of adaptive advances over its progenitors. The embryo, for example, uniformly fills the seed and is not dependent on an enveloping endosperm (nutrient tissues in the seed), as are the embryos of some less advanced families.

The wide range of fruit types in the order also is thought to have had a high selective advantage in its evolution, because a variety of different forms would increase the chances of successful dispersal and establishment. In addition to the familiar edible types of fruits (*e.g.* apples and pears, cherries and plums, and strawberries), many species groups produce fruits with hooks and barbs that ensure their transport on animal's hair or on man's clothing from place to place.

Asexual reproduction

More important than the foregoing, however, is the frequency of a form of asexual reproduction known as apomixis, a term often used loosely to include all forms of nonsexual reproduction. Some species in the order may reproduce exclusively bv such nonsexual methods as runners (strawberries) or horizontal leafless stem structures called rhizomes (raspberries) and still retain their capacity, entirely or in part, to reproduce sexually by seed. Other species have lost all capacity for sexual reproduction. Obviously, apomictic reproduction confers considerable evolutionary advantage, for a species may achieve extensive distribution even in the absence of its principal pollinators or under conditions unfavourable to seed development and germination.

On the other hand, apomixis often produces such a multitude of forms that classification is more a matter of arbitrariness than a scientific attempt to express evolutionary relationships. In the rose and the apple subfamilies (Rosoideae and Maloideae), much of the taxonomic confusion, such as that concerning the genera *Rubus* and Crafaegus, is the result of hybridization, polyploidy (having multiples of the basic number of chromosomes), and apomixis. Two sexual species, each of which may have a multiple set of the normal chromosome complement, can hybridize but produce largely sterile offspring. The latter, however, can become well established in nature, with well-developed geographic distributions, by either total or partial apomixis. At the same time, a doubling of the chromosome complement (polyploidy) may re-establish partial or full fertility. All these routes to survival of new forms appear to have operated during the evolutionary history of the rose order. Grafting and budding are man-made asexual means employed for the perpetuation of desirable varieties of apples, pears, peaches, and other related fruits.

Although knowledge of the cytology (the study of cells, especially of chromosomes) of the order is incomplete, much has been recorded about chromosome numbers in the family Rosaceae. The spirea subfamily (Spiraeoideae) has a basic chromosome number of 8 or 9, and it may be interpreted as more stable evolutionarily than some of the other subfamilies. The rose subfamily (Rosoideae), which includes the highly variable raspberry genus (*Rubus*), has at least three basic chromosome numbers, 7, 8, or 9. The apple subfamily (Maloideae), including the taxonomically confusing hawthorns (Crafaegus), has a basic number of 17, and the plum group (subfamily Prunoideae) has 8 as its base number of chromosomes. The 17 chromosome pairs that are the basic number in the apple group have been explained by some researchers as being the result of hybridization and subsequent polyploidy between ancestral species of the spirea group (with its nine pairs of chromosomes) and the plum group (with its eight pairs).

Ecology. The ecological preferences of the Rosales order are diverse and wide-ranging. Altitudinally, species of herbs, shrubs, or small trees extend from sea level to Alpine meadows beyond the tree line. Many species are weedy, occurring in association with man's activities, and some are clear indicators of soil sterility. Fruit-eating birds drop seeds or other propagula (*e.g.*, whole fruits or plant pieces that can grow to establish new plants) along fencerows or in hedgerows, so blackberries, hawthorns, and wild cherries are commonly found in such places. Members of the order are found in deserts, in the tropical rain forest, and in every intermediate situation.

Not infrequently the distribution of a species is determined by the chemical nature of the soil, for many species are able to succeed only on soils derived from limestone. The cultivated fruits of the order, for the most part, prefer well-drained situations, and, for that reason, much of the commercial production is in hilly or rolling countryside. Soil type is less important, for these fruit trees thrive on even heavy clay soils, but they must be well aerated and freely draining.

FORM AND FUNCTION

Leaf position and forms

Vegetative characters. In habit the species of Rosales are about as diversified as possible, with every life-form from small creeping herbs to giant hardwood trees. Many of the included groups have thorns and prickles, but this is scarcely a universal or even distinguishing characteristic. Much more typical is the alternating position of the leaves on the stems; only rarely do they occur opposite each other. The leaves are often compound (*i.e.*, with several leaflets to a common leafstalk), as in roses, strawberries, and cinquefoils, but the pome-fruited members such as apples and peaches have simple leaves. Stipules, small leaflike appendages at the base of the petiole (leafstalk), also are characteristic of the order. They sometimes fall very soon after expansion, or they are not obvious because they are often joined with the petiole, as in the roses.

Hairlets on various parts of the plants are usually one-celled and either glandular or, more commonly, not. The leaves sometimes bear nectaries (nectar-producing glands) on the petiole, on the blade surface, or on each tooth along the leaf margin. Crystals of calcium oxalate are found in the leaves of many species, either solitary or in clusters. The wood in species of the Northern Hemisphere has small, numerous vessels (water-conducting cells) that often have spirally thickened walls, simple perforations (*i.e.*, open-end walls joining the cells into tubes), and wood fibres that have many bordered pits (tiny openings along the sides of the cells; the "borders" are structural complexities that distinguish this type of pit from "simple pits" found in many other plant groups).

On the basis of anatomy, only the plumlike species can be separated from the other groups of the rose family. The family Chrysobalanaceae, however, is quite distinct anatomically in numerous respects. The stomates (small pores in leaf surfaces) are of a type distinct from the rest of the order, and the leaf hairs are arachnoid (cobwebby) or even stellate (with radiating branches, starlike) or peltate (with a shield-shaped structure at the apex). Glands are often present at the base of the leaf in many species, and silica bodies (deposits of the glassy mineral silica) are found in leaf or young stem cells of many species. Wood vessels are large and occur as solitary elements lacking the spiral thickenings of the vessels in the rose family.

Reproductive structures

Flower and fruit characters. Flowers of the rose order are almost always bisexual; that is, with stamens (male reproductive structures) and pistil (the female reproductive structure) in each flower. When this is not true, the male flowers occur on one plant and the female ones on another individual of that species (the dioecious condition). The flowers are regular (*i.e.*, radially symmetrical) in most genera, but several of the Chrysobalanaceae have irregular (bilaterally symmetrical) flower structure, either in the perianth (the collection of both sepals and petals) or in the reproductive parts. The flowers are borne singly or in a wide variety of inflorescence (flower-cluster) types, which are either determinate or indeterminate. Determinate inflorescences mature from the top downward; that is, the central or terminal flower blooms first, and, since this is located at the growing tip of the inflorescence, no further buds are produced. Indeterminate inflorescences mature from the bottom upward, and new buds are continually produced at the growing tip of the inflorescence.

Sepals and petals usually are present, but, in some species, one or both series are lacking. The perianth parts are usually joined with each other, in part, to form a cupular to elongate tubular structure called a floral cup or hypanthium, the wall of which may also involve the bases of adnate (fused, joined) stamen filaments. At the bottom of the cup, there is often a ring of secretory or glandular tissue that produces nectar.

The stamens are commonly numerous, occurring in one to several whorls of five stamens each and arising from the inner surface or rim of the hypanthium. The filaments are free from each other except in several genera of the family Chrysobalanaceae, in which they are joined basally to varying degrees. In one genus the filaments are joined completely in a ribbonlike structure on one side of the flower.

The gynoecium (female structural complex—the pistil or collection of pistils) in this order is usually superior; that is, the sepals, petals, and stamens are attached below the base of the ovary. The large group of stone- or pome-fruited genera (apples, plums, pears), however, have a distinctly inferior ovary with the perianth and stamens arising above the compound, two- to five-parted gynoecium. The remainder of the genera have two to ten carpels (structures comprising the ovary) in each superior gynoecium. The style, or styles when multiple, is terminal on the ovary or arises from its base. The erect or pendulous ovules (seed precursors) number one or more per carpel.

Fruits of the rose order are small and dry achenes (dry one-seeded structures that do not split open along definite lines) as in the strawberry and cinquefoil, follicles (dry structures that split open along one definite suture) as in spirea, pomes (fleshy fruits with cores) as in the apple and quince, drupes (fruits with stony pits) as in the plum and peach, or an aggregation of drupelets as in the blackberry and raspberry. The seeds have small embryos and endosperm is usually lacking.

Biochemical features. Although concerted biochemical studies of the Rosales have not been undertaken, there are numerous data, some of which have special interest for taxonomists. A few examples may be instructive, but generalities at the level of the order are virtually impossible at present.

Certain amino acids occur with such widespread frequency in stem tips as to be of little systematic (*i.e.*, relating to the classification of species) value. On the other hand, one of these (arginine) is the amino acid commonly present in the family Rosaceae. Elsewhere it makes only sporadic appearance. Similarly, certain fatty acids such as licanic acid are characteristic of some members of the family Chrysobalanaceae.

While relatively few sugar alcohols are found naturally, some are distinctive of major flowering-plant groups. Thus, sorbitol is infrequent generally but common among members of the rose family as a precursor of starch. It has been reported that leaves of certain Rosaceae, kept in the dark for a period to eliminate stored starch, will manufacture starch in the leaves when they are floated on a sorbitol solution.

Cyanide formation

Members of another class of compounds yield cyanide plus a simple sugar when they are broken down by enzyme action. These cyanogenetic materials are probably of systematic importance by virtue of their distribution, but the supporting data are incomplete. It is interesting to note, however, that these compounds are apparently almost totally restricted to the advanced vascular plants. Of those with cyanide-generating capabilities, the family Rosaceae has more species (150) than any other family. There may also be plant populations that produce cyanide within species that are otherwise not cyanogenetic. Still more significant is the distribution of this characteristic within the rose family: cyanogenesis is common among the pome-fruited species and the plum subfamily but is much less frequent in the rose and spirea subfamilies.

Numerous groups of species of the rose family are dangerous because of the cyanide compounds contained in various parts of the plant. Mountain mahogany (species of *Cercocarpus*), evergreen, small-leaved shrubs of arid areas in the western United States, is browsed by animals, often with impunity, but at some sites the concentration of cyanide is high enough to be fatal to livestock. Similarly, but much more dangerous to livestock and to humans, several kinds of wild cherry, peach pits, and bitter almonds are poisonous. Ordinarily, grazing animals avoid such poisonous foliage as that of chokecherry (Prunus virginiana) and wild black cherry (Prunus serotina), but either by accident or by elimination of other more desirable browse, animals are sometimes lost to cyanide poisoning from these plants. The cyanide is higher in wilted leaves and branch tips than it is in mature or dried leaves.

Some phenolic compounds also are distinctive of the family Rosaceae or of some of its subgroups. Isoflavones, for example, have been found only in the families Rosaceae and Leguminosae (also called Fabaceae, order Fabales), which is not surprising to phylogeneticists who consider the two families to be closely related.

EVOLUTION

Fossil record. As with most flowering-plant groups, fossil remains become increasingly common in the Tertiary Period (beginning about 65,000,000 years ago). Fossils are frequent in both European and North American formations of Eocene to Pleistocene age (*i.e.*, from about 54,000,000 to within the last 2,000,000 years). It is significant to note that fossilized parts have been found of plants readily assignable to all the subfamilies of the rose family and of the largely tropical family Chrysobalanaceae as well.

The genus Spiraea is known from fossils of its fruits, and there are leaf remains probably assignable to this group also. Ninebark (Physocarpus), a genus of modern-day shrubs, is represented in fossils of the mid-Tertiary. In the pome-fruited subfamily (Maloideae), seed remnants have been recognized of the hawthorn (*Crataegus*) and pear (Pyrus) genera. Leaf fossils are described for the quince (*Cydonia*), serviceberry (*Amelanchier*), and hawthorn. In the rose subfamily (Rosoideae), fruits of cinquefoil (Potentilla) and raspberry (*Rubus*) are known from the Pliocene Epoch (2,500,000 to 7,000,000 years ago) and Oligocene Epoch (26,000,000 to 38,000,000 years ago) of western Europe. The stone-fruited subfamily (Prunoideae) is represented by fossil fruits of both plum (Prunus) and peach (*Amygdalus*) in the Pliocene of Germany and France. The family Chrysobalanaceae is represented by leaf fossils in the late Eocene flora of west central Oregon. This group, now largely tropical, was a part of a flowering-plant assemblage that unquestionably flourished in a tropical environment there, about 50,000,000 years ago.

Phylogeny. The fossil record characteristically is not very helpful either for relating the family to other families or the order to other orders. The speculations that follow are based more on comparative morphology and anatomy than they are on fossil evidence.

Evolutionary position

The order Rosales is considered to be one of the more primitive orders, derived from the Magnoliales through the modifications of one intermediate order, the Dilleniales. The genus Spiraea has sufficient similarities to the latter order and to the Saxifragales to indicate common ancestry. Likewise, the family Chrysobalanaceae is seen by many as the connecting link between the Rosales and the Fabales.

Within the order, the most primitive family is often considered to be the Chrysobalanaceae, with its superior ovary composed of a single carpel and the style arising from the ovary base. The family Rosaceae is probably even more basic, ranging as it does from having quite simple floral structures to very complex, from woody to herbaceous life-forms, and with so many fruit types that give the family as a whole great selective advantage potentially over groups with less adaptive flexibility. The rose family consists of at least four subfamilies, which differ principally in the characters of their fruits. The family Neuradaceae probably is most nearly related to the rose family through the rose subfamily, but it is quite distinct in the wall structure of its pollen grains, as well as in the structure of its flower and fruit.

CLASSIFICATION

Distinguishing taxonomic features. The features that distinguish the Rosales from related orders are several. Vegetatively, many species have compound or divided leaves that commonly are stipulate. The flower parts are usually in fives, arranged in concentric circles. Although the relationship among the floral organs varies from hypogyny (superior ovary) to epigyny (inferior ovary), the most characteristic condition is perigynp (half-inferior ovary), the sepals, petals, and stamens arising from the rim of a cupular to tubular hypanthium. The stamens are usually in many whorls, and the gynoecium consists of free to united pistils.

Annotated classification. The classification presented here is a recent one that is relatively conservative in approach. Only three families are included, in contrast to as many as 17 families in older treatments. Likewise, the family Rosaceae has been divided by some authors into six separate families, which are treated here as four subfamilies. Although it represents perhaps the narrowest view taxonomically of what the order includes, it is thought to be more natural evolutionarily than more inclusive treatments.

The following is a more technical, highly abbreviated summary of the descriptions of the major groups comprising the order Rosales.

ORDER ROSALES

Usually stipulate, often woody plants with pentamerous (parts in 5s) flowers, many free stamens, and hypogynous to epigynous arrangement of the floral organs. Seeds usually lacking endosperm. Wood vessels with simple perforations. Three families with 115 genera and 3,200 species distributed in Africa, Northern Hemisphere, and Neotropics.

Family Rosaceae

Trees, shrubs, and perennial or (rarely) annual herbs with simple or compound, mostly stipulate leaves, the seeds, bark, and leaves with abundant hydrocyanic acid, as well as amygdaiin, prunasin, prulaurasin; chromosomes 7, 8, 9, or 17 in each egg and sperm cell. Flowers usually complete, sometimes apetalous (without petals), mostly perigynous, rarely epigynous or almost hypogynous, usually with a disk, the hypanthium flat, cupular, urceolate (urn-shaped), or tubular; perianth mostly 5-parted, rarely 3- or 4- or more than 5-parted, the calyx (sepals) sometimes with an outer calyx of an equal number of alternating members; stamens indefinite in number, or 2 to many times as many as the sepals, arranged in 2 or more series, stamens rarely equal in number to the sepals or only 1; carpels numerous to only 1, mostly free at the base of a concave floral cup or of the expanded, invaginated apex of the floral axis, 1- or (rarely) incompletely 2-celled; ovules 2 or 1, infrequently several or numerous, anatropous (inverted and straight), rarely almost orthotropous (erect); fruit a follicle, achene, or drupe, sometimes united with the floral cup or expanded floral axis to form a compound fruit; endosperm sparse or lacking, rarely abundant. About 100 genera and 3,000 species, almost cosmopolitan, mostly in temperate zones, especially richly developed in the Northern Hemisphere.

Subfamily Spiraeoideae

Shrubs, rarely trees, perennial herbs, or half-shrubs, stipules often lacking; carpels 1 to 8 but mostly 5, free or incompletely connate (fused) basally or rarely higher, sometimes adnate (joined) dorsally to the inner wall of the more or less concave floral cup, which is bowl- or bell-shaped, rarely almost plane, the ovules 2 to numerous, arranged in 2 rows; fruits mostly follicles, rarely achenes or capsules. Seventeen genera in South America, Mexico, North America, Siberia, Asia, and Malaysia.

Subfamily Rosoideae

Shrubs or perennial herbs, rarely trees or annual herbs; carpels few to many, free on the expanded, capitate to more or less cylindric floral axis, or the carpels basal, mostly free but rarely weakly united with each other and with the floral cup, the ovules 1 or rarely 2, mostly with only 1 integument; fruit an achene, rarely a drupe, often united in a compound structure with the remnants of the floral cup or of the swollen axis apex. About 34 genera and 2,000 species, in most temperate to subarctic areas of the world.

Subfamily Maloideae

Shrubs or trees with simple, distinctly stipulate leaves; carpels 2 to 5, dorsally united to varying extent with the inner wall of the hollow floral tube, free among themselves or at most only incompletely connate, sometimes the carpels incompletely 2-celled by development of false walls, the styles free or (rarely) more or less united, the ovules erect-apotropous and mostly 2, rarely 1 or 3 per carpel, or horizontal and 20 to 24 per carpel; floral cup thickened at maturity in a more or less fleshy compound fruit enclosing the carpels. About 14 genera and 600 species in temperate Asia, East Indies, Mexico and Central America, and in North America.

Subfamily Prunoideae

Trees or shrubs with simple leaves, the stipules often small and deciduous, hydrocyanic-acid-derived glycosides often present; petals sometimes absent; carpel 1, very rarely 2 or 5, free at the base of the flat, bowl-shaped to tubular floral cup, the styles usually terminal, the ovules usually 2, pendulous; fruit a drupe with succulent, rarely leathery, mesocarp (middle layer of fruit tissues) and almost always a stony endocarp (internal layer of fruit tissues), the seeds solitary or exceptionally 2 in 1 carpel. Three genera with about 100 species, mostly in the North Temperate Zone, especially in east Asia; also in the Old World tropics.

Family Neuradaceae

Prostrate, sympodially branching (with 1 fork of a dichotomy assuming the status of the main plant axis), annual, xerophytic (with dry-habitat adaptations) herbs with lobed or sinuate (wavy-margined) simple, alternate leaves, the stipules mostly lacking; flowers solitary, 5-parted, more or less epigynous; carpels connate with each other and adnate with the inner wall of the mature, dry floral cup, the ovary (because of bisection of each carpel) 10-celled, with 10 persistent, indurate (hardened) styles, the ovules 1 or 2 in each cell, apotropous, the seeds lacking endosperm, germinating within the fruit. Three genera with 10 species in Africa and northwest India.

Family Chrysobalanaceae

Trees and shrubs, rarely half-shrubs with alternate, sometimes 2-ranked, almost always stipulate leaves; flowers rather inconspicuous, usually in various kinds of inflorescences but rarely solitary, mostly perigynous, the floral tube cupular to tubular, mostly with a disk, the perianth 5-parted or the petals sometimes absent; stamens many to 2, sometimes in concentric circles of 5 to 10 stamens in each; in irregular flowers the stamens often rudimentary and staminodial (sterile or partly so) on the side of the flower opposite to the ovary, the filaments free or more or less united; carpels 3 or 2, of which only 1 usually develops, the ovary rarely central, almost always unilateral and excentric because of the lateral union of the more or less elongate gynophore (stipe or stalk of the ovary) with the floral cup, the carpel 1-celled, rarely 2-celled by the development of a false wall, the style basal, sometimes with an indistinctly 3-lobed stigma. About 12 genera and 300 species, pantropical but mostly in the Amazon region of South America.

Critical appraisal. This classification, as noted earlier, is a most conservative one, differing from others in the few families included in the order. The difficulty in extending the ordinal concept beyond these three families is that distinctiveness of the order is lost as family after family is added. This treatment is based on what appears to be a logical point of separation from related families. Some recent authors include up to 19 families, some of which are subdivided by other workers into still other families. Solution of this situation may not be possible in a manner satisfactory to even the majority of botanists. Much of the present lack of agreement is based on the interpretation of abundant data of many kinds. Biochemical, paleobotanical, and anatomical researches perhaps hold greatest promise for further refinement of the order.

BIBLIOGRAPHY. L.H. BAILEY, *The Standard Cyclopedia of Horticulture,* 3 vol. (1947), a layman's guide to cultivated plants, identity, and culture, now somewhat out of date but still useful; J.J. and F.C. CRAIGHEAD, JR., and R.J. DAVIS, *A Field Guide to Rocky Mountain Wildflowers from Northern Arizona and New Mexico to British Columbia* (1963), an introduction to the field study of wild plants in nontechnical language with excellent colour plates; M. GRIEVE (comp.), *A Modern Herbal,* 2 vol. (1931), a popular treatment of common plants with folk uses for medicine, food, etc.; A.F. HILL, *Economic Botany* (1952), a classic text for the study of plants of the world that are useful to man; G.H.M. LAWRENCE, *Taxonomy of Vascular Plants* (1951), one of the best texts available for the study of taxonomy historically and in terms of learning characteristics of plant families; H.W. RICKETT, *Wild Flowers of the United States* (1966– ; with additional volumes added from time to time), a series of large volumes written in an easy to read, nontechnical style, profusely illustrated by colour plates and line drawings; E.T. WHERRY, *Wild Flower Guide: Northeastern and Midland United States* (1948), a popular account in lay language for the identification and appreciation of wildflowers.

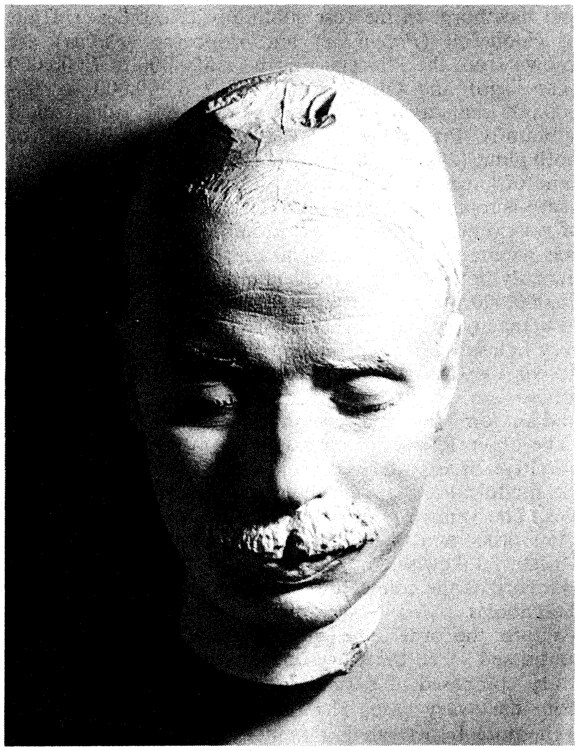
(R.S.C.)

Rosenzweig, Franz

Franz Rosenzweig, German-Jewish religious existentialist, was one of the most influential modern Jewish theologians. He was born December 25, 1886, in Kassel, Germany, the only child of Georg and Adele (*née* Alsberg) Rosenzweig. His father was a well-to-do dye manufacturer and member of the city council; his mother, a deeply sensitive and cultured woman. Franz grew up in an environment of civic responsibility and cultivation of literature and the arts; religious beliefs and observance were no longer evident, beyond perfunctory participation on some occasions. In his university days the gifted young man first started to study medicine (at Gottingen, Munich, and Freiburg) but after a few semesters turned to his real interest: modern history and philosophy (at Berlin and Freiburg). In 1910 he embarked on a study of Hegel's political doctrines. His doctoral dissertation (1912) was to become a section of *Hegel* und der *Staat* ("Hegel and the State"), a comprehensive work completed some years later. Yet while steeped in this research, Rosenzweig developed a critical attitude toward Hegel's overemphasis on history and his treatment of the individual person's life as irrelevant to the "whole." In rejecting Hegel, Rosenzweig opposed the philosophic movement known as German Idealism, with its attempt to construct reality out of abstract concepts. Increasingly he tended toward an "existential" philosophy that found its starting point in the experience and concerns of the concrete individual person.

Rejection of Hegel

By courtesy of N.N.Glatzer



Rosenzweig, death mask, 1929.

Some of his friends (especially the jurist and historian Eugen Rosenstock-Huessy), who were equally critical of the academic philosophy of the day, had found the solution to the problem of man in religious faith (specifically, conversion to Christianity) and in a dialogical relationship between man and God. After an intense inner struggle Rosenzweig decided in July 1913, to relinquish his Jewish heritage (barely known to him), to accept his friends' interpretation of modern Protestantism as an existential, dialogical faith, and to undergo Baptism. At this critical point in his life, however, he attended the Day of Atonement service in a small, traditional synagogue in Berlin (October 11, 1913). The liturgy of this fast day focusses on the motifs of human sinfulness and divine forgiveness, the realization of life as a standing before God, the affirmation of the oneness of God and of his love, The drama of the liturgy had a powerful effect on Rosenzweig. What he thought he could find only in the church—faith providing an orientation in the world—he found that day in the synagogue. He felt he had to remain a Jew. There followed a period of self-examination to determine whether the emotional experience of that Day of Atonement would stand up to rational criteria. After this clarification, Rosenzweig was determined to devote his life to the study, teaching, and practice of Judaism. The academic year 1913–14 was entirely devoted to an intensive reading of classical Hebrew sources and to attending lectures by Professor Hermann Cohen, an eminent German-Jewish thinker, the founder of the Neo-Kantian school in philosophy.

The making of *The Star of Redemption*

With the outbreak of World War I, Rosenzweig joined the armed forces and spent most of the duration of the war at the Balkan front, in an anti-aircraft gun unit. The not-too-demanding service allowed Rosenzweig time for study and writing. In 1916–17 he engaged in an exchange of letters with Rosenstock-Huessy on core theological problems in Judaism and Christianity, published in *Judentum und Christentum* (Eng. trans., *Judaism Despite Christianity,* 1969), wrote newspaper articles on political and strategic questions, drew up a plan for a reform of the German school system, and wrote "Zeit ist's" ("It Is Time"), a program for a reorganization of Jewish education and scholarship (included in *On Jewish Learning,* 1955). In 1918, while attending an officers' training course near Warsaw, in German-occupied Poland, he had opportunity to observe the life and the customs of east European Jews and was deeply impressed by the vitality and richness of their faith. Upon returning to the trenches he felt ready to embark on what was to become his magnum opus: an existentialist religious philosophy demonstrating the mutual relationships between God, man, and the world. This "new thinking" is based on human experience, common sense, and the reality of language and dialogue. The central point of the architectonically arranged work in which this thought is expressed is the act of "revelation" in which God in his love turns to man and awakens within him the consciousness of an "I." *Der Stern der Erlösung,* completed in 1919, appeared in 1921 (Eng. trans., *The Star of Redemption,* 1971). The work has been ignored by the various trends in academic philosophy, but highly regarded by existentialist and, especially, younger Jewish theologians.

The paralytic years

In early 1920 Rosenzweig married Edith Hahn of Berlin and wrote "Bildung und kein Ende" (included in On *Jewish Learning* as- "Towards a Renaissance of Jewish Learning"), outlining a plan for a Jewish adult study centre. Later in the year he was appointed head of such a centre (the Freies Jiidisches Lehrhaus) in Frankfurt am Main. There students were encouraged to examine classical Hebrew sources, searching for what is vital and relevant. The school became a model for similar institutions elsewhere in Germany. Rosenzweig's active directorship did not last long; early in 1922 he was afflicted with a progressive paralysis (from a form of sclerosis). In September 1922 his son Rafael was born. The child brought comfort to the father, whose paralysis affected his whole body, including the vocal organs. In a true heroism of the spirit, although unable to speak or write in a direct physical sense, he managed to continue living as an active scholar, writer, and friend, deeply concerned for his fellowman and community. With the help of his wife, a system of signals between them, and a specially constructed typewriter, he produced important essays and an annotated German version of the medieval Hebrew poetry of Judah ha-Levi. From 1925 on he embarked, together with Martin Buber, the eminent German-Jewish philosopher and biblical interpreter, on a new German translation of the Hebrew Bible (Old Testament). The translation occasioned a series of articles by him on aspects of biblical thought and style. As a hobby he also wrote reviews of records of classical and sacred music. Nowhere in these works of his paralytic years did the reader detect that the author was mortally ill. Everywhere in them there is evidence of a fresh, keen spirit, intellectual clarity, religious faith, and a sense of humour. He died at Frankfurt on December 10, 1929. His influence on Jewish religious thought grew remarkably in the decades after his death.

BIBLIOGRAPHY. FRANZ ROSENZWEIG, *Judaism Despite Christianity* (1969), the Rosenzweig/Rosenstock-Huessy correspondence; *On Jewish Learning* (1955), contains Buber-Rosenzweig correspondence and essays on the title theme; *Der Stern der Erlösung* (1921; 3rd ed., 1954; Eng. trans., *The Star of Redemption,* 1971); ALEXANDER A. ALTMANN, "Franz Rosenzweig and Eugen Rosenstock-Huessy: An Introduction to Their Letters on Judaism and Christianity," *Journal of Religion,* 24:258–270 (1944), the biographical background of this correspondence; NAHUM N. GLATZER (ed.), *Franz Rosenzweig: His Life and Thought,* rev. ed. (1961), containing a biography, based on letters, diaries, and autobiographical fragments and selections from Rosenzweig's works.

(N.N.G.)

Rossetti Family

An Anglo-Italian family in Victorian England, the Rossettis formed a remarkable group. All its members were endowed with unusual intelligence, were equally at home with the languages and literary traditions of both England and Italy, and were united among themselves by close ties of affection and mutual understanding. The talents and characteristics they shared were combined with creative gifts of a high order in Dante Gabriel Rossetti and his sister Christina Georgina Rossetti, both of whom wrote inspired poetry. The literary critic Sir Edmund Gosse did not exaggerate in acclaiming Christina as "[a] wonderful woman who stands almost alone in the forefront of the world's female poets." Dante Gabriel Rossetti was not only both poet and painter but was also possessed of great personal magnetism. His creation of a stimulating intellectual climate, his divination and encouragement of talent in others, and his poetry cast him in a role in England comparable with that of the poet Charles Baudelaire in France.

The elder Rossettis. The father of Dante Gabriel and Christina was Gabriele Pasquale Giuseppe Rossetti, an exiled Italian patriot and man of letters. Born in 1783 at Vasto in the Abruzzi, he was the son of a blacksmith and was clever enough to be able to study at the University of Naples. In 1807 he was librettist at the San Carlo opera house in Naples and was later appointed curator of Ancient Marbles and Bronzes in the Museo e Gallerie Nazionale di Capodimonte, Naples.

Gabriele frequently improvised spirited verses on contemporary politics; one indignant outburst directed against Ferdinand II, the tyrant king of Naples who had revoked the Constitution in 1821, added to Gabriele's membership in the revolutionary society Carbonari, provoked the sentence of death.

After a time in hiding, he escaped to England via Malta in 1824. There he supported himself by giving Italian lessons and in 1831 was appointed professor of Italian at King's College, London, a post he held until 1847, when his sight was seriously impaired. When Gabriele became professor, he received fellow exiles, busied himself with propaganda for a liberally governed and united Italy, and—studying the writings of Italy's 14th-century national poet Dante Alighieri-elaborated a theory of Dante's profound symbolic meaning (*La Beatrice di Dante,* 1842). His love of Italian poetry and in particular his reverence for Dante's *La vita nuova* and *La divina commedia* were transmitted to Gabriele's sons and daughters.

In 1826 he had married Frances Mary Lavinia Polidori, deemed the cleverest and best looking of four daughters of another Italian teacher and man of letters, Gaetano Polidori, Tuscan by birth but Londoner by adoption. Since Polidori had married an English girl, it is possible to consider the children of Gabriele and Frances as three parts Italian and one part English. Frances wished for a husband and children distinguished by intellect—yet later remarked that, though her wish had been granted, she could have done with "a little less intellect in the family so as to allow for a little more common sense!"

The Rossetti children

There were four Rossetti children: Maria Francesca (born February 17, 1827–died November 24, 1876); Gabriel Charles Dante, who later called himself Dante Gabriel (born May 12, 1828–died April 9, 1882); William Michael (born September 25, 1829–died February 2, 1919); and Christina Georgina (born December 5, 1830–died December 29, 1894). All were born in London and baptized in the Church of England.

The younger Rossettis. *Maria.* Maria, who for a long time taught Italian and was a household authority on niceties of grammar and translation, also produced a scholarly commentary, *A Shadow* of *Dante* (1871). Dante's conception of a spiritual and ideal love strongly influenced Maria's brother Dante Gabriel.

Dante Gabriel. After a general education in the junior department of King's College (1836–41), where he may have had lessons from the watercolourist John Sell Cotman, he hesitated between poetry and painting as a vocation. When about 14 he went to "Sass's," an old-fashioned drawing school in Bloomsbury (Central London), and thence, in 1845, to the Royal Academy schools, where he became a full student.

Meanwhile, he read omnivorously—romantic and poetic literature, Shakespeare, Goethe, Byron, Scott, and Gothic tales of horror. He was fascinated by the work of the American writer Edgar Allan Poe, and a dramatic illustration of Poe's poem "The Raven" was one of a number of early drawings inspired by his reading. In 1847 he discovered the 18th-century English painter, poet, and visionary William Blake through the purchase of a volume of Blake's designs and writings in prose and verse; the volume has since been known as the Rossetti MS. The diatribes of Blake against the painter and art critic Sir Joshua Reynolds delighted Rossetti and encouraged him to attempt lampoons of his own against convention and the triviality of early Victorian paintings of anecdotal subjects, those of Sir Edwin Landseer being a special target of his derision.

By the time he was 20, he had already done a number of translations of Italian poets and had composed some original verse, but he was also much in and out of artists' studios and for a short time was (in an informal way) a pupil of the painter Ford Madox Brown. He acquired some of Brown's admiration for the German "Pre-Raphaelites," nickname of the austere Nazarenes, who had sought to bring back into German art a pre-Renaissance purity of style and aim. It remained to initiate a like reform in England.

Aims of the Pre-Raphaelites

Largely through Rossetti's efforts, the English Pre-Raphaelite Brotherhood was formed in 1848 with seven members, all Royal Academy students except for William Michael Rossetti. They aimed at "truth to nature," which was to be achieved by minuteness of detail and painting from nature outdoors. This was, more especially, the purpose of the two other principal members, William Holman Hunt and John Everett Millais. Rossetti expanded the Brotherhood's aims by linking poetry, painting, and social idealism and by interpreting the term Pre-Raphaelite as synonymous with a romanticized medieval past.

In this the High Church movement, leaning toward the ritual and ornament of the Gothic period, influenced him, its aesthetic aspect attracting him apart from any question of religious belief. He entitled his manuscript poems *Songs of an Art-Catholic;* and, while his first two oil paintings—"The Girlhood of Mary" and "Ecce Ancilla Domini" ("The Annunciation")—were simple in style, they were elaborate in symbolism. The "art-catholic" appears, though not with any specifically religious intent, in the rich word-painting and emotional force of his poem "The Blessed Damozel," published in 1850 in the first issue of *The Germ,* the Pre-Raphaelite magazine.

Poor engravings of 14th- and 15th-century frescoes in the Campo Santo at Pisa had vaguely suggested an ideal for the Brotherhood to follow. But Rossetti gained a better idea of the past from seeing the work of the early Flemish masters at Bruges, which he visited in company with William Holman Hunt in 1849.

Exhibited in 1850, "Ecce Ancilla Domini" received severe criticism, which Rossetti could never bear with equanimity. In consequence, he ceased to show in public and gave up oils in favour of watercolours, which he could more easily dispose of to personal acquaintances. He also turned from traditional religious themes to scenes from Shakespeare, Robert Browning, and Dante, which allowed more freedom of imaginative treatment. After 1856 he was led by Sir Thomas Malory's *Morte Darthur* and Tennyson's *Idylls of the King* to evoke an imaginary Arthurian epoch, with heraldic glow and pattern of colour and medieval accessories of armour and dress.

The 1850s were eventful years for Dante Gabriel Rossetti. They began with the introduction into the Pre-Raphaelite circle of the beautiful Elizabeth Siddal, who served at first as model for the whole group but was soon attached to Rossetti alone and, in 1860, married him. Many portrait drawings testify to his affection for her.

Second phase of Pre-Raphaelite movement

In 1854 he gained a powerful but exacting patron in the art critic John Ruskin. By then, as Christina Rossetti signalized in some quietly humorous verses, the Pre-Raphaelite Brotherhood was at an end, splintered by the different interests and temperaments of its members. But Dante Gabriel Rossetti's magnetic personality aroused a fresh wave of enthusiasm. His contributions to the Oxford and Cambridge magazine in 1856 brought him into contact with the then Oxford undergraduates Edward Burne-Jones and William Morris. With these two young disciples he initiated a second phase of the Pre-Raphaelite movement. The two main aspects of this fresh departure were a romantic enthusiasm for a legendary past instead of the realism of "truth to nature" and the ambition of reforming the applied arts of design. Rossetti's influence not only led to easel pictures illustrating Arthurian legend but also into other fields of art. A new era of book decoration was foreshadowed by Rossetti's illustration for the Moxon edition of Alfred Lord Tennyson's *Poems* (1857). His commission in 1856 to paint a triptych ("The Seed of David") for Llandaff Cathedral was a prelude to the ambitious scheme of 1857 to decorate the Oxford Union debating chamber with mural paintings of Arthurian themes.

Though Rossetti and his helpers (Burne-Jones, Morris, and others) failed through want of technical knowledge and experience, the enterprise was fruitful in suggesting that the scope of art could be expanded to include the crafts. The idea of an association to make by hand objects of practical use superior to the mechanical products of Victorian industry became Morris' great ambition. The outcome, in 1861, was a firm of "fine art workmen." The firm known as Morris, Marshall, Faulkner & Co. depended mainly on Morris, Burne-Jones, Rossetti, the architect Philip Web, and Madox Brown for the design of stained glass, wallpapers, chintzes, tiles, embroideries, and furniture. Although Morris was the moving spirit, Rossetti's designs for stained glass for a number of new churches were among the early successes of the firm.

William Michael. Meanwhile, the support of the Rossetti family devolved upon Dante Gabriel's brother William Michael Rossetti. Since 1845, when at the age of 16 he became a clerk in the Excise (later Inland Revenue) Office at £80 a year, he had been a mainstay. His appointment as art critic to *The Spectator* magazine *in* 1850 and subsequent modest advancement in the civil service enabled him, in 1854, to establish his father, mother, and two sisters in a more comfortable home. His mother and his sister Maria helped by giving lessons in Italian, though efforts to start a school, first in Camden Town and then at Frome in Somerset, came to nothing.

William Michael was 45 when he married, in 1874, and would have continued to keep his mother and Christina as part of his household if his wife, Lucy, daughter of Ford Madox Brown, had not demurred. But, though quietly self-sacrificing in the family interest, he was by no means a nonentity. He had sterling qualities as a critic and Pre-Raphaelite historian. Independent in judgment, he hailed Walt Whitman's controversial *Leaves of Grass* (1855) as a work of genius, in spite of the scoffing remarks of his brother Dante Gabriel. His enthusiasm for Blake was no less great than Dante Gabriel's. He was the judicious editor of numerous collections of the work of English poets and could make learned comparison between such medieval writers as Geoffrey Chaucer and Giovanni Boccaccio. He dealt conscientiously with a vast amount of family correspondence and material relating to Pre-Raphaelism and his brother's place in the movement, proving himself an indispensable chronicler.

Christina. If William Michael was in many ways a contrast to Dante Gabriel—in his calm and rational outlook, financial prudence, and lack of egotism, for example—Christina Rossetti in certain ways resembled the elder brother. She, too, was a born poet. No invasion of Polidori aunts could distract her from the exquisite lines she often wrote while standing at the washstand of her bedroom. Like Dante Gabriel, she combined a melancholy and tormenting conscience with a lively sense of humour and a keen critical perception. She suffered much from ill health and, in the 1870s, from the painful and disfiguring Graves' disease (a disease of the thyroid gland); but religious scruple also caused suffering.

Effectof religion on the family

Religion affected the members of the family differently. William Michael, who had imbibed Gabriele's antipapalism, disliked and rejected religion in any form. On the other hand, Maria Francesca was devout, with a yearning toward convent life that impelled her to join the (Anglican) Community of All Saints three years before her death, which occurred in 1876. Christina was haunted by an ideal of spiritual purity that demanded self-denial. This, perhaps combined with a sectarian strictness, has been suggested as the reason for her never having married. There may have been other reasons, too, but ostensibly she broke off her engagement to James Collinson, a painter whom her brother had introduced into the Pre-Raphaelite Brotherhood, because he left the Church of England and became a Roman Catholic. Fourteen years later, in 1864, she refused Charles Bagot Cayley, one of her father's old pupils, supposedly because he had no religious belief, though they remained on friendly terms until Cayley's death, in 1883.

Christina began to write poetry at an early age; the verses written before she was 17 were printed in 1847 on Gaetano Polidori's private press. The transience of material things is a motif that recurs throughout her poetry. The verses beginning "Passing away, saith the World, passing away," from *Old and New Year Ditties 111,* were hailed by the poet Algernon Swinburne as "the noblest of sacred poems in our language." The title poems of two major collections, *Goblin Market . . .* (1862) and *The Prince's Progress . . .* (1866), are more than fairy tales: the vendors of magic fruit in the one are symbols of temptation; the "progress" in the other represents the pilgrimage and trials of the soul.

Dante Gabriel's later years. From 1860 onward, such trials were part of Dante Gabriel's much disturbed life. His marriage to Elizabeth Siddal, clouded by her constant ill health, ended tragically in 1862 with her death from an overdose of laudanum. Grief led him to bury with her the only complete manuscript of his poems. That he considered his love for his wife similar to Dante's mystical and idealized love for Beatrice is evident from the symbolic "Beata Beatrix," painted in 1863.

Dante Gabriel's life and art were now greatly changed. He moved from riverside premises in London's Blackfriars to Chelsea. The influence of new friends—Swinburne and the American painter James McNeill Whistler—led to a more aesthetic and sensuous approach to art. Literary themes gave way to pictures of mundane beauties, such as his model and mistress of long standing, Fanny Cornforth, gorgeously apparelled and painted with a command of oils he had not previously shown. These paintings were popular, and Rossetti grew affluent enough to employ studio assistants to make copies and replicas. He collected antiques and filled his large, ill-kept Chelsea garden with a menagerie of animals and birds. He had enjoyed a modest success in 1861 with his published translations, *The Early Italian Poets;* and, toward the end of the 1860s, when his weakening eyes inclined him to give up painting for a while, his thoughts turned to poetry again. At that time, he and Christina were frequent visitors to Penkill Castle in Scotland, home of their friend Alice Boyd. There Dante Gabriel composed new poems and planned the recovery of the manuscript poems buried with his wife in Highgate Cemetery. Carried out in 1869 through the agency of his unconventional man of business, Charles Augustus Howell, the exhumation visibly distressed the superstitious Dante Gabriel. Publication followed in 1870. The *Poems* were well enough received until a misdirected, savage onslaught by "Thomas Maitland" (pseudonym of the journalist-critic Robert Buchanan) on "The Fleshly School of Poetry" singled out Rossetti for attack. Rossetti responded temperately in "The Stealthy School of Criticism," published in the *Athenaeum;* but the attack, combined with remorse and the amount of chloral and alcohol he now took for insomnia, brought about his collapse in 1872. He recovered sufficiently to paint and write, though the last ten years of his life were passed in semi-invalid seclusion. Until 1874 he spent much time at Kelmscott Manor (near Oxford), of which he took joint tenancy with William Morris in 1871. His lovingly idealized portraits of Jane Morris at this period were a return to his more poetic and mystical style. But his links with Morris were severed when the Pre-Raphaelite shareholders in his firm were bought out in 1875, the firm becoming simply Morris & Co. Dante Gabriel's life in Chelsea was subsequently that of a recluse, varied only by visits to his mother and sister and the company of his old friend Ford Madox Brown, a number of late disciples, and Theodore Watts-Dunton, a lawyer and man of letters who did his best to put Rossetti's financial affairs in order. He occupied himself with a replica of an early watercolour, "Dante's Dream" (1880), a revised edition of *Poems* (1881), and *Ballads and Sonnets,* containing the completed sonnet sequence of "The House of Life," in which he described the love between man and woman with tragic intensity. From a visit to Keswick (in northwest England) in 1881, he returned in worse health than before. He died on April 9, 1882, at Birchington-on-Sea (in southeast England), tended at the last by members of the family and a few loyal intimates.

John Ruskin described Dante Gabriel Rossetti as "a great Italian lost in the Inferno of London," though in fact he was never happy out of London and never visited Italy. With his appreciation of Italian poetry went a mastery of the various forms and rhythms of English as a poetic medium, ranging from the ballad style to the philosophical reflections of *The Burden of Nineveh* (1856). Quite different in tone is the easy and genial style of his intimate correspondence, which ranks him among the great letter writers. As a painter he claims respect, in present-day perspective, as the initiator of a form of symbolism influential in continental Europe as well as England and as a catalyst in the development of the concept of design.

MAJOR WORKS

Christina Rossetti

POETRY: *Goblin Market and Other Poems* (1862); *The Prince's Progress and Other Poems* (1866); *Sing-Song: A Nursery Rhyme Book* (1872, enlarged 1893); *A Pageant and Other Poems* (1881); *Verses* (1893); *New Poems* (1896); *Poetical Works* (1904).

OTHER WORKS: *Commonplace and Other Short Stories* (1870); *Annus Domini* (1874), a book of prayers; *Seek and Find* (1879), religious meditations; *Time Flies* (1885), a reading diary of mixed verse and prose; *The Face of the Deep* (1892). a commentary on the Apocalypse.

~anteGabriel Rossetti

POETRY: "The Blessed Damozel," with six sonnets and four lyrics in *The Germ,* the Pre-Raphaelite periodical (1850); *The Early Italian Poets from Ciullo d'Alcamo to' Dante Alighieri (1100–1200–1300) in the Original Metres Together with Dante's Vita Nuova* (1861); *Poems* (1870); *Ballads*

and Sonnets (1881), including "The House of Life," "The White Ship," and "The King's Tragedy."

PAINTINGS: "The Girlhood of Mary" (1849; Tate Gallery, London); "Ecce Ancilla Domini" ("The Annunciation"; 1850; Tate Gallery, London); "Found" (1854; Samuel and Mary R. Bancroft Pre-Raphaelite Collection, The Wilmington Society of the Fine Arts, Wilmington, Delaware); "Paolo et Francesca" (1855; Tate Gallery, London); "The Tune of the Seven Towers" (1857; Tate Gallery, London); "The Wedding of St. George and Princess Sabra" (1857; Tate Gallery, London); "Arthur's Tomb" (1860; Tate Gallery, London); "Beata Beatrix" (*c.* 1863; Tate Gallery, London); "Monna Vanna" (1866; Tate Gallery, London); "Dante's Dream" (1871; Walker Art Gallery, Liverpool); "The Blessed Damozel" (1871–79; Lady Lever Art Gallery, Port Sunlight, England); "The Bower Meadow" (1872; City of Manchester Art Galleries); "Proserpine" (1874; Tate Gallery, London); "Astarte Syriaca" (1877; City of Manchester Art Galleries).

OTHER ART WORKS: "How They Met Themselves" (1851–60; Fitzwilliam Museum, Cambridge), drawing; "Portrait of Miss Siddal" (1855; Ashmolean Museum, Oxford); "Maids of Elfinmere" (1855; wood engraving, engraved by Dalziel); "The Palace of Art," sometimes called "St. Cecilia" (1857; engraved by Dalziel), wood engraving; decoration of the Oxford Union with William Morris and Sir Edward Burne-Jones (1857, now largely obliterated). The major collection of Rossetti's drawings is in the City Art Gallery, Birmingham.

Gabriele Rossetti

Two volume commentary on the ***Divina Commedia*** (unfinished, 1825 and 1826); ***Sullo spirito antipapale che produsse la Reforma . . .*** (1831); ***Disquisition on the Antipapal Spirit which Produced the Reformation: Its Secret Influence on the Literature of Europe in General, and of Italy in Particular,*** trans. by C. Ward, 2 vol., 1834); ***Iddio e l'uomo, salterio*** (1833), poems; ***Il mistero dell'amor platonico nel medio evo,*** 5 vol. (1840), a philosophico-literary treatise; ***La Beatrice di Dante*** (1842; complete ed., 1935), study of Dante; ***Il Veggente in solitudine poema polimetro*** (1846; ***A Versified Autobiography,*** trans. by W.M. Rossetti, 1901); ***Versi*** (1847); ***L'Arpa evangelica*** (1853), lyrics; ***La vita mia*** and ***Opere inedite e rare*** (1910; 3 vol., 1931, both ed. by D. Ciampoli).

William Michael Rossetti

Inferno (1865), blank verse translation; ***Fine Art, Chiefly Contemporary: Notices Reprinted with Revisions*** (1867), art criticism; ***Troylus and Cryseyde*** (1875), parallel text to Chaucer; ***Filostrato*** (1883), parallel text to Boccaccio; ***Ruskin: Rossetti: Preraphaelitism. Papers 1854–62*** (1899); ***Preraphaelite Letters and Diaries*** (1900); ***Dante and His Convito: A Study with Translations*** (1910).

BIBLIOGRAPHY. A comprehensive bibliographical reference to members of the family is contained in WILLIAM E. FREDEMAN, ***Pre-Raphaelitism: A Bibliocritical Study*** (1965). Of primary importance for the study of Rossetti's art is VIRGINIA SURTEES, ***The Paintings and Drawings of Dante Gabriel Rossetti: A Catalogue Raisonné,*** 2 vol. (1971). Another standard work is H.C. MARILLIER, ***Dante Gabriel Rossetti*** (1899). Kelmscott Manor, now restored by the Society of Antiquaries, recalls the joint tenancy of Rossetti and Morris in many details of furnishing and decoration. Major editions of literary works are: ***The Works of Dante Gabriel Rossetti,*** rev. ed. by WILLIAM MICHAEL ROSSETTI (1911); and ***The Poetical Works of Christina Rossetti,*** ed. with memoir and notes by W.M. ROSSETTI (1904, reprinted 1971). R.D. WALLER, ***The Rossetti Family (1932),*** is an account of Gabriele Rossetti and the family in his lifetime (1824–54). The nearest complete collection of D.G. Rossetti's correspondence published to date is the ***Letters of Dante Gabriel Rossetti,*** ed. by OSWALD DOUGHTY and J.R. WAHL, 4 vol. (1965–67). W.M. ROSSETTI edited and added a memoir to ***The Family Letters of Dante Gabriel Rossetti,*** 2 vol. (1895, reprinted 1970) and ***The Family Letters of Christina Georgina Rossetti*** (1908, reprinted 1968). Numerous selections of letters and reminiscences illuminate different phases of D.G. Rossetti's career; the letters and diaries arranged by W.M. ROSSETTI in ***Ruskin: Rossetti: Preraphaelitism. Papers 1854–62*** (1899, reprinted 1971), give a vivid commentary on Ruskin as patron, Madox Brown as friend, and Miss Siddal as an inspiration. H. TREFFRY DUNN, ***Recollections of Dante Gabriel Rossetti and His Circle*** (1904, reprinted 1971), describes the Cheyne Walk (Chelsea) period when Dunn was Rossetti's assistant. WILLIAM BELL SCOTT, ***Autobiographical Notes . . . , 2*** vol. (1892, reprinted 1970), gives somewhat unkind emphasis to Rossetti's disturbed condition at Penkill and subsequent collapse. THOMAS GORDON HAKE, ***Memoirs of Eighty Years*** (1892), includes reminiscences of Rossetti by his doctor friend. JANET CAMP TROXELL (ed.), ***Three Rossettis*** (1937), studies the brothers and sister through their letters and has much interesting detail of the stay at Penkill and the arrangements for exhuming Rossetti's manuscript poems. HALL CAINE, ***Recollections of Dante Gabriel Rossetti*** (1882), describes the reclusive last years. His letters, often grotesquely amusing, to Fanny Cornforth were edited by P.F. BAUM (1940); in the biographies of other Pre-Raphaelites, D.G. Rossetti inevitably figures. The unsympathetic view of him in EVELYN WAUGH, ***Rossetti: His Life and Works*** (1928), was followed by the overt hostility of VIOLET HUNT in ***The Wife of Rossetti*** (1932). Against this prejudice HELEN ROSSETTI ANGELI, daughter of W.M. Rossetti, came to her uncle's defense in ***Dante Gabriel Rossetti: His Friends and Enemies*** (1949). OSWALD DOUGHTY, ***A Victorian Romantic: Dante Gabriel Rossetti,*** 2nd ed. (1960), is an effort to give overall proportion. LONA MOSK PACKER, ***Christina Rossetti*** (1963), supplements the ***Life*** by MACKENZIE BELL of 1898 with fresh research and revised psychological analysis.

(W.Ga.)

Ross Ice Shelf

The world's largest tabular body of floating ice, the Ross Ice Shelf lies at the head of Ross Sea, an enormous indentation in the continent of Antarctica far beyond the waters of the South Pacific. The ice shelf lies between about 155° W and 160" E longitude; and about 78" S and 86" S latitude. Estimates of its area range from about 192,000 to 208,000 square miles (496,000 to 540,000 square kilometres), making it approximately the size of France. The shelf has had considerable historic importance, having served as a gateway for explorations of the Antarctic interior, including those carried out by many of the most famous expeditions. The behaviour of such a vast sheet of ice is of exceptional interest to scientists for the light it sheds on such matters as the formation of icebergs, and the glacial budget of Antarctica, and also because its movements are analogous to those of huge geological formations. For those familiar with the region, the whole vast area has an austere but magnificent beauty. (For related information, see ROSS SEA; ANTARCTICA; ICEBERGS AND PACK ICE.)

Exploration. The great white barrier wall of the shelf's front, first seen in 1841 by the British rear admiral and polar explorer James Clark Ross, rises in places to 160 or 200 feet high and stretches about 500 miles between fixed "anchor points" on Ross Island to the west and the jutting Edward VII Peninsula on the east. With its immense, gently undulating surface reaching back nearly 600 miles southward into the heart of Antarctica, the ice shelf provides the best surface approach into the continental interior. The McMurdo Sound region on the west thus became headquarters for British expeditions in the early 1900s, including R.F. Scott's 1911–1912 epic sledging trip to the South Pole. In the early 1960s, it was the site of the headquarters of the United States Antarctic Research Program (USARP), the United States Navy's Operation Deep Freeze, and the New Zealand Antarctic Research Program. Eastern barrier regions were headquarters for the Norwegian Roald Amundsen's first attainment of the South Pole on December 14, 1911; for Richard E. Byrd's three U.S. expeditions of 1928–41 at Little America I–III stations; for the U.S. naval expedition Operation High Jump at Little America IV in 1946–47; for U.S. International Geophysical Year expeditions at Little America V in 1955–57; and for summer research groups in the 1960s.

The shelf as a base for expeditions

Physiography and glaciology. The Ross Ice Shelf lies between spectacular, clifflike escarpments of the central Transantarctic Mountains on the west and the grounded West Antarctic ice sheet on the east. Ranges more than 13,000 feet in elevation, including the Royal Society and Queen Alexandra ranges and the Queen Maud Mountains, form an enormous bedrock "dam" holding back the high polar-ice sheet of East Antarctica. The dam is breached at a number of places so that polar ice locally escapes in spectacular giant ice streams, principally the Byrd, Beardmore, Mulock, Nimrod, Scott, Amundsen, Ready, and Liv glaciers, to feed the ice shelf below. Byrd Glacier, the largest, flows at 7.5 feet per day, making it the fastest known Antarctic valley glacier.

The Ross Ice Shelf has been likened to a vast triangular raft, relatively thin and flexible and only loosely attached to adjoining lands. It appears from the air as an incredibly vast flat plain, disturbed here and there by intensely crevassed areas (marking local grounded-ice conditions, or zones of differential shear where physical forces have torn the ice apart). Pioneering studies on ice-sheet seismology, made on Byrd's second Antarctic expedition (1933–35), demonstrated that a large northeastern section of the shelf (Roosevelt Island) is actually grounded. Gentle tidal changes continually break the shelf ice from moorings to grounded ice, thus forming around the inland margin of the shelf a system of filament-like "strand" cracks that are continually rehealed by freezing of intruded seawater. Giant rifts develop behind the barrier and occasionally rupture completely to spawn the huge tabular icebergs that are so characteristic of the Antarctic Ocean. (Little America I–IV stations, in Bay of Whales calved in this manner and disappeared by 1954.)

Movement and deformation of shelf ice

The shelf ice, confined between walls, moves seaward by spreading as a result of both surface accumulation and an input of land ice. The ice (which may, in fact, be technically defined as an easily deformable crystalline rock) buckles where the shelf impinges against the bedrock buttresses of Roosevelt and Ross islands. Under compression, the ice is deformed into the great upwarped and downwarped structures known, respectively, as anticlines and synclines, as well as a great variety of faults and joints that duplicate those seen in more familiar crustal rocks.

Although the barrier's position appears almost stationary, it actually undergoes continual change by calving and melting that accompany northward movement of the shelf ice. The net mass budget of Ross Ice Shelf (a technical expression that reflects the net surface ice accumulation plus the inflow of land ice plus the increments by bottom freezing; less, on the other hand, the loss by calving, surface melting, and bottom melting) seems to be in approximate equilibrium, but some factors, particularly bottom melting or freezing and calving rates, are of highly uncertain quantity. The mean ice thickness is about 1,100 feet along a line at about 79" S latitude across the ice shelf. (In a southward direction along about 168" W longitude, the thickness gradually increases to more than 2,300 feet.) Ice velocity reaches a maximum of more than 3,000 feet per year in central sectors of this line and averages about 2,180 feet per year. The total ice discharge across the line is estimated to be 45 ± 5 cubic miles per year. The net surface accumulation on the ice shelf and on that part of the West Antarctic ice sheet draining into it is believed to be about $46 \pm 10 \times 10^{13}$ pounds per year. Added to this is the input to the shelf of an estimated $11 \pm 3 \times 10^{13}$ pounds per year by valley glaciers from the polar plateau to the west. Although at the leading edges of the ice sheet bottom melting may be high, perhaps as much as 30 inches per year near Little America Station, at a distance of from 30 to 60 miles to the south, the bottom regimen probably changes from one of melting to one of freezing. Recent estimates suggest that at distances of from 100 to 200 miles from the barrier 15 to 20 inches of ice may be added each year by bottom freezing. An estimated net of 4×10^{13} pounds per year is added to the whole shelf by bottom freezing. Estimates such as these suggest that the Ross Ice Shelf drainage system has a positive net budget. The net budget of Ross Ice Shelf itself depends additionally on the calving rate which is presently unknown.

Ice growth rates

Geology. The terrain of the adjoining Transantarctic Mountains consists of Precambrian to Early Paleozoic metamorphic, sedimentary, and granitic rocks dating from 500,000,000 years or more ago. These are unconformably (*i.e.*, erosionally) overlain by flat-lying Devonian to Jurassic strata laid down from 350,000,000 to about 190,000,000 years ago and known as the Beacon Group. Geophysical studies suggest that the Beacon rocks, or other sedimentary rocks perhaps of a Mesozoic or Tertiary basin (from 225,000,000 to perhaps a few million years ago), underlie Ross Ice Shelf.

At the glacial maximum, the ice shelf probably became grounded and spread as an ice sheet far northward, reaching possibly to Pennell Bank (74" S), which is believed to be a terminal moraine, or ridge of glacier-deposited material. At that time, it probably overrode islands (such as White and Black islands) and reached far up on the slopes of Mount Discovery and Ross Island. Morainal evidence shows that four times within the last 1,200,000 years (in ice ages known as Ross Sea Glaciation I–IV stages), ice pushed westward up the McMurdo dry valleys from the direction of the Ross Sea. With the melting of the ice sheet of Ross Sea Glaciation IV, an ice shelf again formed in Ross Sea.

Future scientific work, it appears, will focus on gaining a third-dimensional view of the Ross Ice Shelf environment, by geophysical measurements and especially by drilling, to study not only this glacial sequence as recorded by sediments on the sea floor but also the deeper bedrock structure and the unique biological environment beneath an ice shelf about which little is now known.

BIBLIOGRAPHY. M. MELLOR (ed.), *Antarctic Snow and Ice Studies* (1964), contains several papers on the Ross Ice Shelf; M.B. GIOVINETTO and J.H. ZUMBERGE, "The Ice Regime of the Eastern Part of the Ross Ice Shelf Drainage System," *Int. Ass. Scient. Hydrol.* no. 79 (1968).

(A.B.Fo.)

Rossini, Gioacchino

The leading composer of Italian opera during the early part of the 19th century, Gioacchino Antonio Rossini virtually embodied the history of opera in Italy, and then in France, for nearly half a century. His comic opera *Il* barbiere di Siviglia (*The* Barber of Seville) and **his** serious opera *Guillaume* Tell (William Tell) remain prime examples of their respective genres.

George Eastman House Collection

Rossini, photograph by Étienne Carjat, *c.* 1868.

Every salient fact in the life of Rossini was a theatrical event—to be born on February 29 and to die on Friday the 13th seems to bespeak a sense of humour and a sense of theatre. His birth, his family background, his childhood, his personal life, his career, his retirement, all had connection with the theatre. Admired throughout Europe during his lifetime, he was cast into the shadows only for a moment by the popularity of the opera composer Giuseppe Verdi.

Born in Pesaro, Italy, on February 29, 1792, Gioacchino Antonio Rossini was the son of Giuseppe Rossini, an impoverished trumpeter who played in miscellaneous bands and orchestras, and Anna Guidarini, a singer of secondary roles. Thus Rossini spent his entire childhood in the theatre. A lazy student, the young Gioacchino found it easy to learn to sing and play. By the age of 15, he had learned the violin, horn, and harpsichord and had often sung in public, even in the theatre, to earn some money. At 14 he entered Bologna's Liceo Filarmonico (now the Conservatorio Statale di Musica G.B. Martini) and composed his first opera—*Demetrio* e *Polibio* (staged in 1812)—for the Mombelli, a family of singers.

When his voice broke and he was unable to continue singing, Rossini became an accompanist, then a conductor. He had already realized the importance of the German school, perceiving the new elements by which Haydn and Mozart had enriched music. These influences can be found in the early cantata he wrote for the Liceo Filarmonico, performed there in 1808. During the next 20 years (from 1808) this genial lazybones was to compose more than 40 operas.

The Italian period. By taste, soon by obligation, Rossini threw himself into the genre then fashionable: opera buffa (comic opera). His first opera, La *cambiale* di *ma*trimonio (The Bill of Marriage, 1810), was performed in Venice and had a certain success, although his unusual orchestration made the singers indignant. Back in Bologna again, he gave the cantata La morte di *Didone* (Dido's Death, 1811) in homage to the Mombelli family, who had helped him so much, and scored a triumph with the two-act opera buffa L'equivoca stravagante (*The* Extravagant Misunderstanding, 1811). The following year, two more of his comic operas were produced in Venice.

Rossini had already broken the traditional form of opera buffa: he embellished his melodies (he was the true creator of bel canto, a florid style of singing), animated his ensembles and finales, used unusual rhythms, restored to the orchestra its rightful place, and put the singer at the service of the music. In 1812 Rossini wrote the oratorio Ciro in Babilonia (Cyrus in Babylon) and La scala di *seta* (The Silken Ladder), another comic opera.

Operas for La Scala

The same year, Marietta Marcolini, who had already sung in Rossini's operas and who was interested in the young composer, recommended Rossini to the committee of La Scala opera house, Milan. It was under contract to them that he wrote La pietra del *paragone* (The *Touch*stone, 1812), a touchstone of his budding genius. In its finale, Rossini—for the first time—used the crescendo effect he was later to use and abuse indiscriminately.

By this time Rossini's experience in writing seven operas and several cantatas and his intimate contact with the theatre had given him a profound knowledge of his profession. Singers no longer held terrors for him. He was now ready for his major works. Venice, the most refined city in Italy, was to offer him his true glory. After the comic opera *Il* signor Bruschino (1813), written for the Teatro San Moisè, he next wrote—for La Fenice—his first serious opera, Tancredi (1813), in which he tried to reform opera seria (the formula-ridden, serious operas of the 18th century), and he composed an authentically dramatic score. This work, spirited and melodious, was an instant success. Tancredi's famous song, "Di tanti palpiti," was whistled all over town. La Malanotte, the famous singer, was also bent on proving to him the quality of her natural charms, and rumours began to abound. The success of L'Italiana in Algeri (The Italian Girl in Algiers, 1813) followed, showing further refinements in his reforms of opera buffa. These two successes opened wide the doors of La Scala. With Aureliano in Palmira (1814) the composer affirmed his authority over the singers; he decided to prescribe and write the ornaments for his arias, but the work was not a success. After *L'Italiana* he wrote *Il* Turco in *Italia* (The Turk in Italy, 1814) for the Milanese and a cantata for Princess Belgioioso, "one of the most likeable of protectresses," as the French novelist Stendhal referred to her. His next work, *Sigismundo* (1814), was a failure.

Neapolitan operas

Rossini's fame soon spread to Naples, where the reigning impresario was Domenico Barbaia, an ambitious former coffeehouse waiter, who by gambling and running a gaming house had amassed a fortune and was now in charge of the two great Neapolitan theatres. Barbaia realized Rossini's growing fame and went to Bologna to offer him a contract. Impressed by the terms of this contract—security, two operas a year—as well as by Barbaia, a millionaire rather than the customary fourth-rate impresario on the verge of bankruptcy, Rossini did not hesitate to accept. How could anyone refuse a tempting impresario whose favourite was none other than the imposing diva Isabella Colbran? Her first Rossini opera, *Elisabetta, regina d'Inghilterra* (Elizabeth, Queen of England, 1815), was a triumphant success. Rossini admired Colbran very much and soon fell in love with her. The brilliant success of Elisabetta prompted an invitation from Rome to spend the carnival season of 1816. The first of Rossini's Rome operas was unsuccessful. So was the second, *Almaviva*, soon to become *Il* barbiere di *Siviglia*. The Romans, who knew and lsved Giovanni Paisiello's version of Beaumarchais's play, took a dislike to this new setting, but when it was given elsewhere in Italy it was received with unbounded success. Written in 16 days, the work is a piece of inspired inventiveness that has delighted opera lovers ever since. There followed La *cenerentola* (Cinderella, 1817). As with The Barber, this work uses a coloratura contralto for the heroine's role; it, too, proved no less successful. In between these two comedies came *Otello* (Othello; Rome, 1816), a setting of Shakespeare's play that held the stage until superseded by Verdi's greater opera of the same name. La gazza ladra (*The* Thieving Magpie, 1817), a semi-serious work, was a triumph in Milan.

Armida, a grand opera requiring a trio of tenors and a dramatic soprano (Colbran) appeared in 1817. Rossini was now finding interpreters that suited his music. Colbran, the tenor Manuel Garcia, the bass Filippo Galli ("the most beautiful voice in Italy"), the contralto Benedetta Pisaroni (whose art had no equal in depth) were his usual exponents and carried forward his art of bel canto.

La donna del *lago* (based on Sir Walter Scott's poem "The Lady of the Lake") failed at its premiere in 1819 but soon came into favour. After several more or less successful works, he left Naples for Vienna, along with Colbran (whom he had just married), anxious to meet Beethoven. Disappointed by the economic situation of the composer of Fidelio, he returned to Venice, where he attempted to crown his Italian career with Semiramide (1823). The old-fashioned Venetians, however, did not understand the astonishing work, his longest and most ambitious, and so he resolved not to write another note for his countrymen. Following his resolution, he decided to leave Italy.

The Parisian period. Rich, married, unstable, and by nature an epicurean, Rossini wanted to travel. He arrived in Paris in November 1823 and was enthusiastically welcomed in the French capital. The Academy in Paris received him; all of the town fawned upon him. At the end of the year, he visited London, where he conducted and sang in concerts with his wife and met King George IV. Back in Paris, he embarassed the old musicians. "Rossini," wrote the Escudier brothers, Paris music publishers,

> was then 31 years old and in his prime. His countenance revealed a lofty and congenial expression. His subtle, quick penetrating eye held one magnetically before him. His smile, benevolent and caustic at the same time, reflected his whole disposition. The clear line of his aquiline nose, his vast and prominent brow, which his prematurely receding hairline entirely revealed, the even oval of his face enclosed in jet-black sideburns, all formed a kind of virile and fascinating beauty. He has a marvelously shaped hand, which he displayed somewhat coquettishly through his cuff. He dressed in a simple manner, and, under his clothes, which were more proper than elegant, the appeerance of a newly disembarked provincial into the capital.

If the old nicknamed him "Monsieur Crescendo," the young very quickly paraded their admiration for him. Paris was then the centre of the world and Rossini knew it. After some of his works had been staged, he composed *Il* viaggio a Reims (The Journey to Reims), a cantata improvised for the coronation of Charles X.

For a long while Rossini hoped to modify his style: to replace the comparative artificiality and coldness of florid opera coloratura with declamatory and lofty singing—that is, with truth and intensity. In order to do that, he also had to reform the orchestra and give more importance to the chorus. Thus appeared Le *Siège* de *Corinthe* (The Siege of Corinth, 1826), a revision of the earlier Maometto *II* (1820), which was saluted by the prominent composer Hector Berlioz. Le *Siège* was followed by *Moïse* (Moses, 1827) and Le *Comte* Ory (Count Ory, 1828), an adaptation of opera buffa style to French opera.

Final work for the theatre

His final opera, ***Guillaume Tell*** (1829), is on the noble themes of nationalism and liberty, and Rossini's music is worthy of the elevated subject. The Parisian public gave him an ovation, and, in a single work, he had responded to all the critics in the most elegant manner. Then he decided, at the age of 37, not to write again for the theatre. ***Tell*** was to have been the first of five operas for the Optra, but the new government following the Revolution of 1830 set aside his contract.

The reasons for his musical silence remain only suppositions. Some cite his legendary laziness as the cause, while others point to the Parisian hostility to his work and Rossini's resulting sulkiness. Another cause might have been his jealousy over the Parisian success of the opera composer Giacomo Meyerbeer.

In 1845, Madame Colbran passed away. In 1847 Rossini married Olympe Pélissier. During his retirement he had written, returning to his first loves, some religious pieces: the *Stabat Mater* (1832), *Petite Messe solennelle* (1864), and a few songs and piano pieces the good taste of which endeavoured to establish their reputation. Rossini never agreed to their publication.

After a period in Italy, he returned to Paris in 1855, never again to leave it. His parents being deceased, his new wife less demanding than the preceding one, and he himself a wealthy man whose retirement was assured, Rossini gave way to the sweetness of life and to being a wise man who permitted himself to shine in society with a few clever expressions and witticisms. His bons mots, in fact, are legendary, as were his caustic wit and low humour. At his Paris home and later at his villa in Passy, Rossini gave superb gourmet dinners attended by many of the greats of the musical and literary world of the mid-19th century. In 1860 the renowned German composer Richard Wagner visited him, and their fascinating conversation was recorded by Wagner in his essay "Eine Erinnerung an Rossini." Rossini died at Passy of pneumonia on November 13,1868.

For years Rossini was virtually known only by the omnipresent ***Barber of Seville*** and an occasional revival of ***William Tell.*** From the 1950s more and more of his operas were revived, particularly at festivals, and nearly always with public and critical acclaim.

MAJOR WORKS

OPERAS: Some 35 including ***La pietra del paragone (The Touchstone,*** first performed ***1812)***; ***Tancredi (1813)***; ***L'Italiana in Algeri (The Italian Girl in Algiers, 1813); Il Turco in Italia (The Turk in Italy, 1814); Elisabetta, regina d'Inghliterra (Elizabeth, Queen of England, 1815); Il barbiere di Siviglia (The Barber of Seville, 1816); La cenerentola (Cinderella, 1817); Armida (1817); La gazza ladra (The Thieving Magpie, 1817); La donna del lago (The Lady of the Lake, 1819); Semiramide (1823); Le Siège de Corinthe (The Siege of Corinth, 1826); Moïse (Moses, 1827); Guillaume Tell (William Tell, 1829).***

CHORAL MUSIC: ***Il viaggio a Reims*** (cantata with ballets, completed ***1825); Stabat Mater (1832,*** revised ***1842); Petite messe solennelle (1853).***

CHAMBER MUSIC: Five string quartets ***(1808).***

SONGS: Various, among them ***Les Soirées musicales*** (published ***1835).***

PIANO: ***Péchés de vieillesse (Sins of Old Age),*** about ***180*** pieces for piano or for various instruments, voice, and piano.

BIBLIOGRAPHY. There are very few works devoted to Rossini. All of them are in fact inspired by A. AZEVEDO, G. ***Rossini, sa vie et ses oeuvres*** (1864), an extremely well-documented work written by an unconditional admirer; and STENDHAL, ***Vie de Rossini (1824;*** Eng. trans. ***1956,*** new ed. 1970), a study of the man and his early works by a contemporary. In English, HERBERT WEINSTOCK, ***Rossini*** (1968), is a full and authoritative account of Rossini's life.

(J.-L.Ca.)

Ross Sea

The Ross Sea forms a southern extension of the Pacific Ocean, which, along with the vast ice shelf (see ROSS ICE SHELF) at its head, makes a deep indentation in the circular continental outline of Antarctica. The sea is a generally shallow marine region, approximately 370,000 square miles (960,000 square kilometres) in area, centred at about 75° S, 175" W, and lying between Cape Adare in northern Victoria Land on the west and Cape Colbeck on Edward VII Peninsula on the east. The northern limit lies approximately along the edge of the continental shelf and the southern limit along a great barrier wall of ice marking the front of the Ross Ice Shelf. (For related information see Ross ICE SHELF; ANTARCTICA; ICEBERGS AND PACK ICE.)

Focal point for exploration

Exploration. The Ross Sea is one of the least iced and most accessible of Antarctica's fringe seas. Relatively easy access made this region the traditional avenue for mounting expeditions into the continental interior, from the 1901–1904 British National Antarctic Expedition under R.F. Scott to the U.S. Operation Deep Freeze of the 1960s and early 1970s. The sea was first penetrated on January 5, 1841, by the HMS "Erebus" and HMS "Terror," commanded by James Clark Ross on an unsuccessful attempt to reach the south magnetic pole. It then remained unprobed until January 24, 1895, when a Norwegian expedition on the ship "Antarctic," searching for new whaling grounds, put a party ashore at Cape Adare for the first landing in Victoria Land.

Activity in the area increased after winter survival was proved possible by the Norwegian Carsten Borchgrevink and his party of Laplanders at Cape Adare in the winter of 1899–1900. Ross Island and McMurdo Sound in the southwest comer of the sea became the focal point for inland exploration by early 20th-century British parties, led by Scott and E.H. Shackleton, in their searches for the geographic and magnetic poles; and the Bay of Whales, in the southeastern comer of the sea, became the base for the south polar quest of Roald Amundsen's Norwegian Antarctic Expedition (1910–1912). The Japanese South Polar Expedition, on the "Kainan Maru," revisited the Bay of Whales in 1911–1912. It was used again by Richard E. Byrd's United States Antarctic Expedition of 1928–1930 in establishment of "Little America I" station and several times after. Since the International Geophysical Year (1957–59), Ross Sea activities have shifted to the McMurdo Sound region. All coastal regions have been geologically explored, at least in reconnaissance, largely by American and New Zealand parties. The southern Victoria Land region near McMurdo Sound has become the most thoroughly known on the continent. Rich with history and scenery, the Ross Sea is now regularly traversed by tourist vessels.

Geology and physiography. Overshadowed by the towering ranges of Victoria Land, the floor of the Ross Sea extends northward as a broad shelf before plunging into the deeps of the Southeast Pacific Basin along a line from Scott Island to Cape Colbeck. The broader, western half of the sea shoals to less than 1,000 feet in several wide areas, the southwesternmost culminating in the small and rocky volcanic pile of Franklin Island. Most of the floor is less than 3,000 feet deep.

Coastal terranes consist, in large part, of strongly folded and variably metamorphosed sandstone and slate, probably dating from the Late Precambrian (over 570,000,000 years old). Fold trends in northern Victoria Land (Robertson Bay Group) are generally oriented southeasterly, truncated at the Ross Sea coast, and project toward similar terranes with similar fold trends in Marie Byrd Land. The flat-lying Beacon Group of Victoria Land has no known counterpart across the Ross Sea. The coastal region is dotted with modern volcanos and older dissected volcanic piles of an extensive alkaline–basalt area (McMurdo Volcanics) consisting of Cape Adare, Cape Hallett, Mount Melbourne, Franklin and Ross islands, on the western coast, and a number of lesser known centres in western Marie Byrd Land, on the eastern coast.

Possible origins of the continental embayment

Several possibilities exist with regard to the origin of the Ross Sea continental embayment. Early geologists believed the Ross Sea to extend as a subglacial trench, the hypothetical Ross–Weddell Graben, to connect with the Weddell Sea. Since then, discovery of the intervening Ellsworth Mountains has disproved the hypothesis. Geophysical studies on the Ross Ice Shelf indicate the presence of an underlying thick section of seismically low-velocity, probably sedimentary, rocks. The embayment, therefore, may be either a down-faulted block of continental rocks, including the Beacon Group, or it may be a downwarped basin filled with sedimentary rocks.

Various types of glacial-marine sediment carpet the floor of the sea and suggest that much of it was formerly occupied by grounded ice. The many irregularities are probably glacial moraines, and it has been postulated that the vast Pennell Bank is a terminal moraine from the maximum ice advance. Paleomagnetic dating of sediment cores shows that the first influx of glacial sediments occurred at least 5,000,000 years ago. Micropaleontological studies support evidence from the Antarctic continent itself that glacial onset was much earlier, perhaps several tens of millions of years ago. Probes in 1972–73 by the deep-sea drilling vessel "Glomar Challenger" into the marine sediment record of the Ross Sea were expected to provide some answers to such diverse questions as the age of uplift of the nearby Transantarctic Mountains; the ages of onset of glaciation and of fluctuations in the great Antarctic Ice Sheet; and the nature of sedimentation and of organic evolution in polar marine environments.

Biology and oceanography. Flora and fauna is typical of other southerly antarctic marine regions. The nutrient-rich sea water supports abundant planktonic life which in turn provides food for larger forms, including fish, seals, whales, and sea- and shore-birds. Among the latter, hordes of Adélie and emperor penguins populate rookeries at a number of places around Ross Sea.

The Ross Sea is strongly influenced by the coastal East-Wind Drift that sets up a vast, clockwise gyre accompanied by deep-water upwelling. Surface currents move generally westward along the ice-shelf front and thence northward along Victoria Land, where they meet the West-Wind Drift. Movements are complicated by shoals and tidal currents. Water less than 1,000 feet deep has a minimum temperature of about 28.2" F (—2.1" C). Mean surface water temperature in McMurdo Sound is about 28.8" F (—1.8" C). Salinities vary directly with sea-ice formation, decreasing to values of 34.37 ‰ (parts per thousand) with melting in summertime and increasing to 34.84 ‰ when maximum ice formation occurs in McMurdo Sound. (A.B.Fo.)

Rothschild Family

The most famous of all European banking families, the Rothschilds, whose very name was long synonymous with prodigious wealth, exerted for almost two hundred years a considerable influence on the economic history of Europe. Their continuing success, though far less marked than during the 19th century, when they were the world's most powerful bankers, is still based on intimate cooperation between the remaining branches of the family and their ability to engage in large-scale finance.

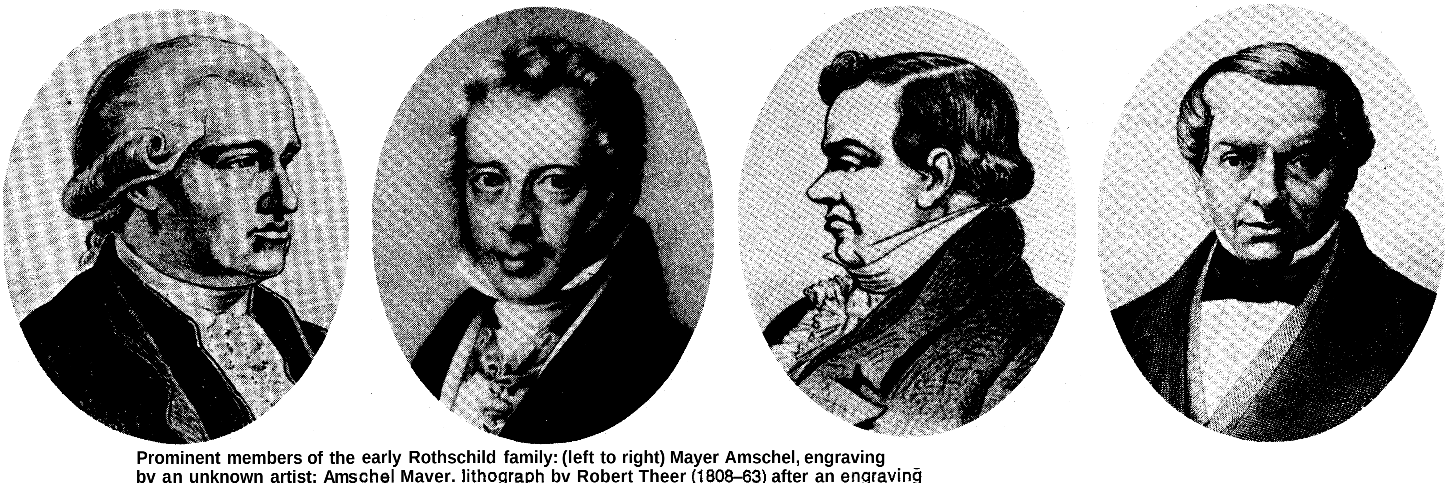
The founder of the house was Mayer Amschel, born in the Jewish ghetto of Frankfurt am Main on February 23, 1744. The family's name derived from the red (rot) shield on the house in the ghetto in which Mayer's ancestors had once lived. Intended for the rabbinate, Mayer studied briefly, but his parents' early death forced him into an apprenticeship in a banking house. Soon after becoming court factor to William IX, landgrave of Hesse-Kassel, Mayer set the pattern that his family was to follow so successfully—by preference to do business with reigning houses and to father as many sons as possible who could take care of the family's many business affairs abroad.

Starting as dealers in luxury items and traders in coins and commercial papers, Mayer and his sons eventually became bankers to whom the French Revolutionary and Napoleonic Wars of 1792–1815 came as a piece of great good fortune. Mayer and his oldest son, Amschel Mayer, supervised the growing business from Frankfurt, while Nathan Mayer established a branch in London in 1804, James (or Jacob) settled in Paris in 1811, and Salomon Mayer and Karl Mayer opened offices in Vienna and Naples, respectively, in the 1820s. The wars, for the Rothschilds, meant loans to warring princes; smuggling as well as legal trading in key products such as wheat, cotton, colonial produce, and arms; and the transfer of international payments between the British Isles and the Continent that Napoleon vainly attempted to close to British trade. Peace transformed the growing Rothschild business: the banking group continued its international business dealings but became more and more an agent in government securities (Prussian or English, French or Neapolitan), in insurance-company stocks, and in shares of industrial companies. Thus, the family successfully adapted to the Industrial Revolution and participated in economic growth throughout Europe with their railway, coal, ironworking, and metallurgical investments. The banking group continued to expand after the 1850s and, in particular, achieved an important position in the world trade of oil and nonferrous metals. But its previous "oligopolistic" position was seriously threatened by new joint-stock banks and "commercial," or deposit banks both in England and in France, as well as in the German states. By the last quarter of the 19th century, the Rothschild group was no longer the first banking consortium. Other groups, in Europe and in the United States, had become stronger, richer, more enterprising.

The five brothers

Yet the two guidelines laid down by Mayer Amschel for the Rothschild business operations (which, indeed, became a family tradition)—to conduct all transactions jointly and never to aim for excessive profits—helped to compensate to a notable extent for the inevitable risks inherent in handing down a business to future generations not all of whose members are qualified to run it. Amschel Mayer (died 1855), Nathan (died 1836), James (died 1868), Salomon (died 1855), and Karl (died in Naples in 1855)—the founders of the Rothschild consortium—were themselves unequally endowed: Nathan and James

BY courtesy of (left, right) the Historisches Museum, Frankfurt am Main; (centre left, centre right) Osterreichische Nationalbibliothek. Vienna



Prominent members of the early Rothschild family: (left to right) Mayer Amschel, engraving by an unknown artist; Amschel Mayer, lithograph by Robert Theer (1808–63) after an engraving by Franz Lieder (1780–1859); Nathan Mayer, detail of a lithograph by Karl von Stur, 1886; James, engraving by an unknown artist.

stood out among their brothers by the force of their personalities, particularly Nathan, who was hard, deliberately boorish, and sarcastic. James, who was his brother's equal in all these things, possessed an alleviating air of some refinement as a result of living in the more polished atmosphere of Paris. The five founders in turn had unequal successors. For example, if Alphonse in Paris (died in 1905) was a worthy successor to his father James, his own son, Edouard (died in 1949), was not as strong a figure as his position required. But Édouard's son (Guy, born in 1909) and his two cousins (Alan, 1910, and Elijah, 1917) have shown exceptional adaptability and ambition, thus confirming the constant element in the group's history for a ceniury and a half: a remarkable capacity for seizing opportunities and for adapting as much in business as in politics.

The second generation

In separating the circumstantial from the personal, individual aspects of the dynasty's hegemony during the 19th century, it must be noted that, although the first group of Rothschilds arrived as strangers in their new countries, unfamiliar with the languages and the customs and subject to the jealousy and competition of local bankers, they stood out from those around them by their fierce will to acquire a place in the sun. By the second generation, when the sons of the five founding brothers (notable among them Anthony and Lionel Nathan in London, Alphonse and Gustave in Paris) entered the business, the Rothschilds were polished and refined, as well as naturalized and nationalized to the point of blending into leadership positions without losing any of their family attributes. It ıs possible that the young Rothschilds' education and the extremely worldly existence of the heads of the various houses helped to create this true mutation. On the other hand, the Rothschilds were influencing the national economy and politics of their countries as greatly as they were being influenced themselves. Alphonse, for example, as the head of the international banking syndicate that in 1871 and 1872 placed the two great French loans known as liberation loans after France's defeat by Prussia, could boast without immodesty that his influence had maintained the chief of the French government, Adolphe Thiers, in power. At the same time, on the other side of the Channel, in the autumn of 1875, Lionel, in London (where he had been a member of the House of Commons since 1858), was able to give Prime Minister Disraeli on a few hours notice the £4,000,000 that allowed the British government to become the principal stockholder in the Suez Canal Company. Obviously, the two cousins had become important citizens in their respective countries.

There were frequent marriages between Rothschild cousins. and marriages generally were—with very rare exceptions––with Jews. As a result, in spite of the number of their descendants and the complexity of their family tree, the Rothschilds, particularly those of Vienna and Paris during the Nazi period, preserved the kind of family unity necessary to weather great misfortunes.

The Rothschilds were much honoured. Mayer's five sons were made barons of the Austrian Empire, a Rothschild was the first Jew to enter the British Parliament, and another was the first to be elevated to the British peerage. The head of the British branch of the family has always been considered the unofficial head of British Jewry. Members of the British and French families—the only ones still engaged in banking after the seizure by the Nazis of the Austrian house—have distinguished themselves as scientists and often as philanthropists.

BIBLIOGRAPHY. JULES AYER, *A Century of Finance: 1804 to 1904* (1905), a brief, official work; CHRISTIAN W. BERGHOEFFER, *Meyeɪ Amschel Rothschild der Gründer des Rothschildschen Bankhauses* (1922), a very good study on the father of the "five brothers"; JEAN BOUVIER, *Les Rothschild*, new ed. (1967), portrays the Rothschilds' history up to the present time; EGON C. CORTI, *La Maison Rothschild, 2* vol. (1931), an excellent study of the Rothschilds in the middle of the 19th century; BERTRAND GILLE, *Histoire de* la *maison Rothschild,* 2 vol. (1965–67), an exhaustive work on the French house that makes use of the firm's records; FREDERICK MORTON, *The Rothschzlds* (1962), a brief, lively history.

(Je.B.)

Rotterdam

t one time it was ıp e to refeı to Rotterdam as a harbour with a it attached, t :hi is no longer true of the of The Netherlands. During the 1960s, Rotterdam outstripped London and New York harbours to become the busiest seaport in the world in tonnage of goods handled. Its position at the mouths of the Rhine and Maas (Meuse) rivers made it for centuries the water gate to the North Sea and the world for the industrial heartland of northern Europe. In the last years of the 19th century it began to handle shipments of oil, and its present-day petrochemical industries, offshoots of this shipping, have replaced shipbuilding and repair as the city's major economic activity.

Post-World War II reconstruction

Though its life is heavily oriented toward its harbour, Rotterdam as a city has literally risen from its ashes following the virtually complete devastation of its centre by German and Allied bombing during World War II. A totally new inner city came from the drawing boards, with a spacious and functional architecture oriented toward the river and a series of experiments at complete city planning that have attracted both professional and touristic admiration. The Lijnbaan Shopping Centre, known as the "Fifth Avenue of Europe" when it was erected in 1953, became the prototype for similar centres in Europe and America that allowed only pedestrian traffic.

A major concern of the planners was to provide the approximately 700,000 Rotterdamers with opportunities for a richer quality of life than is ordinarily encountered in a heavily industrial environment. A sense of openness in the new city and a proximity to recreational areas represent a distinct change from the 19th-century slums that infested much of the prewar city. Amsterdam remains the largest city and the cultural centre of the nation, but Rotterdam is experiencing a growing artistic life, perhaps best symbolized in De Doelen, a music centre completed in 1966 whose acoustical perfection has become world famous.

History. The name Rotterdam was mentioned for the first time in 1283, when a small tract of reclaimed land, or polder, was created by draining the mouth of the Rotte River. A fishing village sprang up, but its favourable position on a curve in the Maas River was the basis for further development. The outer portion of the dike soon became a home port for merchant vessels trading with England, France, and Germany.

Early settlement and growth. As early as 1328 the growing community was granted the prerogatives of a settled town. In 1340 the town received permission to dig a canal to the Schie, and it became the major port of the province. Industry began to develop, and in 1358 the first of many annexations of adjacent districts took place. Gradually a triangular walled town took shape and remained clearly recognizable as "the old town" until the German bombardment in May 1940.

Maritime growth in the 17th century

Although the connection with the sea remained poor until the second half of the 19th century, Rotterdam rapidly expanded as a harbour and an industrial city because of its geographical position. In the 17th century, when the discovery of the sea route to the Indies gave an enormous impetus to commerce and shipping, Rotterdam expanded its harbours and accommodations on the silty soil along the Maas. Before the end of the century it was, after Amsterdam, the second merchant city of the country. The cultural prosperity of this Golden Age in Dutch art did not completely pass up Rotterdam, for inside its walls lived painters such as Pieter de Hooch. Its most famous son was the Humanist philosopher Desiderius Erasmus.

New commercial emphasis. Rotterdam adjusted laboriously to the changed circumstances after the French occupation, which, from 1795 until the fall of Napoleon in 1815, had halted most trade. It no longer was a shipper of staple goods but was able to change to an economic life of transshipment and commissions. The colonial monopoly that had made Rotterdam an entrepreneur of goods to and from the the far-flung Dutch colonies had

The harbour at Rotterdam.
Dick Huffman—Pix

ceased to exist, and free trade had won. In response to this, Antwerp's position as a major shipping competitor began. The advent of the larger steamships made necessary a better connection with the sea. The New Waterway was dug directly through the dunes to the sea between 1866 and 1872.

Railroad connections, too, came late for Rotterdam. A line to the northern part of the country was finished in 1847 and, in 1856, to the eastern region, where commercial relations with Germany were becoming increasingly important. It was not until 1877, however, that the city was connected with the southern Netherlands by a railroad over the Maas. The viaduct, which runs directly through the city, is still unique.

The 20th century. The simultaneous construction of a traffic bridge across the Maas opened the south bank, where large harbour facilities, extending westward, sprang up between 1892 and 1898. From 1908 to 1914 a number of other facilities were built on the north bank. Between 1906 and 1930 the Waal Harbour was built, its 741 acres (300 hectares) making it the largest dredged harbour in the world.

The care for the city was always in painful contrast with the attention paid to the harbour. Around 1860 a major redevelopment featuring wider streets was accomplished, but an enormous influx of working people during the next 50 years nullified these efforts and raised the population from 100,000 in 1857 to 400,000 in 1907. Flagrant speculative building took place, saddling the city with extensive slums. It was only after World War I that architectural and city planning was able to produce attractive working class districts.

World War II devastation

World War II brought about a traumatic break in the history of Rotterdam. The city suffered a German aerial bombardment on May 14, 1940, that completely destroyed its centre, and resulted in an area of 500 acres (200 hectares) going up in flames. More than 1,000 people were killed, nearly 25,000 houses and 11,000 other buildings were destroyed, and 77,000 people were made homeless. After the German occupation, Rotterdam underwent more than 100 Allied air raids, which killed nearly 900 people and destroyed about 2,500 more homes. The city's harbour was largely saved from the bombing of 1940, but in September 1944 the Germans caused great devastation before their retreat, reducing its capacity by one-third.

The contemporary city. In the years since World War II, Rotterdam's inner city has acquired the functions and the status of an extremely modern city.

Physical layout. The triangular pattern of the traditional city has been radically broken. About 700 acres of once-fallow land now hold a completely replanned development detached from the past. The reduced acreage used for buildings in the reconstruction have provided space for broad streets, more open traffic routes, and public squares.

The essential functions of an inner city—predominantly office buildings, banks, and shopping centres—have been used to determine the utilization of land. Buildings with similar functions have been brought together systematically. The two large shopping centres, which differ in quality and price of merchandise, are on both sides of the Coolsingel, the major boulevard. The theatre and the concert hall are opposite one another on a large square that has an underground parking garage. Banks are arranged at points convenient for commercial and individual use, and wholesale merchants are brought together in a special building, the largest such business premise in Europe. These centres are connected with the other parts of the city by large shopping streets. The city hall (1918), spared in the holocaust, and the 15th-century Grote of St. Laurenskerk, which was burned in 1940 but has been restored, and a number of architecturally impressive commercial buildings form the landmarks of the new inner city.

New residential areas

This rearrangement necessitated the construction of new residential areas. North of the city, space was found in several incorporated villages. East of Rotterdam, a new satellite city, Alexanderstad, which lies 22 feet below sea level and is the lowest point in The Netherlands, houses 175,000 people. Along the southern edge of the city a belt of residential quarters was built. One of them, situated close by the oil refinery in the village of Hoogvliet, has a large and important shopping centre. Each of these new residential quarters has been designed as an independent social unit, and, as a result, Rotterdam itself has lost a considerable degree of its traditional character as a bedroom city.

Demography. Within the municipality itself a maximum population of about 700,000 has been reached. Many people who work in the area of Rotterdam reside in the rapidly growing neighbouring municipalities. Incorporation of these into the boundaries of the core city no longer fits the administrative philosophy of The Netherlands. Rotterdam and those municipalities that belong to the harbour and industrial districts are merged into a larger regional district, Rijnmond. Its population consists largely of working class people.

The economy. Rotterdam's economy is almost completely based on shipping. Lying at the heart of the densely populated and industrialized triangle of London, Paris, and the German Ruhr district and at the mouth of two important rivers, Rotterdam is open to the North Sea, the world's most heavily navigated sea. Throughout the centuries the people of Rotterdam have assisted nature by providing deep channels and harbours that can receive the largest ships comfortably. The harbour and the industries as well as the city have been revived and extended on a grand scale in the years since World War II. Before 1940 Rotterdam had specialized almost exclusively in harbour activities, and, as a result, it experienced repeated cyclical economic crises, notably the worldwide depression of the 1930s.

Attempts to attract industry

In 1947, therefore, the authorities set out deliberately to attract industry. It had to be an industry that could benefit from Rotterdam's position, and the answer was found in the oil-processing, or petrochemical, industry. Its development gave an enormous stimulus to shipping, which in turn initiated a period of tempestuous growth in the harbour district. Whereas in 1940 the total area of Rotterdam's harbour and industrial district amounted to about 6,500 acres, in slightly over 20 years it increased to about 25,000 acres. The harbours and grounds are owned by the municipality and are rented to business.

The amount of sea-transported goods assessed in this harbour is the largest in the world, amounting in 1970 to approximately 225,000,000 tons, about 60 percent of it crude oil or oil products. It is the largest grain and general-cargo harbour on the continent. As a transshipment port, it is the junction of world shipping and river transports for inland Europe, with about 225,000 Rhine barges a year in addition to trains, cars, and pipelines. It is the home port of 40 percent of the international Rhine merchant fleet. On the north bank of the 19th-century New Waterway, a large harbour was under construction in the early 1970s that would add about 1,200 acres to the harbour, an area to be used exclusively for forms of containerized merchandise. In addition, Rotterdam is a free port, meaning that goods move in and out of the port without duty, and raw materials can be brought in, processed, and reshipped with a similar exemption from duty charges.

Industrial districts

The expansion of the oil industry made necessary construction of three new large industrial districts, Botlek, Europoort, and Massvlakte, all situated on the south bank of the New Waterway. The construction of the Botlek district had begun in 1947, bordering the old harbour, through which large seagoing vessels can reach it. The district had been rented completely long before its completion, so construction of the Europoort district, nearly three times the size of Botlek, was begun. It, too, was completely filled in a short time with oil-storage facilities and petrochemical and other chemical plants. With this expansion to the west, the harbour had almost reached the sea. The only possible consequence was the construction of still another district, the Maasvlakte, which is built into the sea and gives access to the largest ships with the deepest draft. For some time, part of this district has been reserved for such metallurgical industry as blast furnaces, but the growing concern for the quality of the environment has raised increasing opposition to polluting industries.

The New Waterway, dating from the 1870s, was no longer sufficient for new industrial districts, which must be accessible to the largest tankers. South of the Waterway are the new Caland and Hartel canals, which can be freely navigated by mammoth tankers. To make these canals accessible, a channel 72 feet deep has been dredged far into the sea to accommodate tankers of 500,000 tons. In the future, tankers of 700,000 tons will find their harbour at the north side of the Maasvlakte.

Petrochemical transportation

Because of this development, Rotterdam has become a vitally important industrial centre of western Europe. There are five refineries, whose total capacity increased between 1968 and 1970 from 40,000,000 to 62,000,000 tons. Pipelines transport raw oil to Amsterdam, Antwerp, and West Germany, refinery products to West Germany, naphtha to South Limburg—the former Dutch mining district, which has switched to the chemical industry—and ethylene gas to Terneuzen in the southern island district of Zeeland, where industrial development is beginning.

Ground and air transportation. As much as the river is essential for navigation, it is an obstacle for land traffic. It divides the city in half, and cross-river connections have never caught up with the growing traffic. The city has a tunnel and a bridge, the latter to be replaced eventually by a second tunnel. Around the city a diamond-shaped beltway is being constructed to eliminate through traffic. The north–south legs of this beltway cross the river east of the city over a bridge, the Brienenoordbrug, which has the largest clear span in Europe, 974 feet, and west of the city by the Benelux Tunnel. Public transportation is maintained by trams and buses, which do not cross the river. Most of them connect with the main stations of a partially underground metro system, opened in 1968, that runs from the Central Station under the river to the new residential areas on the south side. East–west connections of the metro are in the preparatory stage.

Northwest of the city is the Zestienhoven Airport, opened in 1956, which serves planes engaged in the European air traffic. It also is used to supply intercontinental transportation from other airports, is a base for charter flights and has facilities for private aviation.

Government. Rotterdam is governed by a 45-member elected municipal council, in which all shades of political persuasion, including the Communist, are represented. Seven full-time aldermen carry out the executive functions of the municipality together with the mayor, who holds his appointment from the crown.

Cultural life and recreation. For a long time, culture and recreation were among the weakest aspects of life in Rotterdam. Reconstruction and renewal of the city and harbour, however, gave a strong stimulus to these features. Because of a lack of material, the reconstruction of the Schouwburg Theatre, begun during the war, was accomplished with the debris from the destroyed city. Plans for a new municipal theatre were in preparation in the early 1970s. A subsidized theatrical company, the Nieuw Rotterdams Toneel, which is one of the most important companies of The Netherlands, performs in the Schouwburg. For smaller performances, two small theatres, the Piccolo and Hofplein, are run by the city. The south side has its own theatre, the Grote Schouwburg, with a large variety of programs.

Drama and music

Opposite the Schouwburg is the concert hall, De Doelen, opened in 1966. Through its perfect acoustical quality, its accommodations, and its carefully considered programming, De Doelen quickly came to play an important role in international musical life. The Rotterdam Philharmonic Orchestra, which gives 80 largely sold out concerts a year, performs permanently in De Doelen. Rotterdam has a centuries-old tradition of organ playing, which draws many people to the Grote of St. Laurenskerk and to the splendid organ of De Doelen.

Artistic and historical collections and buildings

Rotterdam is particularly noted for the plastic arts, both historical and contemporary. The Boymans-van Beuningen Municipal Museum has a famous collection that includes works by Hieronymus Bosch, Pieter Brueghel the elder, Jan van Eyck, Vincent van Gogh, the Rotterdamer Kees van Dongen, and other Dutch and Flemish masters. It often organizes expositions of international importance. The city is gradually acquiring an open-air collection of sculptures by an international array of artists, including works by Auguste Rodin, Alexander Calder, Marino Marini, Henry Moore, and Naum Gabo. The work "Destroyed City" by the Russian-French sculptor Ossip Zadkine symbolizes the destruction of the city in 1940. A famous bronze statue of Erasmus was made by the 17th-century master Hendrik de Keyzer.

The most important of the other museums are the Museum voor Land en Volkenkunde (Museum for Geography and Ethnography), which has a collection of objects of fine and applied art from non-European countries, and the Maritime Museum Prins Hendrik (Prince Henry Maritime Museum), whose displays and large collection of ship models illustrate the history of navigation. The Historisch Museum (Historical Museum), a splendid building from the 17th century, was spared during the bombing. Its annex in the Delfshaven ward of the city has a workshop in which pewter is fashioned after old models. Still standing also is the church in which America's Pilgrims said farewell to Europe before starting out on the first stages of their long sea voyage.

Rotterdam holds a central position in education. Since 1913 it has had a school of economics, which has been expanded into a school for social sciences and, since 1970, a medical school. Technological education is quite varied, with a technical high school and, in the heart of the city, eight training schools. These have a regional function and are united in a single building. There is a school for training dock workers in their increasingly specialized jobs. Finally, Rotterdam has a school for the plastic arts.

Recreational and sporting facilities

For a city that seizes more and more of the surrounding grass and foliage for harbours and industry, outdoor recreation is a great problem. Rotterdam has only one old park, which is situated along the river. In the rest of the city, most grass and foliage has been artificially laid out. A forest of 630 acres (255 hectares) was planted when 34,000,000,000,000 cubic yards of sand, which came from the dredging of the Waalhaven on the south side, were deposited into a polder to the east of the city. After World War II a similar belt of woods was laid out on the south side between old and new residential areas; it has a variety of recreative functions. The municipal parks are part of an experiment in "fencing forests" to protect the living quarters in the industrial districts against air pollution.

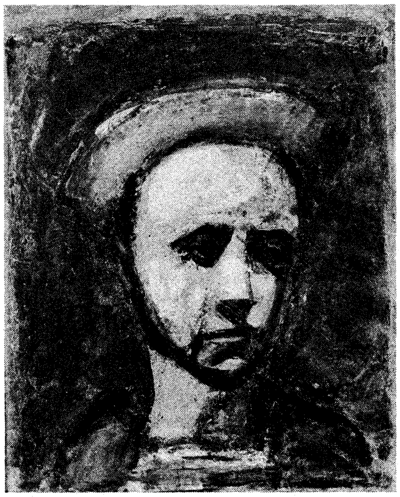
Two large lakes in the eastern part of the city offer facilities for many kinds of water sports. Their artificial beaches draw hundreds of thousands of visitors on warm summer days. Many Rotterdamers travel to the North Sea beaches that are close by in the Hook of Holland at the mouth of the New Waterway. Since 1914 it has belonged to the district of Rotterdam. The Blijdorp, opened in 1940, is one of the most modern gardens in Europe. For spectator sports, the city has 250 fields. In 1970 the sports palace Ahoy, seating up to 10,000 spectators, was opened on the south bank.

BIBLIOGRAPHY. For general history, see *Gedenkboek . . . Rotterdam 1328–1928* (1928); KEES HAZELZET, *Rotterdam van Vissersdorp tot Wereldhaven* (1943); L.J.C.J. VAN RAVESTEYN, *Rotterdam in de Twintigste Eeuw* (1948); JAN VERSEPUT, *Kamer van Koophandel en Fabrieken voor Rotterdam, 1928–1953* (1955); and HANS REINHARDT, *The Story of Rotterdam* (1955). The port is treated in JAN SCHRAVER *et al.*, *Rotterdam, de Poort van Europa* (1946; Eng. trans., *Rotterdam, the Gateway to Europe,* 1948); for more recent development, see REINDER BLIJSTRA, *Rotterdam Stad in Beweging* (1965); and JAN W. DE BOER and CAS OORTHUYS, *Rotterdam* (1959; Eng. trans., 1963). On World War II and reconstruction, AAD WAGENAAR, *Rotterdam Mei '40* (1970), is of interest; see also Ivo BLOM, *Rotterdam in Pictures* (1970).

(H.W.E.)

Rouault, Georges

Among the major artists of the 20th-century school of Paris, the French painter Georges Rouault was an isolated figure in at least two respects: he practiced Expressionism, a style that has never found much favour in France, and he was chiefly a religious painter—one of the most convincing in recent centuries—at a time when most modern painters were preoccupied with aesthetic problems. Both statements, however, need qualification. Rouault was not as fiercely Expressionistic as some of his Scandinavian and German contemporaries; in some ways his work can be regarded as a late flowering of 19th-century Realism and Romanticism (which helps to explain his wide popularity). And he was not an official church artist; his concern with sin and redemption was deeply personal.



"The Workman's Apprentice (Self-Portrait)," in oils by Georges Rouault, 1925. In the Musée National d'Art Moderne, Paris.

Giraudon

Apprenticeship to a glazier

Georges-Henri Rouault was born in Paris on May 27, 1871, in a cellar during a bombardment of the city by the forces opposed to the Commune. His father was a cabinetmaker. A grandfather took an interest in art and owned a collection of Honor6 Daumier's bitterly satirical lithographs; Rouault said later that he "went first to school with Daumier." In 1885 he enrolled in an evening course at the Paris École des Arts Décoratifs. From 1885 to 1890 he was apprenticed in a glazier's workshop; his mature style as a painter was undoubtedly influenced by his work on the restoration of medieval stained-glass windows, including those of Chartres Cathedral. In 1891 he entered the Ecole des Beaux-Arts, where he soon became one of the favourite pupils of the Symbolist painter Gustave Moreau, in a class that also included the young Henri Matisse and Albert Marquet. After the death of Moreau in 1898, a small Paris museum was created for his pictures, and Rouault became the curator.

During this early phase of his career Rouault's style was academic. But around 1898 he went through a psychological crisis, and, during the following years, partly under the influence of Vincent van Gogh, Paul Gauguin, and Paul Cézanne, he evolved in a direction that made him, by the 1905 Paris Salon d'Automne, a fellow traveller of the Fauves (Wild Beasts), who favoured the arbitrary use of strong colour. At this time and up until the beginning of World War I, his most effective medium was watercolour or oil on paper, with dominant blues, dramatic lighting, emphatic forms, and an expressive scribble.

Religious evolution

Rouault's artistic evolution was accompanied by a religious one, for he had become, around 1895, an ardent Roman Catholic. He became a friend of the Catholic intellectuals Joris-Karl Huysrnans and Léon Bloy and took to heart the latter's warning that modern art had to "go on its knees." Through another friend who was a deputy public prosecutor, he began to frequent, as had Daumier, the Paris law courts, where he had a close-up view of human beings apparently fallen from the grace of God. His favourite subjects became hideous prostitutes, tragic clowns, and pitiless judges.

Paintings on the Passion of Christ

Without completely abandoning watercolour, after 1914 Rouault turned more and more toward the oil medium. His paint layers became thick, rich, and sensuous, his forms simplified and monumental, and his colours and heavy black lines reminiscent of stained-glass windows. His subject matter became more specifically religious, with a greater emphasis on the possibility of redemption than he had put into his pre-1914 work. In the 1930s he produced a particularly splendid series of paintings on the Passion of Christ; typical examples, well-known through best-selling reproductions, are "Christ Mocked by Soldiers,""The Holy Face," and "Christ and the High Priest". During these years he got into the habit of reworking his earlier pictures; "The Old King", for instance, has to be dated 1916–36.

Between World Wars I and II, at the instigation of the

Paris art dealer Ambroise Vollard, Rouault devoted much time to engravings, illustrating Les Rkincarnations *du* Pkre *Ubu* by Vollard, Le Cirque de *l'étoile filante* by Rouault himself, Les *Fleurs du mal* by Charles Baudelaire, and *Miserere* et *guerre* (his masterpiece in the genre), with captions by Rouault. Some of this work was left unfinished for a time and published later. In 1929 he designed the sets and costumes for a production by Sergey Diaghilev of Sergey Prokofiev's ballet The Prodigal Son. In 1937 he also did the cartoons for a series of tapestries.

During and after World War II, he painted an impressive collection of clowns, most of whom have features—long noses, pursed mouths, and domed foreheads—that make them virtual self-portraits. He also executed some still lifes with flowers; these are exceptional not only because of their quality but also because of the subject, for three-quarters of his lifetime output is devoted to the human, or divine, figure. In 1947 he sued the heirs of Vollard to recover a large number of works left in their possession after the death of the art dealer in 1939. Winning the suit, he established the right of an artist to things never offered for sale, and afterward he publicly burned 315 canvases that he felt were not representative of his best work.

During the last 10 years of his life, he renewed his palette, adding some greens and yellows unusual for him, and painted some almost mystical landscapes: a good example is "Christian Nocturne". Rouault died in Paris on February 13, 1958.

MAJOR WORKS

"The Road to Calvary" (1891; Wadsworth Atheneum, Hartford, Connecticut); "The Ordeal of Samson" (1893; Los Angeles County Museum of Art); "Duet" (1904; Musée d'Art Moderne de la Ville de Paris); "Head of Christ" (1905; Walter P. Chrysler, Jr. Collection, New York); "Aunt Sallys" (1907; Tate Gallery, London); "Judges" (1908; Statens Museum for Kunst, Copenhagen); "Portrait of Mr. X (1911; Albright-Knox Art Gallery, Buffalo, New York); "Christ Mocked" (1912; Musée d'Art Moderne de la Ville de Paris); "The Old King" (1916–36; Carnegie Institute, Museum of Art, Pittsburgh); "Three Clowns" (1917; Joseph Pulitzer Jr. Collection, St. Louis, Missouri); "Crucifixion" (1918; Henry P. McIlhenny Collection, Philadelphia); "Pierrot" (1920; Stedelijk Museum, Amsterdam); "Circus Trio" (1924; Phillips Collection, Washington, D.C.); "The Workman's Apprentice (Self-Portrait)" (1925; Musée National d'Art Moderne, Paris); "Christ Mocked by Soldiers" (1932; Museum of Modern Art, New York); "The Holy Face" (1933; Musée National d'Art Moderne, Paris); "Christ and the High Priest" (1937; Philips Collection, Washington, D.C.); "Christ and Fishermen" (1937; Musée du Petit Palais, Paris); "Sunset" (1937; Worcester Art Museum, Massachusetts); "The Three Judges" (1937–38; Tate Gallery, London); "Head of Christ" (1938; Cleveland Museum of Art); "Diplomat" (1938; Musées Royaux des Beaux-Arts de Belgique, Brussels); "Christ Mocked" (1942; Staatsgalerie, Stuttgart); "De Profundis" (1946; Musée National d'Art Moderne, Paris); "Head of a Clown" (1948; Museum of Fine Arts, Boston); "Christian Nocturne" (1952; Musée National d'Art Modem, Paris); "Decorative Flowers" (1953; private collection, Paris).

BIBLIOGRAPHY. Standard biographical and critical studies are LIONELLO VENTURI, *Georges Rouault*, 2nd ed. (1948); and PIERRE COURTHION, *Georges* Rouault (1962). The catalog of the ARTS COUNCIL OF GREAT BRITAIN, *Rouault:* An Exhibition of Paintings, Drawings, and Documents, 2nd ed. (1966), is an excellent factual survey.

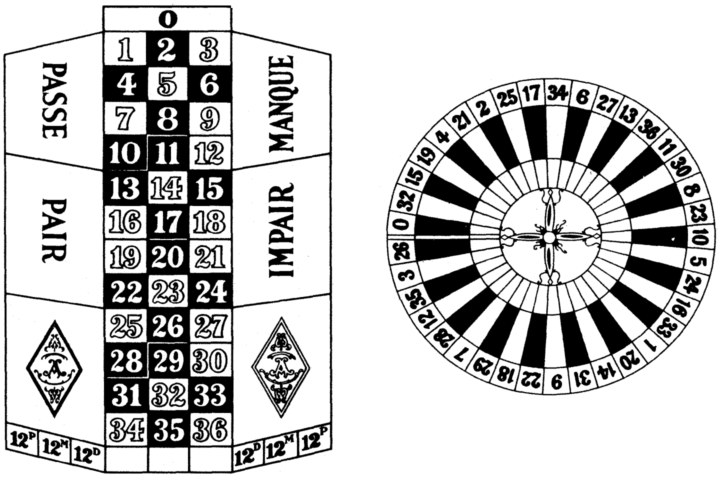
(R.McMu.)

Roulette

Roulette is a casino gambling game of obscure but probably French origin. It is universally played in the gambling casinos of Europe, North and South America, Asia, and Africa. Roulette is a banking game, and all bets are placed against the bank—in this case the "house" or proprietor of the game. As a big-time betting game its popularity has been superseded in the United States and the Caribbean Islands by others, notably craps and black jack.

History. Roulette is said by some historians to have been invented in 1655 by the French scientist Blaise Pascal during his monastic retreat and first played in a makeshift casino in Paris. Others say that it was invented by a French monk to help break the monotony of monastery life. Still others say it originated in an old Chinese game whose object was to arrange 37 statuettes of animals into a "magic square" of 666, but they fail to describe the method of play. They add that the game was later played in Tibet and eventually by French Dominican monks, one of whom transposed the statuettes into numbers, from 0 to 36, and arranged them haphazardly along the rim of a revolving wheel. Whatever its antecedents and origins, it was only in the late 18th or early 19th century that roulette emerged as a glamourous attraction in the casinos of Europe, where it has long been associated with the gaming rooms at Monte Carlo.

Equipment. The roulette table is composed of two sections, the wheel itself and the betting layout, better known as the roulette layout. There are two styles of roulette tables. One has a single betting layout with the roulette wheel at one end; the other has two layouts with the wheel in the centre.



Roulette layout and (right) roulette wheel with 37 compartments, both European style.

European- and American-style layouts and wheels

Heading the layout design, which is printed on green baize, is a space containing the figure 0 (European style), or the figures 0 and 00 (American style). The main portion of the design is comprised of 36 numbered rectangular spaces, coloured red and black alternately, and arranged in three columns of 12 spaces each. Directly below the numbers are three blank spaces (on some layouts these are marked 2 to 1 and are located on the players' side of the table). On either side of these, or along one side of the columns, are rectangular spaces marked 1st 12, 2nd 12, and 3rd 12. On European-style layouts these terms are 12^{p} (premidre), 12^{m} (milieu), and 12^{d} (*dernière* douzaine). Six more spaces are marked red (rouge), black (noir), even (pair), odd (impair), 1–18 (low or manque), 19–36 (High or passe).

The roulette wheel consists of a solid wooden disk slightly convex in shape. Around its rim are metal partitions known as separators or frets, and the compartments or pockets between these are called canoes by roulette croupiers. Thirty-six of these compartments, painted alternately red and black, are numbered from 1 to 36. On European-style wheels, a 37th compartment, painted green, carries the sign 0, and on American wheels two green compartments carry the signs 0 and 00. The wheel, its spindle perfectly balanced on a single ball bearing, spins smoothly, in an almost frictionless manner.

The standard roulette table employs five, six, or seven sets of wheel checks (usually called chips). Each set is differently coloured, each consists of 300 chips, and there is one set for each player. The chips usually have a single basic value, although some casinos also sell chips of lesser value. The colour of the chips indicates the player, not the value of the chips. If a player wishes to buy chips of slightly higher value, the croupier places a marker indicating that value on top of the table's stack of chips of the colour corresponding to the chips purchased. Most casinos have high value chips that can be wagered at any gaming table. Unlike roulette chips, these have their numbered values printed on them.

Payoff odds

Bets. It is possible to place the following bets in roulette: (1) straight, or single-number (en plein), in which the chips are placed squarely on one number of the layout, including 0 (and also 00 on American layouts), so that the chips do not touch any of the lines enclosing the number; a winning single number bet pays 35 to 1; (2) split, or two-number (*à* cheval), in which the chips are placed on any line separating any two numbers; if either wins, payoff odds are 17 to 1; (3) street, or three-number (transversale pleine), in which the chips are placed on the outside line of the layout, betting the three numbers opposite the chips; payoff odds on any of the three numbers is 11 to 1; (4) square, quarter, corner, or four-number (en *carre*), in which the chips are placed on the intersection of the lines between any four numbers; payoff odds are 8 to 1; (5) line, or six-number (transversale six), in which the chips are placed on the intersection of the side line and a line between two "streets"; payoff odds are 5 to 1; (6) column, or twelve-number, in which the chips are placed on one of the three blank spaces (some layouts have three squares, marked lst, 2nd, and 3rd) at the bottom of the layout, thus betting the 12 numbers above the space; payoff odds are 2 to 1; (7) dozens, or twelve-number, in which the chips are placed on one of the spaces of the layout marked 12, betting the numbers 1–12, 13–24, or 25–36; payoff odds are 2 to 1; (8) low-number or high-number, in which the chips are placed on the layout space marked 1–18 (*manque*) or on the space marked 19–36, (passe); payoff is even money; (9) black or red, in which the chips are placed on a space of the layout marked black (noir), or on a space marked red (rouge; some layouts have a large black or red diamond-shaped design instead of the words); payoff is even money; (10) odd-number or even-number, in which the chips are placed on the space of the layout marked odd (impair), or on the space marked even (pair); payoff is even money.

On layouts with a single zero (European style), the 0 may be included in a two-number bet with any adjoining number, in a three-number bet with 1 and 2 or with 2 and 3, and in a four-number bet with 1, 2, and 3 at the regular odds for these bets; with the American style 0 and 00, a five-number line bet also is possible, the player placing his chips on the corner intersection of the line separating the 1, 2, 3, from the 0 and 00, with payoff odds of 6 to 1.

The play. The game begins when one of the croupiers (dealers) in attendance calls, "Make your bets." The players begin making their bets (as many may bet as can get near the table) by placing chips on the spaces of the layout, on any number, group, or classification they hope will win.

Role of the croupier

The croupier starts the wheel spinning in a counterclockwise direction, then spins **a** small ivory or plastic ball on to the bowl's back track in the opposite direction. Players may continue to place bets while the wheel and ball are in motion until, as the ball slows down and is about to drop off the back track, the croupier calls, "No more bets!"

When the ball falls and comes to rest between any two metal partitions of the wheel, it marks the winning number (or a zero or double zero), the winning colour, and any other permitted bet that pertains to a winning number or symbol. The dealer immediately announces the winning number and its colour, and he points with his index finger to the corresponding number on the layout. He first collects all losing bets, not disturbing the chips that are resting on winning spaces. Then he pays off the winner or winners with the correct amount of chips due to each winning bet.

House **odds.** When using the American-style wheel with the 0 and 00, the advantage for the bank arises when either the 0 or 00 shows. Only bets on the winning symbol (0 or 00) and the line or five-number bet (combination bet 0, 00, 1, 2, 3) are paid. Thus the bank should win 2 parts in 38 or $5\frac{5}{19}$ percent of all bets. The only exception is the line or five-number bet where the house percentage is $7\frac{17}{19}$ percent.

Roulette as played in Europe, Africa, South America, and Asia is the same as that played in the United States, with the exception that the wheel and layout contain only a single zero (0). When the zero appears, all even-money bets, such as Red, Black, Odd, or Even, are "imprisoned." On the next spin of the wheel, if the zero appears again the house collects half of each imprisoned bet; if not, it collects all losing bets and returns the original bets to any winners. The bank's percentage on red, black, odd or even is $1\frac{13}{37}$ percent, and on all other types of bets it is $2\frac{26}{37}$ percent.

Systems. The oldest and most common betting system is the Martingale or "Doubling-up" system, in which bets are doubled progressively. This probably dates back to the invention of the roulette wheel, but every day of the week some gambler somewhere reinvents it, or some variation of it, and believes he has something new. Over the years hundreds of "sure-fire" winning systems have been dreamed up, but regardless of what system is used, in the long run it cannot overcome the house's advantage of the 0, or 0 and 00. This house advantage is the only system that works in the long run.

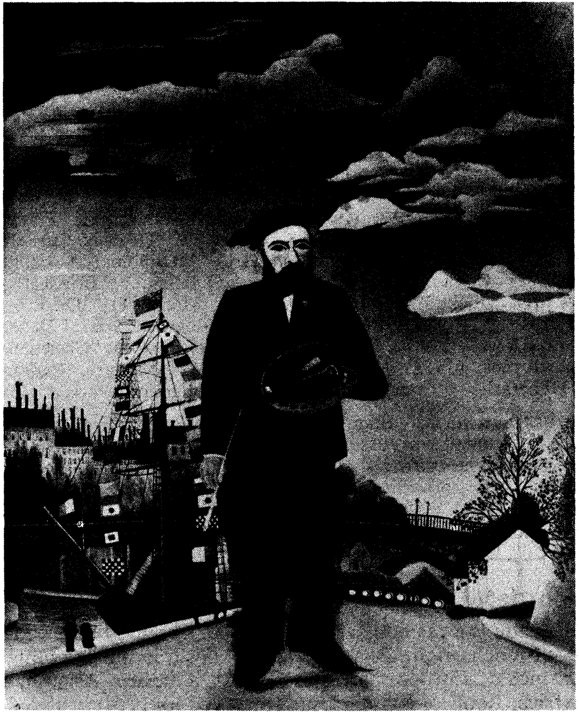
BIBLIOGRAPHY. JOHN SCARNE, Scarne's *Complete* Guide to Gambling (1961), The Woman's Guide to Gambling (1967); N. SQUIRE, *How* to Win at Roulette (1968).

(J.S.)

Rousseau, Henri

The French painter Henri Rousseau represents a special case in modern art. His painting, virtually untaught, was of the primitive, "Sunday painter" genre; yet, despite its definite technical naïveté, it reached brilliant heights of expression and imagination and was greatly admired by the most advanced painters. An unsophisticated man and by profession a petty clerk, he lived most of his life in obscurity, but in his last years he moved, unchanged, in the most sophisticated artistic circles of his day.

Giraudon



Rousseau, self-portrait, oil painting, 1890. In a private collection, Paris.

Born in 1844 in Laval, France, and raised there, Rousseau, the son of a tinsmith, was from a modest background. **A** mediocre student, he left the secondary school in Laval without having completed his studies and soon entered military service, in which he remained for four years. During his term of service he met soldiers who had survived the French expedition to Mexico (1862–65) in support of Emperor Maximilian, and he listened with fascination to their recollections. Their descriptions of the subtropical country were doubtless the first inspiration for the exotic landscapes that later became one of his major themes. The vividness of Rousseau's portrayals of jungle scenes led to the popular conception, which Rousseau never refuted, that he travelled to Mexico. In fact, he never left France.

Civil-service career and early paintings. Released from military service upon the death of his father to support his widowed mother, he settled in 1868 in Paris. The following year he married the daughter of a cabinetmaker, Clémence Boitard. In Paris he began a career as a petty official, eventually becoming, in 1871, a tax collector in the Paris toll office; from this post came the name by which in later years he was well-known, *le douanier,* or "customshouse officer," in spite of the fact that the toll office had no real customs functions. Working as a bureaucrat and busy with family affairs, he still somehow found time to draw and paint. Although no works remain as evidence, he had probably drawn and painted since childhood, and his stated ambition was to be a painter in the style of the academicians of his day. In 1884 he obtained permission to copy paintings at the Louvre. In 1886 he exhibited his first painting, not at the official Salon, which would never have admitted a painter of such naïveté, but at the Salon des Indépendants, an annual exhibition established by young painters to allow themselves and other painters to exhibit and still be free from the narrow official requirements of style and subject matter.

"Carnival Evening" and the early Salon des Indépendants

The picture with which Rousseau made his debut, "Carnival Evening," was, in fact, a masterpiece of primitive painting and an impressive beginning for the artist. This work exhibits an approach to representation that is typical of the primitive artist––everything is literally and deliberately drawn; every branch of the trees is traced, the clouds have a curious solidity, and greater attention is paid to the details of costume than to the figures themselves. The design of the painting, however, is effectively poetic, and a striking quality of atmosphere and mood is achieved through the accurate and sensitive observation of the colours of the evening and through the literal treatment of trees and clouds that is ultimately unreal and contributes to an air of mystery.

In spite of this auspicious beginning, Rousseau's work still went largely unnoticed, except for the consistent ridicule of the critics, for the next seven years. During this period he exhibited some 20 paintings at the Salon des Indépendants, but he remained essentially an amateur, dividing his time among painting, work at the toll house, and family life. His wife, who had been ill for some time, died in 1888, and within several years he lost all of his family except for a daughter, whom he sent to live with relatives.

The exposition of 1889

This period of personal hardship was also a period of increased artistic activity. An important event in Rousseau's life at this time was the Universal Exposition, held in Paris in 1889. Being a simple man but one with great imagination, he was profoundly impressed by what he saw there. It is probable that the reconstructions of Senegalese, Tonkinese, and Tahitian landscapes at the exposition provided further inspiration for the exoticism of his later paintings. His enthusiasm for the fair was so great that he wrote a vaudeville play entitled ***A Visit to the Exposition of 1889,*** which, however, he did not succeed in having produced. In this play, as in other theatrical works he wrote, his naïveté revealed itself even more than in the technical aspects of his painting. Also. revealed, however, was an intense desire to express himself artistically; he even attempted to compose music. Still, his only real gift was for painting. Like his contemporaries the Impressionists, he was attracted by landscapes, and at this time he wished to imitate nature. His painting "The Toll-House" (*c.* 1900; Courtauld Institute Galleries, London) shows his place of work; an early photograph of the same location shows that Rousseau conferred on his depiction a profound lyricism while remaining loyal to reality. The most important work of this period in Rousseau's career is his self-portrait "Myself: Portrait-Landscape." Standing in the foreground, palette in hand, Rousseau is surrounded by the Parisian landscape, painted with great accuracy. This is obviously intended as a "portrait of the artist" in the academic tradition; the seriousness of purpose is impressive in spite of the naïveté of execution.

"The Sleeping Gypsy"

Later paintings and recognition. In 1893 Rousseau retired from the toll house to devote himself entirely to painting. Soon afterward, he met Alfred Jarry, a brilliant young writer, also from Laval, whose nonconformity shocked his contemporaries. He was struck by Rousseau's unusual work and introduced the self-taught artist to the circle of intellectuals associated with the avant-garde review ***Le Mercure de France.*** It was this review that first published an article praising Rousseau. The article was written in connection with his painting "War," an allegory exhibited at the 1894 Salon des Indépendants, which demonstrated Rousseau was much more than a minor landscapist. This work also marked the beginning of the recognition of Rousseau as a serious painter. His most important painting of this period was "The Sleeping Gypsy" (1897). It portrays a female Gypsy asleep in a moonlit desert with a huge lion standing over her, seemingly transfixed and unwilling to touch her. The landscape is completely bare except for the Gypsy's jug and mandolin. This painting is exceedingly primitive in its technique; the Gypsy lies stiffly on the ground, still clutching her walking staff, and her smiling face is childishly rendered. The stripes of her dress and the hairs of the lion's mane are individually traced in a naïve but decorative, almost abstract manner. The painting, however, is wonderfully expressive. The Gypsy's smile, the lion's staring eye, the bare, unearthly landscape, and the whimsical twist at the end of the lion's tail unite opposing feelings of peace and danger, solemn mystery and whimsy, into a powerful expression of magical enchantment. When he exhibited the painting, Rousseau wrote to the mayor of his native Laval asking him to purchase it; he intended it to be a tribute to his native town. The mayor, however, was merely amused at the idea. Rousseau, by this time, had enormous confidence in his own work and considered himself to be a great painter. Not only was he unaware of his lack of conventional technical skill, but he believed that his work resembled that of the academic painters. He still dreamed of official glory, but his painting was appreciated only by young avant-garde painters such as Robert Delaunay, Picasso, and their common defender, the poet Guillaume Apollinaire, who also became Rousseau's principal supporter.

In 1905 Rousseau was invited to the Salon d'Automne (a semi-official exhibition created after a schism among the academicians), where his painting "The Hungry Lion" was hung in the same room with the works of some of the most popular painters of the day, who were called Fauves (Wild Beasts): Henri Matisse, André Derain, and Maurice de Vlaminck. The work of these men, with its pure colours and lack of conventional realism, was similar to Rousseau's. At last the critics began to speak of him. Ambroise Vollard, the most important dealer in modern paintings in Paris, bought pictures from him, and Rousseau emerged from the shadows.

Rousseau lived in a humble quarter of Paris, where he gave painting lessons in his home. His second wife died in 1903 (he had married her in 1899). Because of his aggrandized self-image in the midst of such poor surroundings, he seemed to be an eccentric to the people of his district. Among avant-garde artists, however, he became an object of affectionate admiration and curiosity, and he was a popular figure in the most advanced intellectual circles. In 1908 Picasso organized in his studio an unforgettable banquet in Rousseau's honour—partly a joke on the innocent seriousness of the *Douanier* but more a tribute to his indisputable genius—to which the most sophisticated artists and critics of his day were invited.

The exotic landscapes

During these last years Rousseau painted chiefly exotic landscapes, of which "The Hungry Lion" was the first major example. He excelled in these works, creating with his primitive approach a unique mode in depicting this type of scene. The paintings are characterized by a profusion of exotic plant growth painted with great attention to detail; many varieties of plants, undoubtedly studied at the Paris botanical garden, are distinctly differentiated, with obvious fascination for the different leaf forms. Although crowded together, the plants are deliberately treated, with each leaf painted separately and each branch of leaves constituting a beautiful abstract pattern. In the midst of this vegetal density, colourful birds flit about and mysterious animals stare out at the viewer. There is usually some dramatic incident taking place in the centre, such as a lion attacking its prey, which is in keeping with Rousseau's continued predilection toward the grandiose historical, dramatic narratives of the academic tradition of painting of his time.

"Yadivigha's Dream"

Shortly before his death, in 1910, Rousseau painted the most ambitious of these jungle paintings and one of his greatest works, "Yadivigha's Dream." In this impressive fantasy, an enchanting nude rests on a red-plush Victorian sofa in the middle of a dense jungle. Huge flowers wave about her head, two lions and an elephant peer out of the undergrowth, and a darkskinned musician plays a flute behind her. Rousseau's explanation of this scene is that the lady, having fallen asleep on the sofa, dreams fhat she is transported to this improbable region. The very simplicity of Rousseau's conception and of his cast of mind allows a directness of purpose that accounts for the astonishing unconventionality of, the subject. The painting, which exhibits all of Rousseau's descriptive and expressive skill, is also a supreme revelation of his powerful and uncommon imagination.

Rousseau died on September 2, 1910. His career had been remarkable in the history of art, and his work not only inspired a revival of true primitive painting in the 20th century but also influenced a branch of modem art.

MAJOR WORKS

"Landscape with Tree Trunks" (1887; Barnes Foundation, Merion. Pennsylvania): "The Present and the Past" (*c.* 1889; Barnes Foundation); "Myself: Portrait-Landscape" (1890; Museum of Modern Art, Prague); "L'Octroi" (c. 1890; Courtauld Institute Galleries, London); "The Artillerymen" (c. 1893; Solomon R. Guggenheim Museum, New York); "War" (1894; Louvre, Paris); "The Child Among Rocks" (c. 1895; Philadelphia Museum of Art, Philadelphia); "The Sleeping Gypsy" (1897; Museum of Modern Art, New York); "Banks of the Seine" (c. 1898; private collection, Paris); "Spring in the Valley of the Bièvre" (c. 1904; Metropolitan Museum of Art, New York); "The Hungry Lion" (1905; private collection, Switzerland); "L'Herbage" (c. 1906; Fogg Art Museum, Cambridge, Massachusetts); "The Repast of the Lion" (1907; Metropolitan Museum of Art, New York); "The Snake-Charmer" (1907; Louvre); "La Bougie Rose" (1907; Phillips Collection, Washington, D.C.); "Football Players" (1908; Solomon R. Guggenheim Museum); "The Rabbit's Feeding" (1908; Barnes Foundation); "The Muse Inspiring the Poet" (1909; State Pushkin Museum of Fine Arts, Moscow); "Promenade au Luxembourg" (1909; State Pushkin Museum of Fine Arts, Moscow); "The Muse Inspiring the Poet" (1909; Kunstmuseum, Basel, Switzerland); "Vase of Flowers" (1909; Albright-Knox Art Gallery, Buffalo, New York); "View of the Île Saint-Louis" (1909; Phillips Collection); "Yadivigha's Dream" (1910; Museum of Modern Art, New York); "The Cascade" (1910; Art Institute, Chicago); "Landscape" (c. 1910; State Pushkin Museum of Fine Arts, Moscow); "Tropical Forest with Monkeys" (1910; John Hay Whitney Collection, New York).

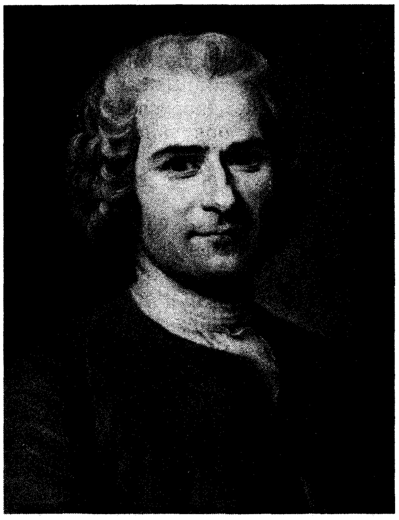
BIBLIOGRAPHY. W. UHDE, Henri Rousseau (1914), the foremost contemporary work (in French) fundamental to studies of Rousseau; C. ZERVOS, Henri Rousseau (1927), a work (in French) written by an eminent critic of modern art, with excellent reproductions; D.C. RICH, Henri Rousseau (1942), a serious study containing valuable information on Rousseau's paintings preserved in the United States; H. CERTIGNY, *La Vérité* sur *le Douanier* Rousseau (1961). D. VALLIER, Henri Rousseau (1961; Eng. trans., 1964), the most important monograph; Tout *l'oeuvre* peint de Henri Rousseau: documentation et catalogue *raissoné* (1970), a scientific study containing all of Rousseau's paintings, classified for the first time in chronological order, with an appendix where the works attributed to Rousseau appear.

(D.V.)

Rousseau, Jean-Jacques

Jean-Jacques Rousseau was one of the greatest of the European thinkers of the 18th century whose writings inspired the leaders of the French Revolution and influenced what became known as the Romantic generation. As a philosopher, he tried to achieve a synthesis between Christianity and the Rationalist and Materialist thought of his time. He named this synthesis "materialism of the wise" or "theism" or "civil religion." In politics, his theory of "social contract" went beyond both the economic liberalism of English thinkers and the Positivist attitude of Montesquieu, a French political philosopher. His effort toward a "natural education" and his search for a freely accepted "contract" between teachers and pupils have been the seed of all modern pedagogical movements. His thought has had broad influence because he expressed himself not only in political and philosophical treatises but also in beautifully written novels and autobiographical works.

By courtesy of the Musee d'Art et d'Histoire, Geneva: photograph, Jean Arlaud



Rousseau, drawing in pastels by Maurice-Quentinde La Tour, 1753. In the Musée d'Art et d'Histoire, Geneva.

Genevan background

Formative years. Jean-Jacques was born on June 28, 1712, the son of Isaac Rousseau, a citizen of Geneva and a watchmaker, and Suzanne Bernard, niece of a Calvinist minister. His mother died a few days after his birth. His education began exceptionally early; at the age of six, he was reading books from his father's library. He began with novels and comedies but immediately passed on to the works of historians such as Plutarch and some of the great moralists. Discussions in his father's workshop and in workers' clubs filled him with republican principles and with puritanical interpretations of the Bible. His Confessions (1782) and his Lettre *à d'Alembert* sur *les* spectacles (1758; "Letter to d'Alembert on the Theatre") are the most precise documents on life in Geneva during the 18th century and on his self-education.

The working class elite in Geneva was concentrated in the Saint-Gervais area, situated in the "lower city," as opposed to the "upper city" of the wealthy merchants and bankers. In the time of Jean-Jacques, the Saint-Gervais part of Geneva was periodically the scene of social troubles. In 1722 Isaac went into exile because he had a quarrel with a member of an influential family; he left his two sons, Jean-Jacques and his older brother François, in the care of their maternal uncle, Gabriel Bernard. François was apprenticed, and Jean-Jacques was sent to live outside Geneva with a minister, J.-J. Larnbercier, who taught him the classics. In September 1724 he returned to his uncle in Geneva and soon after began work as a clerk. In April 1725 he was apprenticed to Abel Ducommun, an engraver, who treated him roughly. Rousseau reacted by reading prohibited books during his working hours, and he joined a gang of youngsters. On March 14, 1728, he fled from Geneva by means of a network organized by the Roman Catholic clergy of Savoy.

From his Genevan childhood, Rousseau retained a few simple but profound ideas calling for the sovereignty of the people, taxation through popular vote, the necessity of free and periodic popular assemblies, but the dangers of the delegation of power. Nevertheless, after he left Geneva, Rousseau had to shed both the Calvinism and the republicanism of his father.

Relationship with Mme de Warens

Rousseau found his way to Annecy, in Savoy, and to Mme de Warens (Louise-Elkanore de la Tour du Pil), who had left her husband and become a convert from Pietism to Roman Catholicism. Under her protection he was sent to Turin, to the Hospice of the Holy Spirit, which he entered on April 12, 1728. According to the register of the hospice, he abjured the heresies of Calvin on April 21 and was received into the Roman Catholic Church two days later, though in the ***Confessions*** he describes a much longer resistance to conversion and says that he finally succumbed in order to escape the pernicious moral atmosphere of the hospice. For several months he stayed in Turin, working as an engraver and as a lackey. He returned to Mme de Warens in the spring of 1729 and went for a time to a seminary, but he soon found that he had no vocation to the priesthood. He then set himself up as a teacher of music to girls from some of the wealthiest families in Annecy and later in Neuchâtel. In 1731 a trip to Paris for a few days led to nothing. Rousseau returned on foot to Mme de Warens, who by this time had moved to Chambéry. She offered herself to him as his mistress in 1733, saying that she thought it would be good for him; he accepted, though he felt himself more her son than her lover. He stayed with her until 1740, and she found him a position in the tax office of Chambéry and then at a small farm nearby called Les Charmettes. It was during this period that he became a diligent but unsystematic reader and that he first began to write.

In May 1740 he went to Lyons to tutor the children of M. Jean-Bonnet de Mably, *prévôt des maréchaux* ("provost marshal") and elder brother of Etienne Bonnot de Condillac, a noted philosopher and psychologist, and of the Abbé de Mably, a well-known political writer; but when his year's contract was over he was not asked to return. In 1741 or 1742 Rousseau set out again for Paris with the scheme for a new system of musical notation, the rough draft of a comedy *(Narcisse),* and two letters in verse dedicated to friends in Lyons. He failed to gain the success he had hoped for in Paris, but the publication in 1743 of his *Dissertation sur la musique moderne* and his *Épître à Bordes* together with the composition of an opera, *Les Musesgalantes* (1745), a comedy, *Les Prisonniers de guerre,* and some works on chemistry opened for him the doors to the wealthy family of Claude Dupin, a *fermier général* (banker) and counsellor to the King. Rousseau attempted to court Mme Dupin and failed. He then left Paris for Venice and became the private secretary to the French ambassador, with whom he quarrelled so violently, however, that he lost his post and returned to Paris in August 1744.

Years in Paris with the Encyclopédistes. Early in 1745 he took up with Thérèse Levasseur, a chambermaid at the hotel where he was staying. He had several children by her, who were all sent to a foundling home. Eventually, in 1768, he married her in a civil ceremony. Shortly after returning to Paris, he also met and established a friendship with Denis Diderot, then still a young philosopher going through an intellectual evolution from Skepticism to Materialism; he was to have a profound influence on Rousseau. A performance of *Les Muses galantes* in the home of La Poplinière, a *fermier général,* and his work on the score of the *Fêtes de Ramire,* an opera with words by Voltaire and music by Jean-Philippe Rameau, once again admitted him to the Dupin family. They employed him as secretary from 1745 to 1752. He did research for Mme Dupin, who was preparing a book on the subject of women, and he helped Dupin in a project to refute *De l'esprit des lois* (1748; Eng. trans., *The Spirit of Laws,* 1750) by Montesquieu.

Work on the *Discours sur les sciences et les arts*

At the same time, however, Rousseau was working for himself. Encouraged by Diderot, who was then in prison in Vincennes, he competed for the prize offered by the Dijon Academy in 1750 for an essay on the question whether the restoration of the sciences and the arts had tended to purify morals. His essay, known under the abbreviated title of *Discours sur les sciences et les arts,* won first prize; its publication at the end of the year made him famous. The published version did not completely reveal the boldness of Rousseau's ideas, however *—e.g.,* the subtitle *Libertk* was not used, and several passages directed against the tyranny of kings and the hypocrisy of the clergy were left out. Nevertheless, he attacked the sciences and especially arts and literature as enslaving and corrupting and as instruments of propaganda and sources of greater wealth in the hands of the rich. His attacks contained the seeds of criticism that would be applied to all the institutions of monarchic Europe. He wrote:

> Princes always view with pleasure the spread among their subjects of the taste for the arts and for superfluities that do not result in the exportation of money. For, besides fostering that spiritual pettiness so appropriate to slavery, they know well that the needs that people create for themselves are like chains binding them. . . . The sciences, letters, and arts . . . wind garlands of flowers around the iron chains that bind them [the people], stifle in them the feeling of that original liberty for which they seemed to have been born, make them love their slavery, and turn them into what is called civilized people.

The *Discours* provoked a series of violent disputes that continued for three years following its publication. Rousseau took advantage of this continuing debate to reveal progressively his ideas. He criticized the system of a mercenary army and advocated the organization of a people's militia. Not satisfied with merely denouncing luxury, he began to search for its causes in the existing social structures. In his *Observations* (1751) he noted:

> The primary source of evil is inequality; inequality has made possible the accumulation of wealth. The words rich and poor are only relative terms; wherever men are equal, there will be neither rich nor poor. Wealth inevitably leads to luxury and idleness; luxury permits a cultivation of the arts, and idleness that of the sciences.

During this period Rousseau pursued his work as a musician and a theorist of music. He wrote all the articles pertaining to music for the *Encyclopédie,* edited by Diderot and Jean d'Alembert, and later published them in his own *Dictionnaire de musique.* In 1752 he composed an opera, *Le Devin du village (The Cunning-Man, 1766),* which was first performed before the royal court at Fontainebleau. It was a great success; Rousseau was to be presented to Louis XV the next day to receive a pension, but he refused to go. The favourable reaction to this work contributed to the increasing popularity of Italian music, which Rousseau described as more natural and more passionate than French music, and earned him the opposition of Rameau and the partisans of French music. Defending himself against the accusation that his music and writing were in contradiction to the message of his *Discours,* Rousseau wrote in his *Prkface de Narcisse* (1752) that because he was not able to suppress the poison of the arts and literature, he was serving as an antidote to the poison, directing the weapons of corruption back on corruption itself. To prove the sincerity of his words, he sold all his valuables, stepped down from his position as cashier at Dupin's bank, and from then on earned his living by copying music; he called this action his "great reform." As a result of his criticism of French music and of tyrannical government, Rousseau was placed under police surveillance, beginning in 1753.

Political thought

In the same year he ventured to reveal his true political thoughts in response to questions raised by the Dijon Academy concerning the origin of inequality among men and whether it is authorized by natural law. In his *Discours sur I'origine et les fondements de l'inégalité parmi les hommes* (1755), he gave a hypothetical description of man's natural state, proposing that, although unequally gifted by nature, men at one time were in fact equal: they lived isolated from one another and were subordinated to no one; they avoided each other as wild animals do. According to Rousseau, geological cataclysms brought men together for the "golden age" described in various myths, an age of primitive communal living in which men learned good together with evil in the pleasures of love, friendship, songs, and dances and in the pains of jealousy, hate, and war. The discovery of iron and wheat initiated the third stage of human evolution by creating the need for private property:

> It is iron and wheat that have civilized men and ruined the human race . . . . From the cultivation of land, its division necessarily followed. . . . When inheritances increased in number and extent to the point of covering the entire earth and of bordering one on the other, some of them had to be enlarged at the expense of others. . . . Nascent society gave way to the most horrible state of war.

He further proposed that this warring state forced wealthy landowners to resort to a system of laws that they imposed to protect their property.

Rousseau did not pretend that man ought to recapture the primitive equality of his natural state, for he knew that a return to the past was impossible; but, in an article commissioned by Diderot for the *Encyclopédie* and published separately in 1755 as ***Le Citoyen: Ou Discours sur l'économie politique,*** Rousseau sought ways to minimize the injustices that result from social inequality. He recommended three ways: first, equality in political rights and duties, or the respect for a "general will" according to which the private will of the wealthy would not impinge upon the freedom or the life of anyone; second, public education for all children grounded in devotion for one's country and in moral austerity according to the model of ancient Sparta; third, an economic and financial system combining the resources of public property with taxes on inheritances and luxuries.

In the summer of 1754, Rousseau made a short visit to Geneva, where he was well received. He once again became a Calvinist and recovered his citizen rights. In Paris he continued for a while to support the Encyclopédistes and Philosophes in their struggle against the forces of the church and the Parlement (the "superior" court of judicature under the *ancien régime*). He gradually, however, detached himself from them. Though he did not return to the traditional Catholic faith, he still did not wish to join his friends in their criticism of all religions. Rousseau elaborated a "theism" of the heart. According to him, only the rich could afford the luxury of atheism, whereas the poor needed the solace that could be found only in popular songs and religion. Moreover, the Encyclopédistes, having common interests and goals with the bourgeoisie and the rising class of technicians and scientists, were trying to increase productivity rather than to relieve the miseries of the people.

The novel *Julie*

Seclusion at Montmorency. Rousseau became increasingly exasperated with life in Paris and wished to escape it. Mme Louise-Florence-PCtronilleTardieu d'Esclavelles d'Épinay, daughter of a *fermier général* and a friend of the Encyclopédistes, offered him the use of l'Ermitage, a country house near Montmorency. He moved there in 1756 to devote himself to longer works, especially a treatise he planned to call ***Institutions politiques*** and another to be called *Matérialisme du sage,* in which he hoped to insert a spiritual goal into the materialistic methods of education. His projects, however, were sidetracked. In the forest of Montmorency, Rousseau dreamed of love and conjured up an outline for a novel based on the correspondence between two young lovers. By accident he met the Comtesse d'Houdetot, sister-in-law of Mme d'Épinay. Their relationship went no further than a kiss by moonlight, but his passion gave an unexpected turn to the plans for his novel, which eventually became ***Julie: Ou La Nouvelle Héloïse*** and had a broader moral, philosophical, religious, and even economic context than he had first conceived. Out of this passion there arose misunderstandings with Mme d'Épinay, which quickly, other causes intervening, led Rousseau to break with her and her lover, Baron de Grimm, a former friend and an important Franco-German literary critic. He left l'Ermitage in December 1757 and settled in Montlouis, another house near Montmorency, which belonged to the Maréchal de Luxembourg, one of the most devoted of Rousseau's powerful friends. He stayed at Montlouis until June 1762; there he completed ***La Nouvelle Héloïse*** which was published in 1761.

Another incident caused Rousseau to break completely from Diderot and the Encyclopédistes. Jean le Rond d'Alembert, at Voltaire's instigation, had written the article "Genève" for the *Encyclopédie,* in which he asserted that the clergymen of Geneva were "Socinians" (referring to the anti-trinitarian followers of two 16th-century Italians known as Socinus) and that they no longer believed in Christ's Resurrection. Furthermore, he proposed that a theatre be set up in Geneva to serve as a propaganda centre for philosophical ideas. Rousseau secretly prepared a reply to this article, the ***Lettre à d'Alembert sur les spectacles,*** which was published in 1758. In it Rousseau showed that the French classical theatre from Molière to Voltaire had been the prisoner of the aristocratic conception of social and cultural life. It did not "purge" evil passions and did not "chasten" morals; on the contrary, it led to idleness and corruption, and the people of Geneva should distrust it if they wished to defend their liberty. In the preface to the letter, Rousseau responded to Diderot, who had written that "only the wicked one lives alone," alluding to Rousseau's seclusion at Montmorency. Rousseau rebutted that social and worldly life was the source of wickedness. Diderot considered the letter a desertion, a betrayal at a time when Rousseau should have been defending the ***Encyclopkdie.***

Émile and ***Du contrat social***

Still at Montlouis, Rousseau began working on his two great projects, which both appeared in 1762. The ***Matérialisme du sage*** became a long reverie about education and acquired the title ***Émile: Ou de l'kducation.*** It has appeared to many that Rousseau was trying to make amends for abandoning his own children by helping other parents raise their children properly. He advised them to shed their social prejudices and to "follow nature." Mothers should breast-feed and should strengthen the bodies of their children by means of severe tests of physical strength and endurance. His method of education was to slow down intellectual growth: it called for the child to demonstrate his own interest in a subject and ask his own questions; at the stage of puberty, however, the sensitivity of the youth should be educated. The adolescent would hopefully accept a free and reciprocal "contract" of friendship with his teacher, who could then help him uncover the joys of religion and the difficulties of coping with society. Émile, who laboriously acquired a sense of property while cultivating his garden, discovered the hard life of a worker when he became apprenticed to a carpenter. Then, through Sophie, Émile discovered the nature of love. He had to leave her, however, to complete his political education, which required looking through the world for the country that would best suit his future family.

In ***Du contrat social*** Rousseau developed the political principles that are summarized at the conclusion of ***Émile.*** Starting with inequality as an irreversible fact, Rousseau tried to answer the question of what compels one man to obey another man or by what right does one man exert authority over another. He concluded that only a contract tacitly and freely accepted by all allows each one "to bind himself to all while retaining his free will." Freedom is inherent in freely accepted law: "To follow one's impulse is slavery but to obey the self-prescribed law is liberty." In order to prevent this law from profiting solely the rich, the condition in which each one has something and no one has too much must exist. The conclusion derived from these principles is that the people alone are sovereign and that they exercise their sovereignty through a government that may be revoked at any time. He stressed the necessity of adapting the forms of government to existing historical and geographical conditions. After analyzing the workings of the Roman Republic, he concluded that "civil religion" is a political necessity. Such a religion consists of two dogmas: first, tolerance of all religious opinions; second, recognition that Providence will reward the good citizens and punish the bad. This religion lays a firm foundation for total commitment to the collectivity.

Attacks on his works and his way of life

Exile in Switzerland and England. *Émile* and *Du contrat social* were condemned by the Parlement of Paris in June 1762 as contrary to the government and to religion. Rousseau had to flee to Switzerland, but there, too, he and his works were banned. He defended himself in his *Lettre à Christophe de Beaumont* (1763), an attack on the Archbishop of Paris, who had condemned *Emile,* and in his *Lettres écrites de la montagne* (1764; "Letters Written from the Mountain"), a reply to J.-B. Tronchin, procurator general of the Genevan Republic, who had written in defense of the executive council of Geneva for having ordered the burning of *Émile* and of *Du contrat social.* In September 1764 Rousseau was asked by Matteo Buttafoco, a friend of Pasquale Paoli, leader of the Corsican nationalists, to prepare a constitution for Corsica. He never completed the task, though he did make a rough draft. On the last day of 1764 he received an anonymous pamphlet, *Le Sentiment des citoyens* ("The Feeling of the Citizens"), attacking him savagely as a hypocrite, a heartless father, and an ungrateful friend. It was written by Voltaire, and its effect on Rousseau was terrible. When he recovered from the shock, he decided to write his autobiography, the *Confessions.*

The *Lettres écrites de la montagne* had turned the Protestant ministers even more strongly against him, and in September 1765, after his house had been vandalized, he decided to leave Môtiers in the territory of Neuchâtel, where he had been living. He fled to the Île Saint-Pierre on Bernese territory, and then, through the help of François-Louis de Bourbon, the prince of Conti, obtained a special passport allowing him to go to England, where he arrived in January 1766. David Hume, a British Empiricist, took him under his patronage, but Rousseau, whose reasoning powers had been affected by his tribulations, began to suspect that Hume had been won over by the Parisian Philosophes into a plot to ruin his name. This fear and Hume's excessive zeal to justify himself publicly at Rousseau's expense led to an open quarrel between them, which excited and amused all cultivated Europe. In May 1767 Rousseau fled in panic to France.

Last years in France. Rousseau took the name Renou and moved to the Château de Trye, near Gisors, lent to him by the Prince of Conti. While he was at Trye his *Dictionnaire de musique,* which he had been working on for years, was at last published. He left Trye abruptly in June 1768, again in a panic, going first to Bourgoin near Lyons, where he married Thérèse Levasseur, and then to Monquin. During this period he wrote his *Confessions.* In 1770 he moved to Paris to defend himself against the "conspirators." There he resumed his own name but was left unmolested. To justify himself to the world, he read extracts from his *Confessions* in various Parisian salons until, at the request of Mme d'Épinay, he was ordered to desist by the police lieutenant of Paris.

In 1771 he was asked by the Confederation of Bar—noble Polish nationalists—to advise them how the Poles should reform their institutions, and he wrote *Considérations sur le gouvernement de Pologne.* Still eager to justify himself, he wrote the *Rousseau juge de Jean-Jacques: Dialogues.* He tried in December 1775 to place this work under God's protection on the high altar of the Cathedral of Notre-Dame but was prevented from doing so by the iron grille surrounding the choir. It seemed to him that God had joined his persecutors, and for a time he was in despair. During the last two years of his life his madness weighed less heavily on him; he led a secluded life with Thérèse, accepted some young visitors, and wrote the most serene and delicate of his works, *Les Rêveries du promeneur solitaire* ("Reveries of a Solitary Walker"), which contains descriptions of nature and of man's feelings for nature. In May 1778 he moved to Ermenonville, to a pavilion on the estate of the marquis René de Girardin, where he died about six weeks later, on July 2. He was buried on the Île des Peupliers in the lake at Ermenonville, but his remains were moved to Paris to the Panthéon during the Revolution.

MAJOR WORKS

NOVELS: *Lettres de deux amants habitants d'une petite ville au pied des Alpes, reccreillies et publiées par J.J. Rousseau,* 6 vol. (1761, in subsequent editions called *La Nouvelle Héloïse; Eloisa;* or *A Series of Original Letters Collected and Published by J.J. Rousseau,* trans. by William Kenrick, 4 vol., 1761); *Emile: Ou de l'dducation, 4* vol. (1762; *Emilius and Sophia: Or, A New System of Education,* trans. by William Kenrick, 4 vol., 1762–63).

AUTOBIOGRAPHICAL WORKS: *Rousseau juge de Jean-Jacques. Dialogues . . . premier dialogue. D'après le* manuscrit de *M. Rousseau, laissé entre les mains de M. Brooke Boothby* and edited by him (1780); *Les Confessions, suivies des rêveries du promeneur solitaire,* 2 vol. (1782; *The Confessions of J.J. Rousseau: with the Reveries of the Solitary Walker,* trans. from the French, 1783; *Confessions, etc.,* Everyman Library, 2 vol., 1960).

ESSAYS: *Discours qui a remporte' le prix à l'Académie de Dijon en l'année 1750. Sur cette question . . . si le rétablissement de sciences et des arts a contribué a épurer les moeurs; Par un citoyen de Genève* (1750; *A Discourse, to Which the Prize Was Adjudged by the Academy of Dijon . . . on This Question . . . Whether the Re-establishment of Arts and Sciences Has Contributed to Purify Our Morals,* trans. from the French by R. Wynne, 1752); *Discours sur l'origine et les fondaments de l'inégalité parmi les hommes* (1755; *A Discourse upon the Origin and Foundation of the Inequality Among Mankind,* 1761); *Emile for Today: The Emile of Jean Jacques Rousseau,* selected, trans., and interpreted by William Boyd (1956); *Le Vicaire savoyard, tiré du livre intituld Emile* (1765; *Rousseau's Admissions Concerning the Gospel and the Character of Christ,* being an extract from Rousseau's *Emile,* 1830); *Du contrat social: Ou, principes du droit politique* (1762; *A Treatise on the Social Compact: Or, The Principles of Political Law,* trans. from the French, 1764); *Considérations sur le gouvernement de Pologne, et sur sa rdformation projettke* (written 1771, published 1782); for further examples of Rousseau's political works, see *The Political Writing of Jean Jacques Rousseau,* ed. by C.E. Vaughan, 2 vol. (1962).

LETTERS: *Lettre sur la musique française* (1753); *J.J. Rousseau . . . à M. d'Alembert . . . sur son article Genève dans le VII^e^ volume de l'Encyclopédie, et particulièrement, sur le projet d'e'tablir un théâtre de come'die dans cette ville* (1758; *Politics and the Arts: Letter to M. d'Alembert on the Theatre,* trans. and ed. by A. Bloom, 1960); *Lettres écrites de la montagne* (1764); for further examples of Rousseau's voluminous correspondence, see the *Catalogue de la correspondence de J.-J. Rousseau, lettres expédiées et reçues, conservée à la bibliothèque de la ville de Neuchâtel* (1963).

MUSICAL WORKS: *Dictionnaire de musique* (1768; *A Dictionary of Music . . .*, trans. by William Waring, 1779?).

BIBLIOGRAPHY

Bibliography: JEAN SENELIER, *Bibliographie générale des oeuvres de J.J. Rousseau* (1949); *Annales de la Société Jean-Jacques Rousseau* (1905–71).

Works: Oeuvres complètes de Jean-Jacques Rousseau, "Bibliothèque de la Pléiade," 4 vol. (1959–69); *Correspondance complète de Jean-Jacques Rousseau,* ed. *by* R.A. LEIGH, 16 vol. (1965–72).

Biography: JEAN GUEHENNO, *Jean-Jacques Rousseau,* new ed., 2 vol. (1962; Eng. trans., 1966), a vivid, accurate biography; F.C. GREEN, *Jean-Jacques Rousseau: A Critical Study of His Life and Writings* (1955), a good study of the relationship between Rousseau's life and writings; LOUIS JOHN COURTOIS, *Chronologie critique de la vie et des oeuvres de Jean-Jacques Rousseau* (1923–24).

Political thought: ROGER D. MASTERS, *The Political Philosophy of Rousseau* (1968); MICHEL LAUNAY, *Jean-Jacques Rousseau, écrivain politique* (1972), the first chronological bibliography of the writings of Rousseau, showing the relationship between his politics and literature.

Religion: PIERRE M. MASSON, *La Religion de Jean-Jacques Rousseau, 3* vol. (1916), a valuable work.

Philosophy: ERNST CASSIRER, "Das Problem Jean-Jacques Rousseau," in *Archiv fur Geschichte der Philosophie* (1932; Eng. trans., *The Question of Jean Jacques Rousseau,* 1954, reprinted 1963), a very clear and prestigious Kantian interpretation; PIERRE BURGELIN, *La Philosophie de l'existence de J.-J. Rousseau* (1952).

Literature: JEAN STAROBINSKI, *Jean-Jacques Rousseau: La Transparence et l'obstacle suivi de 7 essais sur Rousseau,* 2nd ed. rev. (1971), the best psychological and literary interpretation of the thought of Rousseau.

(M.C.L.)

Rubber

Rubber is an organic substance—obtained from natural sources or synthesized artificially—which has the desirable properties of extensibility, stretchability, and toughness. Previously known as caoutchouc, a term that has become limited to the chemically pure form of the substance, rubber acquired its name from its ability to erase pencil marks.

Initially, all rubber was natural. Formed in a living tree, it consisted of the solids separated (coagulated) from the milky fluid latex, which occurs under the bark of many tropical and subtropical trees and shrubs. Today, more than half the rubber on the market, apart from one specialty lubber, is manufactured synthetically by means of chemical processes that were partly known in the 19th century but had not been applied successfully until World War II.

NATURALRUBBER

History. On his second voyage to the New World in 1493–96, Christopher Columbus saw Indians in Haiti play a game with balls made from the gum of a tree. The first mention of rubber being used for purposes other than sport was made in 1615, when a Spaniard related how the Indians, having gathered the milk from incisions made in various trees, brushed it onto their cloaks and obtained crude footwear and bottles by coating earthen molds and allowing them to dry.

The first serious accounts of rubber production and the primitive native system of manufacture were given in the 18th century when a member of a French geographical expedition sent to South America in 1735 described caoutchouc as the condensed juice of the *Hevea* tree. The first accurate account of the botanical characteristics of the *Hevea* species was given in 1775.

At the outset, not chemical knowledge but the ability to devise a suitable method of rubber manipulation was of paramount importance to the development of the rubber industry. Two Frenchmen, a physician and a chemist, attempted to discover a cheap and effective solvent for crude rubber brought from South America and introduced the use of turpentine and ether as solvents.

Invention of the "mackintosh"

Macintosh and Hancock. Important progress came at the beginning of the 19th century from the separate experiments of a Scottish chemist, Charles Macintosh (1766–1843), and an English inventor, Thomas Hancock (1786–1865). Macintosh's contribution was the rediscovdry, about 1820, of coal-tar naphtha as a cheap and effective solvent. His experiments were successful; he placed a solution of rubber and naphtha between two fabrics and in so doing avoided the sticky or brittle surfaces that had been common in earlier single-texture garments treated with rubber. Manufacture of these double-textured, waterproof cloaks, henceforth known as "mackintoshes," began soon afterward.

The work of Hancock, who became Macintosh's colleague and partner, is of even greater importance. He first attempted to dissolve the rubber in turpentine, but his hand-coated fabrics were unsatisfactory in surface texture and smell. He then turned to the production of elastic thread. Strips of rubber were cut from the imported lumps and applied in their crude state to clothing and footwear. In 1820, in an effort to find use for his waste cuttings, Hancock invented a masticator. Constructed of a hollow wooden cylinder equipped with teeth in which a hand-driven, spiked roller was turned, this tiny machine, originally taking a charge of two ounces of rubber, fulfilled Hancock's greatest hopes. Instead of tearing the rubber to shreds, it produced enough friction to weld the scraps of rubber into a coherent mass that could be applied in further manufacture.

Goodyear and vulcanization. Early manufactured rubber softened with heat and hardened with cold (particularly annoying in the U.S., where climate was more extreme than in England). It also was tacky, odorous, and perishable. These fundamental weaknesses were removed by the invention of vulcanization in 1839 by the U.S. inventor Charles Goodyear (1800–60). Developing a compound of rubber, lead, and sulfur and a heat treatment (or curing) process, Goodyear created a product, at first called fireproof gum, afterward vulcanized rubber, that exhibited impressive durability.

Vulcanization made the modem rubber industry possible by permitting use of the substance in combination with machinery and in tires for bicycles, and later for automobiles. Though subsequent discoveries have refined Goodyear's original techniques, the vulcanization process remains fundamentally the same as it was in his day.

Founding of the tire industry

Rise of the rubber industry. Most European countries had established rubber industries by the decade 1820–30. Vulcanization and the increased use of steam and electric power brought a large demand for rubber. A pneumatic wheel was patented as early as 1845. The invention was ahead of its time, and it was left to John B. Dunlop, a British veterinary surgeon, to found the tire industry by patenting and developing pneumatic tires for bicycles and tricycles in 1888. His work had a great influence on popularizing cycling, both as a hobby and as a means of transport. At the end of the century some of the largest tire companies were commencing operations manufacturing solid rubber tires for horse-drawn carriages.

In 1876 the Englishman Sir Henry Wickham collected rubber seeds from the wild trees of the Amazon jungles and had them planted in Kew Gardens. The young trees were subsequently transferred to Ceylon and the Malay Peninsula where they formed the base of the natural rubber plantation industry that produces nearly 3,000,000 tons a year.

Rubber and the automotive industry. The first decade of the 20th century saw the establishment of the motorcar in Europe and the U.S. The dependence of the automotive industry for its tires, as well as many other components, on the drops of latex exuding from trees halfway across the world seems extraordinary in retrospect. Yet, despite the rapid expansion of the auto industry, especially in the U.S., it remained entirely dependent on natural rubber until about 1940. The change from steel wheels to rubber tires on agricultural tractors in the early 1930s added still more to the demand for rubber in the years before World War II.

Development of a synthetic rubber industry

When World War II began, Germany and the Soviet Union were the only two countries with the means of making synthetic rubber; all the other combatants were dependent upon natural rubber from southern Asia. After Japan entered the war, Asian sources, except for Ceylon, were cut off. The U.S., faced with necessity, developed a synthetic rubber industry almost overnight, achieving a production of 800,000 tons per year. At the war's end, with natural rubber again available, the U.S. synthetic rubber industry went into a sharp decline. By the early 1950s, however, a superior and more uniform synthetic became available. This new synthetic rubber, exported to the British and other European markets, stimulated development of a synthetic rubber industry in England. Synthetic rubber equalled natural in world volume in the early 1960s and moved ahead thereafter, as indicated in Table 1.

Table 1: World Consumption of Natural and Synthetic Rubber in Selected Years (long tons)

year	natural rubber consumption	synthetic consumption*
1955	1,890,000	1,063,000
1964	2,260,000	2,748,000
1970	2,888,000	4,530,000†
1972	3,120,000	5,365,000

*All types. †Estimated.

Since World War II, rubber—natural or synthetic—has retained its place as one of the world's most important commodities. While vlastics have made some inroads into markets previously monopolized by rubber, new uses and expansion in traditional applications have brought steady expansion of the industry.

The rubber tree. (For a detailed discussion of the botany of the rubber plant, see EUPHORBIALES.) Commercially, natural rubber is obtained almost exclusively from Hevea brasiliensis, a tree indigenous to South America, where it grows wild to a height of 120 feet (36 metres). Cultivated in plantations, however, the tree grows only to about 80 feet (24 metres) because carbon, necessary for growth, is also an essential constituent of latex. Since only atmospheric carbon dioxide can supply carbon to the plant, that element has to be rationed between the two needs when the tree is in active production. And with foliage limited to the top of the tree (to facilitate tapping), the intake of carbon dioxide is less than in a wild tree. Because other latex-producing trees do not compare with Hevea brasiliensis for efficiency, industry botanists have concentrated their efforts exclusively upon this species.

Other natural sources of rubber

Because of their historical importance, however, other natural sources of rubber of considerable value in presynthetic days deserve brief mention. *Castilla,* a plant of the family Moraceae of the order Urticales (elm), yielded a large quantity of rubber at one tapping, then required a rest for several months. Manihot, a genus of the family Euphorbiaceae, though best known as the source of tapioca, provides another variety of latex. Outside Brazil, producers have constantly searched for alternate sources of rubber. The watery content of many plants contains some rubber; two of importance are guayule (Partheniurn argentaturn), a shrub grown mainly on the U.S.–Mexican border, its chief drawback being its high gum content (up to 50 percent); and the Russian dandelion (kok-saghyz), yielding a latex with a 30 percent dry-rubber content. Both were exploited during World War II.

Modem production. Hevea brasiliensis has long been cultivated. Seeds are carefully selected, often from widely scattered areas to avoid inbreeding. Widespread planting provides a greater selection of young plants with elimination of weak growths. In the scientific layout of plantations, the natural contours of the land are followed and the trees protected from wind. Precautions against disease include inoculation. Cover crops planted adjacent to the rubber trees hold rainwater on sloping ground and help fertilize the soil by fixing atmospheric nitrogen. Standard horticultural techniques, such as nursery growing of hardy rootstocks and grafting on top of them, hand pollination, and vegetative propagation (cloning) to produce a genetically uniform product, are also employed.

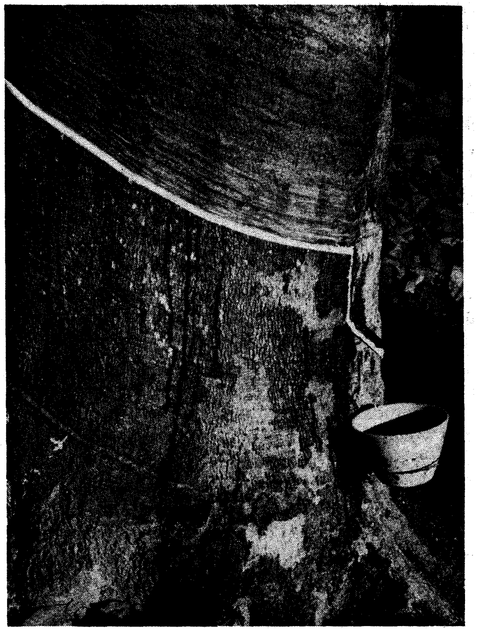
Plantations. Rubber trees grow only within a well-defined area of the tropics—about ten degrees north or south of the Equator—and in most types of soil. Heavy annual rainfall of about 100 inches (250 centimetres) is essential, with emphasis on a wet spring. In consequence, growing areas are limited. Southeast Asia is particularly well situated for rubber culture; so are Liberia and Nigeria in West Africa. Malaysia is the biggest producing area and is considered the centre of the world industry because of the concentration of commercial and technical interests in the capital city of Kuala Lumpur.

Malaysian rubber production

The plantations of Malaysia achieved prominence about 1900; from that time their products steadily replaced the wild rubber of Brazil. Yield initially was about 300 pounds of rubber per acre per year. By effective management, research, and the use of chemical stimulants, the average has been increased to 1,500 pounds (680 kilograms) per acre. An average of 3,000 pounds (1,360 kilograms) is anticipated in the near future; eventually, scientists believe, 6,000 pounds (2,720 kilograms) will be possible. About one ounce (28 grams) of rubber is obtained per tree per tapping before the latex coagulates and flow stops. If this coagulation can be prevented, the latex continues to flow; a chemical, Ethrel, applied to the bark has been shown to be capable of doing this. In consequence, the small collection cup that was emptied every day or two is being replaced by larger polyethylene plastic bags that require attention only every 10 to 14 days.

Tapping. The standard method of tapping a rubber tree for latex is to score it with a knife to a depth of one millimetre for half the circumference of the trunk, slanting the cut down from the left to right at an angle of 30 degrees starting at the highest point convenient to the tapper; each subsequent cut is made immediately below its predecessor. Trees are often rested for a period after heavy tapping. Production commences when a tree is from five to seven years old; with care, the tree's life may extend to more than 30 years.

By courtesy of UniRoyal, Inc., New York



Trunk of a rubber tree showing latex flowing from tapping cut into the collection cup on a plantation in Malaysia.

The watery latex from the scored tree oozes out from between the outside cork layer of the bark and the cambium inner layer, where new cells are formed. Although the function of latex in maintenance of the trees has not been explained satisfactorily, a great deal is known of its composition. Three types of particles are suspended in the watery medium: one rubber and two nonrubber. Of the two latter, the more abundant is the neutral gray particle known as a lutoid, and the other, more highly coloured and named after its discoverer, is called the "Frey Wyssling" particle.

Treating slabs of rubber

Coagulation. After collection, the latex is sieved to remove foreign objects; then water is added. Next, the rubber is coagulated from the latex by the addition of dilute acetic or formic acid and is deposited on aluminum partitions in horizontal tanks having aluminum partitions that allow the rubber to be removed in slabs rather than in a single lump. When necessary, the slabs are washed by passing them between horizontal rollers, moving at different speeds and flushed with a copious supply of water. Excess water is removed by passing the rubber through a succession of rollers. Subsequent treatment depends upon the type and grade of rubber desired.

Wild rubber was originally recovered by heating latex over a fire while stirring it with a stick. The water was removed by evaporation, leaving behind a solid ball of rubber. This method was not welcomed by the market, however, because the solid balls of rubber were hard to store and, in the early days, because of the fear that stones, metals, and other dense objects would be included to falsify weight.

Rubber types and grades. Rubber may be processed into pale crepe, smoked sheet, or skim rubber. The various ways of grading rubber are subject to controversy.

Pale crepe. After emerging from the coagulation tanks, the wet slabs of rubber are run through a succession of creping mills that roughen the rubber's surface and gradually reduce its thickness to facilitate drying. The sheets are then laminated to the desired thickness and allowed to air-dry. Pale crepe, which varies in colour from off-white to deep cream, is used wherever colour is an important factor. Impurities in the rubber give it its colour, which can be controlled by regulating the strength of the acid when the first rubber is deposited by coagulation; the stronger the acid, the more impurities will be precipitated. The lighter coloured rubber can be sold at a premium. A bleaching process is also used. In general, pale crepe is used only where colour is decisively important, as some of its physical properties (tensile strength, for example) are inferior to those of sheet.

Smoked sheet. Slab rubber can sometimes be sheeted out—that is, run through sets of plain rolls until the last set, which prints a simple pattern, such as crisscross. The sheet is then hung up to dry in smoke from burning wood or oil to obtain the "smoked sheet" of commerce. The best quality has an amber or brown shade.

Skim rubber. Normal coagulation methods leave ten percent of the rubber behind in the liquid of the coagulation tank. That residual rubber can be recovered by skimming, but it contains a much higher proportion of impurities and so is less desirable.

Pale crepe, smoked sheet, and skim rubbers are generally made up into bales, most frequently weighing *225* or *250* pounds *(102* or *113* kilograms) and bulking as a cube about two feet *(60* centimetres) to a side. The rubber is coated with clay or similar material to prevent sticking, and the bales are bound with metal strapping.

International grading standards. The three types of rubber described above are graded according to international standards published by the Rubber Manufacturers Association, Inc. of New York. Blending different grades of the same type or of entirely different types of rubber is quite a common practice of manufacturers to obtain the best properties for some specific purpose. The user also may blend bales from different deliveries of the same type of rubber to obtain a more uniform product. It is uneconomical to wash and strain low grades to raise their quality.

Grading has traditionally relied upon visual inspection and, for that reason, has received serious criticism over the years. Consequently, efforts have been made to find a more scientific basis for grading. In *1950,* the Technically Classified Rubber system was devised, employing nine grades based on the physical properties, viscosity and rate of cure. But because viscosity of natural rubber varies with age, industry would not accept the system. Furthermore, a premium was charged for grading; and the plan was initially confined to top-grade rubber. By the time it was extended to the lower grades, it was too late to halt declining interest. Rubber-growing interests always regard top grade as standard and others as off-grades, whereas important sections of the manufacturing industry, such as tire makers, purchase the lower grades.

The Technically Classified Rubber system

In *1965* Standard Malaysian Rubber, a much more ambitious and practical system, based on technical specifications and defined packing requirements, came into use. With some later modifications and extensions, the fundamental principle has taken hold, and rapidly increasing quantities of rubber have been marketed under this system, amounting in *1970* to about *250,000* tons or *10* percent of the total natural rubber produced (more than *20* percent of the Malaysian output). Grading is based on chemical analysis and controlled by the Rubber Research Institute of Malaysia.

Fundamentally, classification is by dirt content, but other chemical standards are imposed. For each grade there are maximum amounts allowable for various impurities. The Plasticity Retention Index has proved the most controversial quality criterion of the Standard Malaysian Rubber system. That index measures the chemical degradation of the rubber over time; the better the rubber, the less it should break down and, consequently, the higher the Plasticity Retention Index. Though the originators of the Standard system now feel that sufficient experience exists to accept the Index as an essential part of their standard, major rubber growers disagree. They say they have been able to provide first-class rubber that fails this test and samples of poor rubber that pass. Nevertheless, the Standard system has been a great advance over visual inspection and has formed the cornerstone of technical rubber grading.

Chemical and physical properties. Rubber has the empirical formula C_5H_8, indicating a molecule in its simplest form with five carbon and eight hydrogen atoms. Many of these simple molecules commonly combine together, producing a structure whose atomic weight varies within wide limits, depending upon conditions. The composition of raw rubber has been chemically determined both by breaking down the molecule into its simplest form, or C_5H_8 monomer, and by the reverse process of building up rubberlike materials from monomers (polymerization).

Rubber has certain peculiar and useful properties of which the most outstanding is its ability to stretch seven to eight times its original length. During extension, rubber gives out heat; as it contracts, it absorbs heat. The extent to which a piece of rubber can contract after being extended or stretched depends on temperature.

Rubber in solvents

The behaviour of rubber in certain solvents is peculiar. In most organic solvents it swells, part going into solution readily (the sol portion) and part resisting (the gel portion). Mastication of rubber encourages its diffusion in the solvent appreciably, producing "rubber solution," the viscosity of which is substantially reduced by further mastication or by the addition of small amounts of other liquids.

When defining the properties of rubber, care must be taken to distinguish between the chemically pure substance and the commercial product. Commercial latex contains a significant proportion of nonrubber material, much of which remains in the coagulated solid. Dried rubber typically contains *92* to *94* percent chemically pure C_5H_8, *0.13* to *1.20* percent water, *2.50* to *3.20* percent acetone extract, *2.50* to *3.50* percent nitrogenous substances, and *0.15* to *0.50* percent ash.

The nonrubber substances in commercial rubber affect its properties to a greater extent than their proportion would suggest. A high moisture content causes mold to form, damaging market value where appraisal is based on appearance, even though usefulness is not impaired.

The acetone extract contains unsaturated fats and fatty acids, which have a softening effect and also accelerate vulcanization even when present in quite small proportions. The yellow colouring material of pale crepe is found to some extent in the acetone extract. Antioxidants occur in minute amounts but have a very beneficial effect on the aging qualities. Proteins, present in varying degrees, have little effect except upon vulcanization, which they accelerate. Complicated alcohols and other organic substances occur in small proportions.

The ash found in latex contains all the mineral matter and varies in quantity and nature according to circumstances. A certain amount of sand and soil is inevitable. More harmful are the traces of copper and manganese, which act as catalysts for chemical reactions that result in the degradation of rubber. Their action depends upon their state of combination. For example, Cambodian rubbers are always high in manganese, but because the manganese is combined organically it has no deleterious effect.

Physical tests of rubber

Chemical examination of rubber is a tedious operation requiring great skill. Hence, it is usually done only where absolutely required as, for example, in arbitration disputes. Of greater practical value and far easier to perform is the physical testing of commercial rubber. It is carried out on vulcanized compounds and varies in type according to the proposed use of the product (*i.e.*, resistance to or conduction of electric current, resistance to water, acids, and other chemicals). Two important physical tests involve measurement of modulus of elasticity and tensile strength. Modulus of elasticity, defined as the response of rubber to a force applied in a given direction, is generally measured by stretching a piece of rubber of known cross section, the stress being the force applied and the strain, the extension, which, of course, will vary with temperature. Modulus may be measured at *100,200,300, 500,* or *600* percent extension or at any other appropriate elongation. The result is recorded in kilograms per square centimetre or pounds per square inch. Modulus is normally a function of the state of cure, and as such is considered more important than tensile strength, which is the ultimate stress required to break a standard specimen, again expressed in the same units. Elongation at break is usually measured at the same time.

One other physical property of cured rubber is frequently quoted because of its simplicity, namely, hardness. It is generally measured by pressing the point of a cone-shaped instrument called an indenter into the rubber and measuring the depth of penetration on an International Rubber Hardness scale.

For specific uses, special laboratory tests have been devised. For example, tire-tread compounds are tested for abrasion and skid resistance. Dynamic testing (*i.e.*, with the rubber in motion) attempts to simulate service conditions; but until it is accepted more widely, full-scale service testing, which is expensive and time-consuming, will continue.

Distinctive uses of natural rubber. Though natural rubber performs well in most applications, some of the newer synthetics surpass it for specialized purposes. For example, acrylonitrile rubber has much better oil resistance, butyl rubber is much less permeable to air, and silicone rubber withstands higher temperatures.

Natural rubber is still preferred in applications that demand elasticity, resilience, tackiness, and low heat buildup. Natural rubber also is chosen when it has no particular technical advantage over synthetics but is competitive for price or availability. Natural rubber is indispensable for the treads of tires for racing cars, bus and truck tires, airplane tires, and any other form of transport where conditions are particularly severe and where a buildup of heat inside the tire could cause a failure. Broadly speaking, the larger the tire, the greater the proportion of natural rubber used in its construction, to minimize heat buildup.

The resilience of rubber is utilized in engine mounts and suspension units of automobiles, and its ability to withstand severe loads for long periods in compression has led to its use in building foundations. These properties, in addition to good cold-temperature resistance, also make it useful in bridge bearings.

Distinctive uses of natural rubber latex. Latex represents rubber in its most tractable form. Criticism has long been made of the time and effort spent at plantations in removing water from latex to obtain solid rubber and the subsequent effort made at its destination to render the rubber processable. Dried latex, in fact, has some better physical properties than conventional solid rubber, but the cost of transporting a material containing 60 percent water across the world would be excessive. Concentration of latex to a content of 60 to 70 percent dry rubber is performed either at the source or in the user country by means of centrifuging, evaporation, or a process known as creaming, in which an agent is added to the latex that causes the rubber particles to swell and rise to the surface of the liquid.

Creaming as a concentration Process

Latex is an excellent adhesive and is widely used in its natural state, but for most industrial uses it needs additives and vulcanization. Accelerating agents for vulcanizing and other essential fillers are added. One of the oldest uses of latex is in the production of such dipped goods as rubber gloves and prophylactics. The process consists of dipping formers of the correct size and shape into compounded latex, then drying, repeating the operations, and vulcanizing.

Rubber (elastic) thread also is produced from latex. Initially, strips were cut from a thin sheet to form a thread of square cross section. Later, the much superior round thread was obtained by extruding compounded latex into a bath of coagulant through glass nozzles.

Foam rubber has been one of the most important latex products since its discovery in the 1920s. The confinement of thousands of gaseous bubbles in natural rubber cells provides ideal resilience. In the simple manufacturing process, excess ammonia is removed from the latex, and, after a vigorous stirring with dispersions of soap, sodium fluorosilicate is added to the frothy mixture as a gelling agent, along with accelerating agents, zinc oxide, and antioxidant. After the liquid has stood in a mold for several minutes to permit gelling, it is vulcanized. In practice, the process requires the most careful control and the judgment of experienced personnel.

Foam products were made originally in individual molds, which progressed along a conveyor in series and into each of which a predetermined quantity of foam was inserted, the mold closing during its progress to the oven. Washing, drying, and finishing completed the process. For large-scale production, a new method has supplanted this early method, which is still in active service. The process consists of filling a fixed mold with a metered quantity of foam, sealing it with its lid, and evacuating the air; the foam expands to partly fill the resulting vacuum, and then is frozen; next, carbon dioxide gas is passed in to fill the remaining vacuum in the foam, and the temperature is raised to 104" C (219" F) for curing.

The carpet industry has made increasing use of latex not only as a separate foam rubber underlay but as an undercoating on the carpet itself and as an anchoring matrix for tufted carpets. The latex may be either a vulcanizable natural rubber compound or a synthetic self-curing compound.

Foams frequently comprise a blend of natural and styrene-butadiene rubbers, the latter reducing the cost with some reduction in resilience.

SYNTHETIC RUBBER

History. In 1826 Michael Faraday assigned the empirical formula C_5H_8 to rubber. For the next 100 years or so chemists endeavoured to reproduce this molecule synthetically without success. It was only when the search for chemical equivalence was abandoned and adequate physical properties were emphasized that synthetic rubber came into being. The choice fell upon butadiene, C_4H_6, a compound related to natural rubber, to form the basis for synthetic rubber. Though Germany produced synthetic rubber (called methyl rubber) during World War I, it was an inferior substitute by present-day standards. Subsequently, Germany produced Buna rubbers, the first two letters "bu" standing for butadiene, and the last two, "na," for the chemical symbol for sodium, the catalyst used for polymerization.

Buna rubber and its successors

In Germany during World War II, product control remained in the hands of the chemical manufacturer, who gave first priority to quality and ease of production, without much regard for the subsequent difficulties of the user. In the U.S., the rubber producer was also the user; hence a more workable product was developed.

Wartime U.S. synthetic rubber was made from butadiene and styrene. Originally known also as Buna, its name was changed to GR–S (Government Rubber–Styrene) when the government took control. This was the forerunner of modem synthetic rubber. Although ideas for better or cheaper products developed, the government held to the GR–S formula for the duration of the war to avoid complications. Meantime, German chemical engineers perfected low-temperature polymerization, producing a more uniform product. The process, which has come into universal use, is known as redox, coined from the words reduction and oxidation. The two products are often referred to, respectively, as "hot" rubber and "cold" or low-temperature polymer synthetic rubber.

Types of synthetic rubber. In addition to styrene–butadiene, there are many specialized rubbers, sometimes difficult to classify because of the varying systems of nomenclature. Table 2 gives the equivalents that have been used: Some of the newer rubbers—*e.g.*, silicone, polyurethane, and Hypalon—have no official or generally agreed symbols.

The approximate world production in millions of tons of various popular types of rubber is given in Table 3. Adding in the smaller amounts of other special-use rubbers brings the annual total to more than 8,000,000 tons.

The world distribution of synthetic rubber production in 1970 is given in Table 4 as percentage of total production.

Table 2: Types of Synthetic Rubber

common name	original code name	American government code	current code	chemical description
Synthetic (unspecified)	Buna S	GR-S	SBR	styrene–butadiene
Butyl		OR-I	IIR	polyisobutylene–polyisoprene
Butadiene			BR	polybutadiene (solution polymer–stereospecific)
Neoprene (trade name)		GR-M	CR	polychloroprene
Nitrile	Buna N	GR-N (previously GR-A)	NBR	acrylonitrile–butadiene
Polyisoprene			IR	polyisoprene (solution polymer–stereospecific)
Thiokol (trade name)		GR-P		polysulfide
EPR				ethylene–propylene
EPDM				ethylene–propylene terpolymer

Styrene–butadiene rubber. This general purpose synthetic rubber, a polymer of butadiene and styrene, exceeds all others in consumption. Its symbol, SBR, covers a range of materials in which proportions of the two ingredients vary. Most styrene–butadiene lubber is manufactured by the emulsion process, in which the styrene and butadiene are brought together in a water solution of soap that acts to disperse, or emulsify, the materials in the solution. Other materials in the solution include catalysts, which initiate the reaction, and stabilizers, which prevent deterioration of the final product. A "short stop" agent is added to stop the reaction at the optimum point, when about 60 percent of the ingredients are converted to synthetic rubber. A modifier produces a material of the desired plasticity, and a coagulant is added to deposit the rubber from the latex.

Table 3: World Production of Popular Types of Rubber (1970)

	total	
	long tons (000,000)	short tons (000,000)
Natural	2.75	3.08
Styrene–butadiene	3.25	3.64
Butyl	0.30	0.34
Polybutadiene	0.70	0.78
Neoprene	0.35	0.39
Nitrile	0.20	0.22
Polyisoprene	0.25	0.28

Table 4: World Distribution of Synthetic Rubber Production (1970)

	percentage of total production
United States	44.3
Japan	13.8
France	6.3
United Kingdom	6.1
West Germany	6.0
Netherlands, The	4.1
Canada	4.1
Italy	3.0
East Germany	2.3
Brazil	1.5
Poland	1.2
Romania	1.1
Czechoslovakia	1.0
Others	5.2
World total	100.0

Styrene, a liquid, boils at 145° C (293" F), but butadiene, a gas at normal temperatures, boils at —4° C (25° F); butadiene may be stored in liquid form under pressure. Both liquids are pumped continuously into reactors and mixed with water, soap, and catalyst by constant agitation at controlled temperature. After the short stop is added, the unconverted butadiene and styrene are recovered and re-used. Then the antioxidant is added and the coagulant to deposit the rubber.

Varieties of SBR, each with its own properties, can be produced either by altering the ratio of the two ingredients; by using hot or cold polymerization; or by varying other elements in the manufacturing process. For example, with 23.5 percent SBR, fatty acid–soap emulsifying agents give a faster curing rubber, while the rosin acid–soap types yield a less tacky final product. Oils of different grades added to the reaction mixture of styrene and butadiene determine the final product's toughness.

A convenient feature of styrene–butadiene rubber is its final packaging in 34-kilogram (75–pound) polyethylene plastic bags, shaped for easy handling and shipping. If the minimum temperature in the user's mixing machine approximates 140" C (284" F), the rubber, bag and all, can be tossed into the mixer where the polyethylene will soften and disappear.

For many purposes SBR directly replaces natural rubber, the choice depending simply on economics. Its particular qualities include abrasion and crack resistance and generally better aging properties. Its limitations are poor strength (without such reinforcing fillers as carbon black), low resilience, low-tear strength (particularly at high temperatures), and poor tack (*i.e.*, it is not tacky or sticky to the touch). These characteristics determine its use in tire treads; its proportions decrease as need for heat resistance increases, until 100 percent natural is reached in the heaviest and most severe uses, such as tires for buses and aircraft, and for "off-the-road" vehicles. SBR, however, is used in great quantities for tire carcasses; generally some natural rubber is admixed to produce the necessary tack in assembly.

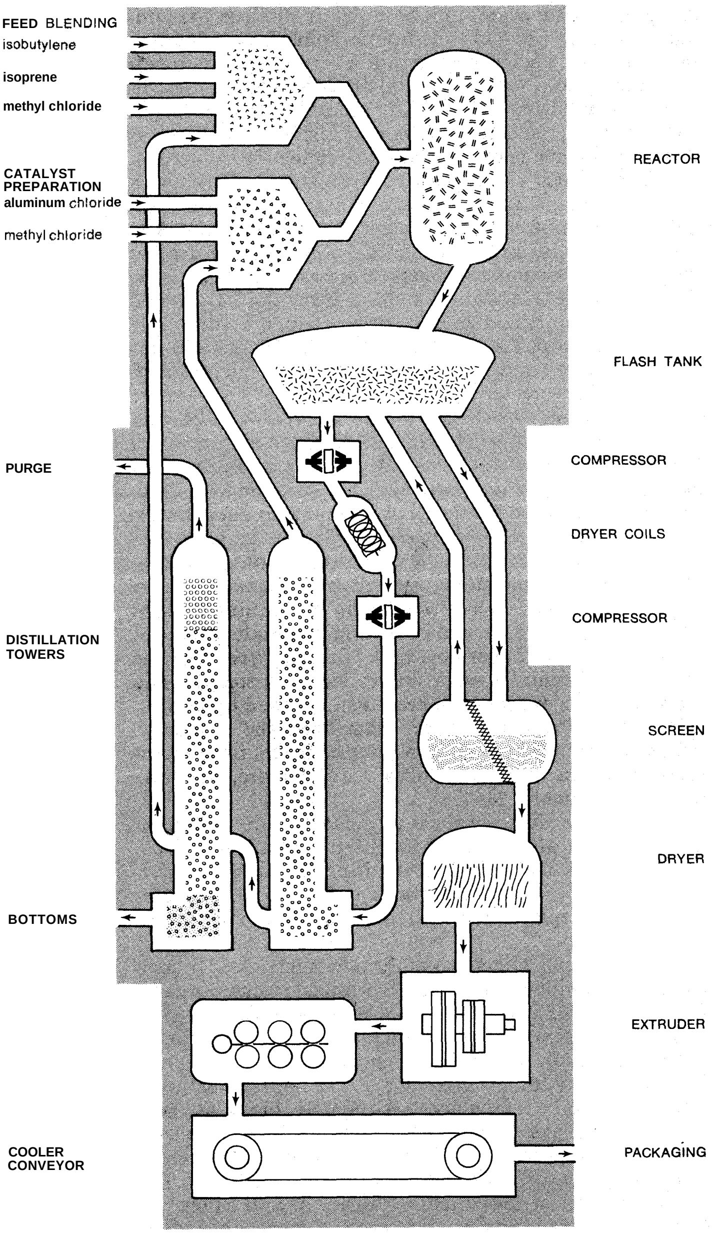
Butyl rubber. Butyl rubber (its symbol IIR) is essentially polyisobutylene, a polymer of isobutylene, which is an isomer of butylene. (Isomers are molecules with the same number of atoms but with different molecular structure and properties.) Isoprene is added during manufacture. Marketed in 1943, after ten years of research work, butyl rubber has enjoyed widespread acceptance because of some of its special properties: low permeability to gases, excellent resistance to oxygen and ozone at normal temperatures, and good resistance to vegetable oils. Butyl rubber is not readily attacked by acids, alkalies, and other chemicals harmful to natural rubber. In addition, it is elastic, odourless, nontoxic, light in colour, and can be worked on normal rubber machinery. Its electrical insulating properties are excellent.

Properties of butyl rubber

Butyl rubber has several disadvantages, especially its incompatibility with natural rubber and the main synthetics. In addition, butyl rubber has a strong affinity for foreign matter and must be carefully stored and screened before manufacture. Finally, it stiffens at low temperatures.

The polymerization of isobutylene is carried out at —95° C (—139" F), with aluminum chloride as a catalyst. Isobutylene and the desired proportion of isoprene are fed continuously into a medium of methyl chloride. After antioxidant is added, crumbs of butyl rubber form directly, without any latex stage. Zinc stearate (to prevent agglomeration, or formation of clumps of butyl rubber) also goes into the mixture. Excess reactants are recovered and re-used.

Because of its excellent air retention butyl rubber quickly took over from natural rubber for inner tubes in all but the largest sizes. Butyl rubber also plays an important part in the inner liners of tubeless tires. Despite excellent road-holding properties under wet conditions, and quiet, comfortable riding qualities, all-butyl tires have not proved commercially successful because of their poor tread durability. Butyl rubber, however, is used for many automobile components, such as window strips, because of its excellent resistance to oxidation. In its resistance to heat, butyl rubber also plays an indispensable part in tire



Principal steps in processing butyl rubber.
Drawing by D. Meighan

manufacture, forming the container for the hot water or steam required to vulcanize the inside of tires.

Polybutadiene. Polybutadiene (its symbol BR) is produced by a hydrocarbon solution process, in contrast with the water-emulsion method for styrene–butadiene rubber (SBR). Indeed, in making polybutadiene, water must be rigorously excluded to enable the catalyst to work. The resulting polybutadiene has better physical properties, particularly tensile strength and resilience, than styrene–butadiene rubber.

The main outlet for polybutadiene is in tire treads. Its high resistance to abrasion, low heat buildup, and resistance to cracking are strongly in its favour, particularly when blended with natural or styrene–butadiene rubber in giant tires. Its lower skid resistance when wet and lower tear resistance limit its proportions in car tires. Polybutadiene can absorb large proportions of oil and carbon black, thus reducing the cost of the final product and improving skid resistance appreciably. It is increasingly used to add strength and flexibility to plastics, highly resilient foams, and footwear; its low temperature performance favours outdoor uses.

Polyckloroprene (neoprene). Polychloroprene (its symbol CR) is the scientific name for the synthetic that bears the proprietary name Neoprene (which, however, has become generic). One of the first synthetic rubbers to be discovered, neoprene was first marketed in 1931. It is prepared by treating acetylene (C_2H_2) with cuprous chloride to form monovinyl acetylene, which, with hydrochloric acid, yields chloroprene. Competitive methods of production exist. It is manufactured in solid form (chips) and as latex.

Because of its general properties of high tensile strength, high resilience, good resistance to oxygen, ozone, and tearing, as well as an oil resistance only slightly inferior to nitrile (see below), neoprene is used extensively in the wire and cable industries, hose, extruded automobile parts, and for protective clothing. It also has excellent resistance to flame.

Nitrile. The original Buna N of German war production, the name nitrile (its symbol NBR) covers a range of copolymers of butadiene and acrylonitrile, the members differing from one another in their acrylonitrile content. The "standard" NBR material contains from 33 to 36 percent acrylonitrile, and others from 20 to 50 percent. The outstanding property of nitrile is its oil resistance, which increases in pace with its acrylonitrile content—but at the expense of low-temperature flexibility. Production is by an emulsion process quite similar to the styrene–butadiene rubber process; it can be made as solid sheets or liquid latex. Nitrile requires the addition of carbon black or other reinforcement to produce satisfactory physical properties, thus reducing the overall cost of nitrile, which is more expensive than conventional rubbers.

Nitrile's
oil
resistance

Nitrile is mostly used where high oil resistance is required, as for seals, gaskets, or other items subject to contact with hot oils. The rolls for spreading ink in printing and hoses for oil products are other obvious uses. Nitrile's reaction to oil is selective, however; it has excellent resistance only to some types. Nitrile applied to woven and unwoven fabrics improves finish and provides waterproofing. It blends readily with the plastic polyvinyl chloride, a mixture whose use for artificial leather and for paper coating is increasing.

Polyisoprene. The first true equivalent of natural rubber was achieved in 1955 by the polymerization of isoprene. Its product was polyisoprene (its symbol IR). Its manufacture follows the hydrocarbon solvent method for making polybutadiene (BR). Synthetic polyisoprene can replace natural rubber for most purposes, except when the nonrubber constituents of natural become advantageous. The decision is therefore usually based on economics, and since the basic monomer isoprene is expensive, either synthetic polybutadiene or natural rubber is usually favoured.

Silicone. A specialty rubber, silicone is limited in use by its cost; but its exclusive ability to retain its physical properties at extreme temyeratures, from −74" C (−101° F) to 315° C (599° F), makes up for its high cost. It is the first "inorganic" rubber; the polymer has a backbone of silicon and oxygen instead of the conventional carbon chain.

Although the raw material silica is cheap and plentiful, the manufacturing process is difficult and expensive. A range of products is available, their properties largely influenced by the molecular structure of the silicone. Its retention of electrical properties, resistance to oxygen and ozone, and inertness at extremely high temperatures recommend uses in aircraft and outer space equipment, and in the automobile field. Its inertness and nontoxicity encourage use for food and surgical applications.

Polysulfide rubber (Thiokol). Polysulfide rubber is the technical term for the substance bearing the trade name Thiokol, which has become generic. Originally synthesized in 1927, it is known for its excellent resistance to organic solvents and is marketed as a liquid, a putty, and in solid form. Its tensile strength does not compare with conventional rubbers, but it offers excellent resistance to weathering and permeabiiity over a wide range of temperature.

Ethylene–propylene rubber. A versatile new synthetic, ethvlene–propylene rubber exhibits excellent resistance to oxygen, ozone, acids, alkalies, and other chemicals over a wide range of temperature. It has good electrical properties and is quite lightweight, with an enormous capacity for acceptance of oil, thereby reducing the final cost. The addition of carbon black used as a filler, along with the oil, acts to raise the low tensile strength. In the early 1970s, ethylene–propylene was not yet sufficiently developed for a full assessment of its potential.

Polyurethane rubber. Polyurethane rubber includes a large range of materials of varied compositions, properties, and uses. In this group of rapidly growing importance, some of the polymers help significantly to bridge the gap between rubber and plastics. All are based on isocyanates, organic compounds that combine with other organic compounds such as alcohols, esters, ethers, and amides. The wide variety of materials available, when properly controlled, offers a series of polymers with a broad range of properties and forms.

Casting, mixing, and milling of polyurethanes

Polyurethanes may be cast from the liquid or semiliquid state, or mixed, milled, and vulcanized like other rubbers. Although they have high tensile strength and good resistance to abrasion and to oil and other solvents, they are not suited for normal tire service because of their rapid drop in abrasion resistance as temperature increases. Polyurethane is, however, finding increased use in small, solid tires. Probably the most popular uses of polyurethane are as foams and surface coatings: rigid foams for lining refrigerators and flexible foams as a substitute for natural rubber and durable finishes for floors and furniture.

Chlorosulfonated polyethylene (Hypalon). Chlorosulfonated polyethylene is normally known by its trade name of Hypalon. It is made by introducing first sulfur dioxide into a solution of polyethylene under strictly controlled conditions and then chlorine at higher temperature. Chlorine content may vary between 20 and 45 percent and sulfur from 0.5 to 2.75 percent. The physical properties depend much more upon the sulfur than the chlorine content. Rubbery compounds are produced with low sulfur and metallic oxide curing. Hypalon is outstanding in its resistance to weathering, ozone, heat, flame, and chemicals. Cost alone restricts its wider adoption.

Fluorocarbon polymers. Probably the best known fluorocarbon polymer is PTFE—polytetrafluoroethylene—although similar products share polytetrafluoroethylene's outstanding qualities of heat resistance, minimal friction, and freedom from surface adhesion. Originally these materials could not be extruded or vulcanized, but rubbery polymers have been developed that can be worked on conventional equipment. A recent use that has popularized this polymer is as a lining for cooking utensils; its nonsticking quality facilitates removal of cooking residues by simple washing.

Reclaimed rubber. Reclaimed rubber, or vulcanized rubber reworked to render it suitable as raw material, has become an important element in the industry. The starting point in reclaim rubber manufacture is the cured rubber scrap, natural or synthetic in origin, which is segregated into separate incompatible rubbers and then graded according to quality and intended use. Reclamation methods differ in sophistication, so that an appreciable range of reclaims is available.

It is not profitable to use reclaim unless it costs no more than half as much as virgin rubber, and when that new rubber is low-cost synthetic, the market for reclaim suffers.

Digestion of scrap rubber

Reclaim is manufactured in various ways. The simplest is digestion of scrap (to which oil has been added) either in caustic soda or in zinc chloride (the more modern method) to remove nonrubber products; it then is milled. Under the newer, continuous "reclaimator" method, any fabric is first removed mechanically; the rubber then is ground, mixed with oil, and extruded.

COMPOUNDING

Blending rubbers in prescribed proportions and adding chemicals in accordance with a "formula," or schedule, is called compounding. Initially, these formulas were guarded with great secrecy, often unnecessarily. Such secrecy has largely disappeared; most suppliers disclose the contents of their products. Although some manufacturers prefer to show each constituent of a formula as a percentage of the whole, the more usual practice is to list ingredients in relationship to 100 parts of rubber hydrocarbon.

The materials added to rubber are known variously as drugs, pigments (not necessarily confined to colours), and fillers. Since all rubbers, natural and synthetic, with the exception of some already containing fluorine, require addition of chemicals, a filler is not to be mistaken for an adulterant or cheapener. Of course, the vulcanizing agent, sulfur, must be added. But by itself it is slow, so that it needs accelerators, which, in turn, can be speeded up with secondary accelerators, or activators. To delay the decomposition of rubber products, antioxidants and antiozonants are required. (These and all other such protective materials are termed antidegradants.) To increase strength, rubbers need reinforcing agents—*e.g.*, carbon blacks, which are available in a wide range and impart a black coloration. Other materials may be added, as well as oils for processing, plasticizing, or softening. Finally, inert materials may be added for bulk, thus reducing cost.

Carbon-black filler

Reinforcing and filling materials. The outstanding material used for the last 50 years or more for improving elasticity and tensile strength of rubber has been carbon black. Early carbon black was made by burning natural gas in multiple fine jets and collecting the product on cold metal. Under the modem method, "furnace type" black is made by burning oil in a controlled atmosphere. The many grades of blacks have been distinguished by initials; HAF (high abrasion furnace) has been the mainstay. Every pound of rubber used in tire treads requires at least one-half pound of black; tubes require more, and carcasses only slightly less.

To abate the dust nuisance associated with the use of carbon black, alternatives have been investigated, especially silica in finely divided condition. Though considerably more expensive than carbon black, it has had success as a partial replacement in treads for large off-the-road tires.

Inert inorganic fillers basically add substance, though clays and whiting have some technical value; organic fillers, more akin to rubber, are now increasingly used.

Colouring agents. Pigments for colouring rubber must withstand curing temperatures and not interfere with or be affected by other ingredients. In high-temperature cures, therefore, the choice is limited to inorganic substances: zinc oxide and some titanium dioxide, plus a trace of ultramarine (to kill any creamy tinge) to achieve whiteness; iron oxide, to produce browns and reds. For rubbers curing at lower temperatures, a wide choice of organic dyes is available. For colour work, any added antioxidant must be nonstaining (see DYES AND DYEING).

Plasticizers and softeners. Adding mineral oils to aid processing is as old as the industry. Usually 5 to 8 percent of oil was added to natural rubber. The introduction of synthetics, with their greater affinity for oil, has led to more scientific standards.

Studies have shown that increasing the oil content of natural rubber improves certain of its properties, particularly traction of tires on snow-packed roads. In addition to mineral oils, a wide variety of softeners has been used, including coal-tar by-products such as bitumen, and Vaseline, fatty acids, and wood resins. Pine tar has always been popular with natural rubber because of its value in dispersing carbon black.

Vulcanizing agent. The outstanding vulcanization agent and the one longest used is sulfur. Cheap, effective, and relatively free from problems, it is added in proportions of up to 3 percent, and much higher for hard rubbers. Sulfur has one peculiarity: regardless of how little or how much is used in vulcanizing, some remains uncombined and is known as free sulfur, which migrates to the surface forming an unattractive sulfur bloom. Insoluble sulfur, considerably more expensive, can be substituted to avoid blooming.

Sulfur bloom

Accelerators of vulcanization. Accelerators or catalysts form an ever-growing list. Mercaptobenzothiazole, which revolutionized vulcanization in the 1920s, is still in use, although much safer and more efficient products have been developed since. As a class, accelerators have enabled curing times and temperatures to be reduced significantly, rendering the operation more efficient and permitting the introduction of valuable materials that would be destroyed under the earlier conditions.

Antioxidants. These, as their name implies, resist the action of oxygen. They form an important section of the group of materials known as antidegradants, which encompasses all materials resisting the deterioration of rubber. The amount of antioxidant added per pound of rubber has steadily risen, including the amounts indigenous to natural rubber and the amounts introduced as stabilizers in styrene–butadiene and other synthetics during manufacture.

Antioxidants may be divided into two main classes: amine derivatives, among which the more powerful are generally complicated condensation products with aldehydes or ketones, which cannot be used in light-coloured stocks because of staining; and phenol derivatives, which stain less and are more efficient. A third and smaller class of antidegradants fulfill such specific purposes as resisting the catalytic action of copper in breaking down rubber.

Derivatives of p–phenylenediamine occupy a prominent position among modem materials. Despite their excellent resistance to oxygen and ozone and good flex resistance, great care has to be taken to minimize staining and more particularly to ensure freedom from skin irritation.

Zinc oxide. Versatile zinc oxide belongs in a category by itself. Mention has been made of its use as a colouring agent. Although its use as a reinforcing agent has declined, zinc oxide finds its way into most formulas, primarily because of its activating powers. It makes accelerators more efficient and generally gives stability to products. Its high density makes it an expensive addition because volume rather than weight is the criterion in filling molds. Stearic acid should be associated with zinc oxide, for although as a fatty acid it might be considered a softener, it is normally added in such proportions as to convert the zinc oxide to zinc stearate, which is thereby made more effective.

Other compounding materials. Resins fulfill a variety of important functions. Mention has been made of wood resin softeners, and also of phenolic resins for high-temperature curing. Other resins are added to rubber to make it tacky and still others to stiffen it.

Several waxes may be added either as softeners or to form protective blooms on the surface of the product.

With the introduction of faster accelerators in manufacture, retarders have been developed; they are designed to make processing safer without slowing down vulcanization.

Peptizers, organic sulfur compounds that save mixing time by breaking down the huge molecular structures of the raw rubbers, are effective in small amounts. They are selective, however, and their performance is dependent on the presence or absence of other materials. Their relatively high cost must be justified by the mixing time saved.

Polyac and Elastopar

Where butyl rubber is used in compounding, the material Polyac serves as a promoter, a substance a minute quantity of which when heated with the polymer speeds subsequent action with the accelerator. A somewhat similar substance is Elastopar, used with styrene–butadiene rubber, but not effective below 149" C (300" F). Both promoters act vigorously and because of the difficulty of control are not always acceptable. To overcome that difficulty, they are sold mixed with an excess of inert material.

Lubricating additives cover a wide range. Dusting powders applied to uncured rubber prevent pieces from sticking together and ease their removal from molds. The standard materials are talc or french chalk. Zinc stearate is much more expensive but has the outstanding advantage of not causing cracking because it is soluble in rubber and, therefore, does not act as a foreign substance at the point of contact between two rubber surfaces. Liquid lubricants fall broadly into two categories: suspensions of talc, clay, or soap in water or alcohol; and the higher costing silicones, which have proved their effectiveness as mold release agents.

Vast quantities of solvents are consumed by the industry. Basically, they are petroleum distillates often known as naphthas, added in the factory to provide or restore tack to uncured rubber. Other solvents include aromatics obtained from coal tar and such chlorinated hydrocarbons as trichlorethylene and carbon tetrachloride.

PROCESSES

Mastication. Mastication is the process of reducing the size of the large organic molecules in natural rubber to increase its plasticity and thus facilitate subsequent mixing. Mastication may be accomplished in a specially designed machine, on open mills, or by merely giving the rubber an extra two minutes in a mixer before the addition of other ingredients. The process, therefore, is not precise but depends upon the condition of the rubber. The necessity of mastication may be minimized by the prior controlled storage of rubber at elevated temperatures, the use of low-viscosity rubber, or the use of special plasticizers.

Mixing. Open mills originally performed mixing. They were machines with two adjacent rollers mounted horizontally and rotating in opposite directions; the clearance between the rollers (the nip) was adjustable. Open mills, however, are slow, hazardous, and not amenable to temperature control; their replacement by internal mixers (see below) was accelerated by the increasing incorporation of powdery carbon black. Open mills still are used for warming up rubber before extrusion or sheeting, for low-temperature mixing, masticating, and other specialized work.

Internal mixers

An intemal mixer consists of two horizontally mounted rotors with blades that intermesh and shear the compound against the lining, with most of the mixing done in the small clearance between the outside of the rotors and the lining of the mixer. Mixing is regarded as complete at the end of a given time period, after a given temperature is attained, or after a certain amount of power is consumed. The mixed batch, normally in the form of discrete lumps, is sheeted out by dropping it on a mill mounted immediately below the mixer.

Forming. Uncured solid rubber not required in sheets is shaped or formed, generally in an extruder or tuber. The apparatus is frequently compared with a sausage machine because raw material is fed into a hopper at one end and forced through an aperture at the other, along a barrel containing a tightly fitting screw. Temperature is controlled by hot or cold water circulating around the barrel, and a variety of dies can be fitted to the head to produce the desired shapes. The rubber must be consistent in plasticity, for the cut of the die is correct for only one set of conditions, including temperature and speed. Extruders, defined by screw diameter, range from 1½ inches (38 millimetres) up to as much as 12 inches (305 millimetres) for very large tire treads.

The compounded rubber is warmed up before being shaped, generally on open mills. The process is extravagant in use of time and material for setting up before the specified dimensions are obtained; and when the product is cut, it shrinks on cooling, with "belled ends," sections that have to be discarded. Extruders may be used in twin or triple formation, each with its own rubber stock, to provide hot unions—*e.g.*, a tough tread on top of a more resilient base or a tread with flexible sidewalls.

Injection molding. The waste and delays associated with extrusion have led to more efficient processes such as cold feed mixer-extruders and injection molding machines similar to those used for plastics. With modification to guard against incipient cure in the barrel, they extrude the rubber, not through a die, but directly into a mold where it is given a very short cure. Success depends on extremely long runs, which are not common. Multiheads that allow four molds to be operated from one machine increase output.

Calendering. Calenders have two main functions: one, to produce sheet rubber of accurate and consistent gauge from a few thousandths to ⅛ inch (3 millimetres) or

Rolling plies and coating fabrics

more by plying up (using multiple plies); and, two, to rubberize fabric with great accuracy. In its crudest form, a calender is essentially a two-roller mill, although the rollers are normally mounted vertically. Modem calenders comprise three or four rollers ground very accurately.

In the typical three-roller calender for rubberizing fabric, all rollers lie in the same vertical plane, the top and bottom rotating in the same direction. Warmed-up rubber is fed into the nip between the two top rollers; it follows around the middle roller, and is applied to fabric entering the machine between the two lower rollers. The speed of the centre roller can be varied in relation to the bottom: when it is even with the bottom roller, it coats the fabric; when it is higher, it frictions, or forces the rubber into the interstices of the fabric.

A three-roller calender treats only one side of fabric at one pass. To rubberize both sides, the alternatives are a second pass, the use of a second calender (preferably in train with the first) or a four-roller calender, which coats both sides in one operation. The four rollers may be in line vertically, or the top roller may be offset to form an inverted "L." A calender must be supported with an array of auxiliary equipment, including variable-speed motors; efficient means of feeding both rubber and fabric and collecting the treated material, spreading the fabric across the roller under uniform pressure; accurate gauging; and temperature control. The rollers are traditionally cored to permit the flow of cold water or steam, as required. In a more modern method of temperature control, the rollers are drilled longitudinally forming a series of bores that allow steam or water to circulate within an inch of the surface.

Vulcanization. Reference already has been made to the principles of vulcanization and the chemicals used in the process. The simplest form of vulcanizer is the heater or autoclave, a vessel capable of retaining moderate pressures of steam and air (or inert gas). Simple shapes, and particularly thick sections of uncured rubber, may be exposed directly to steam, or molds with their cavities filled with uncured rubber may be inserted in the vessels. So-called bag-o-matic curing is the modem method of tire vulcanization, except for extremely large off-the-road tires. Bag-o-matic curing is entirely automatic and has reduced cure times for car tires from about two hours to 12 to 14 minutes. Heat is applied to the inside of the uncured tire through the medium of a one-quarter-inch gauge bladder, which has supplanted the much heavier and less efficient air bag. Time, temperature, and pressure are controlled automatically.

Bag-o-matic curing

Smaller rubber products are normally formed in a mechanical press, usually the daylight press on which a number of individual molds can be mounted. The molds are held in a closed condition by a hydraulic ram operating in a vertical plane. Another technique employs a single press that accommodates one mold, possibly with several cavities, the two parts of the mold being held together by mechanical pressure. The operation of the single press is largely automatic.

Cold curing, or vulcanization at ambient temperature, is still practiced but only where it is unavoidable, as in repairing belts *in situ.* Special accelerators can perform at room temperatures, but controlled heat, time, and pressure are always preferable.

Infrared and high energy radiation. These methods of vulcanization are of scientific rather than commercial interest. Use of internally generated heat to cure thick articles saves considerable time. Because of the poor conductivity of rubber, heat applied to the surface only penetrates slowly. The initial expense of equipment, plus operating costs, has limited interest to specialized high-temperature operations.

MAJOR APPLICATIONS

Tires and tire products. Automobile tires absorb 60 to 70 percent of the available rubber. As the outside diameter of tires has dropped progressively, widths of treads have increased, providing wider contact area with the ground and improved stability. The major strength of the wide-tread radial tires is built in transversely to the direction of motion; a series of inextensible narrow bands under the centre of the tread prevents its spreading. The conventional tire, in contrast, has cross plies. Increasing varieties of carcass materials are available: cotton, rayon, nylon, polyester, steel wire, and glass fibre, the last two for radial tires.

Tire-making remains a blend of craftsmanship and automation. Despite advances in mechanization, personal skill and experience are dominant in extruding, calendering, and in assembling all the components together in the process of tire building. The carcasses are constructed by laying ply upon ply of rubberized "fabric" on a cylindrical drum, followed by the other components, and ending with the tread and sidewalls (generally together as one unit). Adhesion in the uncured state relies upon the natural tack of the rubber.

Rainwear and shoes. Rainwear represented the first commercial use of rubber. Vulcanization overcame the obstacles encountered by Macintosh and the process of rubber coating fabric has continued into modem times. Doughs of rubber, naphtha, and other ingredients are fed between the roller transmitting the fabric and a "doctor" blade that spreads the rubber onto the fabric and controls thickness. The treated fabric then passes over a heated framework and is wound into rolls. If the heat is not sufficient to cure the rubber, an additional operation is performed.

Spreading has largely given way to the greater speed and accuracy of calendering, as has rubber to oil-soap treatment for shower-proofing and to plastics for more serious protection.

Shoes consume more rubber than any other commodity except tires. Styrene–butadiene types of rubber with high styrene content are favoured for soles because their wear characteristics render carbon black unnecessary. Attachment of soles to uppers by stitching has been largely replaced by either polymer adhesives or direct vulcanizing, a process requiring expensive equipment but providing rapid curing of two-and-one-half to three minutes. One-piece soles and heels made of rubber resins have been used successfully in children's footwear. Microcellular rubber, which feels soft and comfortable in wear, also has gone into soling.

Industrial uses. One highly important industrial use of rubber is in belting. The tendency for each piece of machinery to have its own electric motor has limited the use of driving belts, but the conveyance of coal and other minerals over thousands of yards has provided valuable business for the manufacturer. Strength is obtained by plying up any of the tire textiles, with polyester preferred where nonstretch is critical. Where fire is a hazard, the outer cover of belting is usually polyvinyl chloride. Cure may be either continuous or by sections, in long, specially constructed presses. The development of travelling sidewalks for passengers at airports and similar traffic centres opens up new and substantial possibilities.

Hoses of all shapes and sizes form an important sector of the rubber business. Small ones, such as garden hoses, are being replaced by polyvinyl chloride largely on a weight basis; otherwise an endless variety exists. The type of synthetic used depends on the material to be transmitted; some hoses must carry corrosive chemicals at high and low temperatures. Type of construction varies with the weight of the fluid to be transported and its pressure. Small all-rubber hoses may be molded to complicated shapes, but generally they are an extruded product. If reinforcement is required, it is primarily cord fabric plaited just ahead of the extruder. An outer cover, possibly of different composition, designed to resist abrasion, follows, and if further reinforcement is required, wire is wound spirally. Small hoses with strong walls that do not collapse under heat may be coiled and so vulcanized, but the larger types must be cured straight to prevent collapse.

Rubber-cable insulation

Cables consume a substantial amount of rubber as insulation. Wire (normally copper) is drawn through an extruder and the insulation applied. Butyl rubber is preferred because of its resistance to heat and moisture; an outer cover of either butyl rubber or neoprene is added, but both are increasingly being replaced by polyvinyl chloride.

Rubber is useful for many more industrial or engineering purposes, generally to insulate against sound or vibration or both, as in building and bridge foundations. A wide range of rubber springs and flexible seals performs valuable service; rubber fenders are ideal to protect the sides of ships and docks because the rubber retains its resilience for a score or more years despite destructive workings of the sea, oxygen, ozone, and other agencies.

BIBLIOGRAPHY. H.J. STERN, ***Rubber: Natural and Synthetic***, 2nd rev. ed. (1967), is the most up-to-date comprehensive work on the subject. Another good basic text is M. MORTON (ed.), ***Introduction to Rubber Technology*** (1959). Two classic, but somewhat dated works are C.C. DAVIS and J.T. BLAKE (eds.), ***The Chemistry and Technology of Rubber*** (1937); and G.S. WHITBY (ed.), ***Synthetic Rubber*** (1954). See also L. BATEMAN (ed.), ***The Chemistry and Physics of Rubber-Like Substances*** (1963), which contains an excellent article on the botany of the rubber tree; and W.J.S. NAUNTON (ed.), ***The Applied Science of Rubber*** (1961). For history, see P. SCHIDROWITZ and T.R. DAWSON (eds.), ***History of the Rubber Industry (1952).*** Current statistical information may be found in the monthly periodicals, ***Rubber News*** and ***Rubber Statistical Bulletin.***

(L.R.M.)